5-1-2010

Vol. 9, No. 1 (Full Issue)

JMASM Editors

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Recommended Citation
DOI: 10.22237/jmasm/1272688320
Available at: http://digitalcommons.wayne.edu/jmasm/vol9/iss1/33

This Full Issue is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.
Invited Articles
2 – 8    Ronald C. Serlin    Fisher Was Right
9 – 14   Rand R. Wilcox    Inferences about the Population Mean:
                         Empirical Likelihood versus Bootstrap-t

Regular Articles
15 – 27  Tim Moses,        The Influence of Data Generation on
      Alan Klockars    Simulation Study Results: Tests of Mean
                         Differences
28 – 42  Marie Ng,        The Small-Sample Efficiency of Some
      Rand R. Wilcox    Recently Proposed Multivariate Measures
                         of Location
43 – 51  Erik L. Heiny,    The Effectiveness of Stepwise Discriminant
      Daniel J. Mundfrom Analysis as a Post Hoc Procedure to a
                         Significant MANOVA
52 – 63  Marc H. Kroopnick, Assessing Classification Bias in Latent Class
      Jinsong Chen,      Analysis: Comparing Resubstitution and
      Jaehwa Choi,       Leave-Out Methods
      C. Mitchell Dayton
64 – 74  Walter Leite,     The Performance of Multiple Imputation for
      S. Natasha Beretvas Likert-type Items with Missing Data
75 – 89  H. E. T. Holgersson, Model Based vs. Model Independent Tests for
      Peter S. Karlsson  Cross-Correlation
90 – 94  Madhusudan Bhandary On Exact 100(1-α)% Confidence Interval of
                         Autocorrelation Coefficient in Multivariate
                         Data When the Errors are Autocorrelated
95 – 102 Nol Bendermacher   Beyond Alpha: Lower Bounds for the
                         Reliability of Tests
103 – 115 Weihua Fan       Impact of Measurement Model Modification
                         on Structural Parameter Integrity When
                         Measurement Model is Misspecified
Faisal G. Khamis, Abdul Aziz Jemain, Kamarulzaman Ibrahim

On a Comparison between Two Measures of Spatial Association

Carolyn F. Furlow, S. Natasha Beretvas

An Evaluation of Multiple Imputation for Meta-Analytic Structural Equation Modeling

Dimitrios D. Thomakos

Median-Unbiased Optimal Smoothing and Trend Extraction

Samuel B. Green, Marilyn S. Thompson

Can Specification Searches Be Useful for Hypothesis Generation?

Gaetano Ferrieri

Measuring Openness

Haiyan Bai, Wei Pan, Leigh Lihsing Wang, Phillip Neal Ritchey

Another Look at Resampling: Replenishing Small Samples with Virtual Data through S-SMART

Stan Lipovetsky

Nonlinear Parameterization in Bi-Criteria Sample Balancing

Gyan Prakash

Shrinkage Estimation in the Inverse Rayleigh Distribution

Abedel-Qader Al-Masri

Combining Independent Tests of Conditional Shifted Exponential Distribution

Set Foong Ng, Heng Chin Low, Soon Hoe Quah

A New Biased Estimator Derived from Principal Component Regression Estimator

Hassan Pazira, Parviz Nasiri

Estimations on the Generalized Exponential Distribution Using Grouped Data

Carlos N. Bouza

Ranked Set Sampling Using Auxiliary Variables of a Randomized Response Procedure for Estimating the Mean of a Sensitive Quantitative Character
255 – 262  Kouji Yamamoto, Sadao Tomizawa  Symmetry Plus Quasi Uniform Association Model and Its Orthogonal Decomposition for Square Contingency Tables


274 – 286  Breda Munoz, Virginia M. Lesser, Ruben A. Smith  Applying Multiple Imputation with Geostatistical Models to Account for Item Nonresponse in Environmental Data


296 – 303  Reza Kaihani, Ali Reza Seifi  Optimal Meter Placement by Reconciliation Conventional Measurements and Phasor Measurement Units (PMUs)

Brief Reports
304 – 307  Markus Neuhäeuser  An Equivalence Test Based on $n$ and $p$

308 – 313  David Parker  Derivation of Mass Independent Quantum Treatment of Phenomenon

JMASM Algorithms & Code
314 – 331  Dongquan Zhang, C. Mitchell Dayton  JMASM30 PI-LCA: A SAS Program Computing the Two-Point Mixture Index of Fit for Two-class LCA Models with Dichotomous Variables (SAS)
JMASM is an independent print and electronic journal (http://www.jmasm.com/), publishing (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Editorial Assistant: **Julie M. Smith**
Internet Sponsor: **Paula C. Wood**, Dean, College of Education, Wayne State University
INVITED ARTICLES

Fisher Was Right

Ronald C. Serlin
University of Wisconsin-Madison

Invited address presented to the Educational Statistician’s Special Interest Group at the annual meeting of the American Educational Research Association, Denver, May 1, 2010.

Key words: Fisher.

---

Introduction
I would like once again to thank you for awarding me this honor last year. Given the scholars between whom I am sandwiched, the first honoree, Ingram Olkin, and next year’s, Joel Levin, I must try very hard to act as though the committee did not make a serious mistake with my nomination. Tonight, I’d like to focus on some of the work of R. A. Fisher, who would have been 120 years old now, to make a couple of points of my own. I hope that some of what I say will give you the same feeling of fun in the discovery of something neat and surprising as I experienced.

Early Years: Up to 1922
Fisher held two chairs in genetics, at University College in London and then at Cambridge but, surprisingly, was never a professor of statistics. Regarding Fisher’s accomplishments in statistics, Savage (1976) commented that it would be easier to list the few topics in which he was not interested. “In the art of calculating explicit sampling distributions, Fisher led statistics out of its infancy, and he may never have been excelled in this skill” (p. 449).

There is much, of course, about which Fisher was right. Despite his shunning the concept of Type II errors, Fisher (1928) was the first to provide formulas for the noncentral Chi-square, t, and F distributions. (The symbol F was introduced by Snedecor in honor of Fisher, “for which officiousness,” according to Savage, “Fisher seems never to have forgiven him” (p. 449)). There once existed a fair amount of disagreement regarding how to count degrees of freedom in a contingency table, with Karl...
Pearson (among others) claiming $rc - 1$ and Fisher (1922) correcting to $(r - 1)(c - 1)$. Fisher, of course, was right here. Fisher was a pioneer in nonparametric statistics, having suggested the use of the sign test in place of the t-test in certain designs, and having introduced what he called exact tests to avoid the assumption of normality in many circumstances.

According to Stigler (2005, p. 33), of Fisher’s 97 publications from 1912 to 1920, 91 were in the *Eugenics Review*, two were on genetics related to eugenics, two were papers published in *The Messenger of Mathematics*, and the other two (in 1915 and 1920) were on mathematical statistics. I’ll focus briefly on the 1915 and 1920 papers, as described by Stigler (2005, 2006).

Mathematically, the 1915 derivation of the distribution of the sample correlation coefficient was the kind of work to which we all strive. Fisher found the distribution, expressions for moments, transformations (r-to-z) and distributional relationships (including his earlier work on the Student’s t-distribution), expressions for the bias of $r$, and the maximum likelihood estimator of $\rho$.

Right Nice Stuff

This type of work led Neyman (1951), in his review of Fisher’s *Contributions to Mathematical Statistics* (1950), to describe Fisher as “a very able ‘manipulative’ mathematician” (p. 406). The *Contributions* contain prefatory comments by Fisher on the various papers. For the 1915 paper, Fisher wrote “Here the method of defining a sample by the coordinates of a point in Euclidean hyperspace was introduced...” (p. 87). Unfortunately, according to Neyman (1951), representing the sample by a point in space was used for a similar purpose by Karl Pearson in 1900 and - Neyman suspected - had probably been used even before that; thus, Fisher was wrong in this regard.

During the year following the publication of Fisher’s article on the correlation coefficient, Kirstine Smith (1916), working at Karl Pearson’s laboratory, published an article suggesting that when fitting a frequency curve with grouped data, the constants should be estimated using a minimum Chi-square criterion. She illustrated the use of this criterion through a series of examples. She stated that compared to the use of the minimum Chi-square method of fit, other approaches were arbitrary, including what she termed “the Gaussian ‘best’ value,” (p. 262) the maximum likelihood approach from error theory that Fisher had supported in a paper he wrote as an undergraduate student in 1912. According to Stigler (2005), in response to a letter and manuscript that Fisher submitted to *Biometrika*, Karl Pearson as editor told Fisher that he had to demonstrate the logic of maximum likelihood, to justify it being better than Smith’s approach. For a while Fisher could not respond.

The basis for Fisher’s reply came, possibly by accident (Stigler, 2005), in the late spring of 1919. Fisher was considering the relative merits of two alternative estimates of the standard deviation of a normal distribution: one was based on the mean absolute deviation, the other the maximum likelihood solution. He had considered combining the two estimates in some way but instead discovered that the whole of the information regarding $\sigma$, which a sample provides, is summed up in the value of the maximum likelihood estimator. Not only did it have a smaller standard deviation, it was, in a word, sufficient.

On November 17, 1921, Fisher read a paper to the Royal Society of London entitled *On the Mathematical Foundations of Theoretical Statistics*. The paper opened with a set of definitions that were, in 1921, entirely new to statistical theory, but which are now familiar; they include consistency, efficiency, estimation, likelihood, optimum, and sufficiency. Stigler (2005) pointed out that not in the list is “…another, even more basic statistical concept: It is in this paper of Fisher’s that the word ‘parameter’ is first used in the modern statistical sense” (p. 32). Stigler notes that the word parameter appears 57 times.

According to Fisher, a consistent estimate is called efficient if it is asymptotically normal and if it has the minimum asymptotic variance (Neyman, 1951). In his 1908 paper, however, Edgeworth expressed the idea that maximum likelihood estimates are always efficient and made several attempts to prove his conjecture. The proofs, however, “…of the efficiency of maximum likelihood estimates
offered both by Edgeworth and by Fisher are inaccurate, and the assertion, taken in its full generality, is false” (Neyman, 1951, p. 407). So Fisher was wrong in the assertion, the proof, and in not giving Edgeworth some credit for priority.

Summarizing Fisher’s work, Neyman (1951) wrote, “…three major concepts were introduced by Fisher and consistently propagated by him in a number of publications. These are mathematical likelihood as a measure of the confidence in a hypothesis, sufficient statistics, and fiducial probability,” (p. 407) all employed by Fisher in the service of scientific induction.

Inference

Fisher (1947) felt that “the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis” (p. 16). Regarding the rate of error to assign to an incorrect rejection of the null hypothesis, Fisher wrote (1926) that “it is convenient to draw the line at about the level at which we can say: ‘Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials.” “A scientific fact,” he went on, “should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance” (p. 504). Further, Fisher (1973) wrote, “…in the vast majority of cases the work is completed without any statement of mathematical probability being made about the hypothesis or hypotheses under consideration. The simple rejection of a hypothesis, at an assigned level of significance, is of this kind and is often all that is needed, and all that is proper, for the consideration of a hypothesis in relation to the body of experimental data available” (p. 40). This all seems right.

Regarding Type II errors, Fisher (1947) wrote that “the notion of an error of the so-called ‘second kind,’ due to accepting the null hypothesis ‘when it is false’ may then be given a meaning in reference to the quantity to be estimated. It has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true” (p. 17). Thus, Fisher felt that one could not commit a Type II error, because one never drew a conclusion on the basis of a non-rejection of the null hypothesis. As he wrote (Fisher, 1973), “To a practical man, also, who rejects a hypothesis, it is, of course, a matter of indifference with what probability he might be led to accept the hypothesis falsely, for in his case he is not accepting it” (pp. 41-42). Some rightness to this is evident.

Fisher always desired to establish a correct theory of statistical inference. According to Kempthorne (1976) “Fisher really did think that one could develop by logical reasoning a probability distribution for one’s knowledge of a physical constant” (p. 496). Fisher, as Neyman (1951) pointed out, seemed proud to have formulated a measure of rational belief. Thus, Fisher (1973) wrote that the level of significance “in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders” (p. 43). Similarly, Fisher (1925a) had observed that “if the value of $P$ so calculated turned out to be a small quantity such as 0.01, we should conclude with some confidence that the hypothesis was not in fact true of the population actually sampled” (p. 90).

In similar vein, Fisher (1935c) stated “more generally, however, a mathematical quantity of a different kind, which I have termed mathematical likelihood, appears to take its place as a measure of rational belief…” (p. 40). In addition, Fisher (1973) commented that “the actual value of $P$ obtainable from the table by interpolation indicates the strength of the evidence against the hypothesis” (p. 80). And finally he also stated (1973) “What has now appeared is that the mathematical concept of probability is, in most cases, inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate…I have used the term ‘Likelihood’” (pp. 9-10). There is a whole lot of wrong here, as a measure of rational belief - even if obtainable - provides a theory with no level of epistemological virtue.

Note that even Neyman (1956) was not immune to this inductive probability infection, for he wrote in defense of control of the Type II
error rate “...the numerical values of probabilities of errors of the second kind are most useful for deciding whether or not the failure of a test to reject a given hypothesis could be interpreted as any sort of ‘confirmation’ of this hypothesis” (p. 290).

Fiducial Probability and Fiducial Intervals
Fisher (1935b) wrote on fiducial probability and fiducial intervals, about which he stated, “This form of argument leads in certain cases to rigorous probability statements about the unknown parameters of the population from which the observational data are a random sample, without the assumption of any knowledge respecting their probability distributions a priori.” His argument seems basically the same as that which leads to confidence intervals.

Defining $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$, Fisher noted that the probability statement $P(t > t_\alpha) = \alpha$ can be solved in terms of $\mu$ to yield $P(\mu < \bar{x} - t_\alpha s/\sqrt{n}) = \alpha$. Fisher believed that this probability statement holds even after the sample values are substituted. Conversely, Neyman and Pearson contended that at that point, the probability is either zero or one. Neyman (1956) offered a counter-argument in terms of two flips of a fair coin, where the variable $Y$ is the number of heads appearing. So it may be written that $P(Y = 1) = 0.5$ before the experiment. If $Y = 2$ is observed, Fisher would say the probability statement holds after substituting, or that $P(2 = 1) = 0.5$. Fisher appears to be wrong in this case.

To summarize, in Neyman’s (1951) words, “Unfortunately, in conceptual mathematical statistics Fisher was much less successful than in manipulatory, and of the three above concepts only one, that of a sufficient statistic, continues to be of substantial interest. The other two proved to be either futile or self-contradictory and have been more or less generally abandoned” (p. 407). As may be observed, it is fiducial probability that Neyman considered self-contradictory, and I agree that a search for a measure of rational belief is futile. Thus, for Fisher, one out of three right will have to do.

Personality
Fisher was not always charming and gracious, and his running battles with Neyman are well known. Regarding Karl Pearson, he wrote, “Pearson’s energy was unbounded. In the course of his long life he gained the devoted service of a number of able assistants, some of whom he did not treat particularly well. He was prolific in magnificent, or grandiose, schemes capable of realization perhaps by an army of industrious robots responsive to a magic wand” (1973, p. 2).

In similar vein, in a prefatory note on Fisher’s Contributions to Mathematical Statistics is a personal attack on Sir Karl: “If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age. Unscrupulous manipulation of factual material is also a striking feature of the whole corpus of Pearsonian writings, and in this matter some blame does seem to attach to Pearson’s contemporaries for not exposing his arrogant pretensions” (p. 437). On multiple occasions, Fisher (1958) criticized the ability of mathematicians to do science; for example he wrote “…with mathematical symbols, they are of course experts. But it would be a mistake to think that mathematicians as such are particularly good at the inductive logical processes which are needed in improving our knowledge of the natural world, in reasoning from observational facts to the inferences which those facts warrant” (p. 261). Judging by most of those in this audience, I believe that Fisher was wrong in this.

Analysis of Variance
It is not clear why Neyman did not include analysis of variance among Fisher’s major accomplishments. Perhaps, as seems possible, it was due to personal enmity. Fisher’s first paper on this subject, with W. A. Mackenzie, was published in 1923. According to Cochran (1980), “two aspects of this paper are of historical interest. At that time, Fisher did not fully understand the rules of analysis of variance—his analysis is wrong—but the role of randomization” (p. 17), but by the time Statistical Methods for Research Workers came out in 1925, he was back on top of his game.
Fisher was the first to discuss Neyman’s 1935 paper regarding analysis of variance in randomized blocks and Latin Square designs, *Statistical Problems in Agricultural Experimentation*, presented to the Royal Statistical society. In this paper, Neyman formulated a model that allowed each treatment to respond differently in each plot, making no assumption that treatment effects were fixed and additive in the plots. As noted by Holschuh (1980), “the null hypothesis he [Neyman] considered was that the average treatment response over the entire experimental area was the same for all treatments. Under this null hypothesis, he found that the z-test for the randomized block design was unbiased” (p. 43) but that the test for the Latin square design was, in general, not unbiased (z is one-half the natural log of the F-statistic). If it is assumed that the correlation of plot errors is unity, the z-test is unbiased.

Fisher (1935) began his comments by writing, “…he [Fisher] had hoped that Dr. Neyman’s paper would be on a subject with which the author was fully acquainted, and on which he could speak with authority…Since seeing the paper, he had come to the conclusion that Dr. Neyman had been somewhat unwise in his choice of topics” (p. 154). Fisher focused primarily on Neyman’s analysis of the z-test for treatment effects. Fisher scolded Neyman for obtaining the wrong result for the Latin square design and said that he may have been “misled by his excessive use of symbolism” (Holdschuh, 1980, p.43).

Fisher, however, had ignored Neyman’s null hypothesis. The null hypothesis Fisher entertained was that in any plot the treatments have the same effect. In that case the correlation of plot errors is unity and Neyman’s conclusion is correct: the z-test is unbiased. In the course of the discussion, Neyman (1935) exposed Fisher’s error, but Fisher then claimed that the z-test was only intended to test the null hypothesis of identical treatment effects. Neyman replied that he was “considering problems which are important from the point of view of agriculture” (p. 173).

Neyman (1935) began his written response sarcastically, writing:

I am grateful to Professor Fisher for a sentence in the third part of his contribution…: ‘I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose…’ The sentence I have quoted applies to its author, Professor Fisher, himself, who not only criticized my paper, but blamed me for a variety of sins of which I am not guilty—all this before apparently taking the trouble to discover what my paper is about and what are the results. According to him: I was unwise in the choice of my topics, I have been speaking of things with which I am not fully acquainted, I deceived myself on so simple a question, I forgot the meaning of the facts, I confuse the questions of estimation and the tests of significance and I am apparently not able to grasp the very simple argument!” (p. 174)

Here, again, Fisher seems to have been wrong.

It is in his book *Design of Experiments* (1935a) that Fisher described a method that all have come to know to be defective, except in special cases, that being Fisher’s Least Significant Difference (LSD) procedure. Fisher wrote (1935a) that if the F test is not significant, “…they will not often need to be considered further,” whereas if the test was significant, he continued,

…the null hypothesis has been falsified, and may therefore be set aside. We shall thereafter proceed to interpret the differences between the varietal yields as due at least in part to the inherent qualities of the varieties, as manifested on the conditions of the test, and shall be concerned to know with what precision these different yields have been evaluated. …In either case the square root of the variance gives the standard deviation, and provides therefore a means of judging which of the differences among our varietal yield values are sufficiently great to be
regarded as well established, and which are to be regarded as probably fortuitous. If the experiment leaves any grounds for practical doubt, values may be compared by the t test… (pp. 64-65)

He implied that these t tests would each be conducted with a Type I error rate of five percent.

Fisher went on in the next paragraph to describe a method introduced to the literature 26 years later by Dunn. He explained that when the test is not significant, and yet the researcher goes on to examine comparisons suggested by the data, much caution should be used. He wrote (1935a),

…for if the variants are numerous, a comparison of the highest with the lowest observed value, picked out from the results, will often appear to be significant, even from undifferentiated material. Properly, such unforeseen effects should be regarded only as suggestions for future experimentation, in which they can be deliberately tested…Thus, in comparing the best with the worst of ten tested varieties, we have chosen the pair with the largest apparent difference out of 45 pairs, which might equally have been chosen. We might, therefore, require the probability of the observed difference to be as small as 1 in 900, instead of 1 in 20, before attaching statistical significance to the contrast.” (p. 66)

Although testing contrasts, even with a Dunn-Bonferroni adjustment, after a non-significant F test inflates the Type I error rate, it is of interest to discuss the LSD procedure. Fisher maintained that significance tests reveal facts. Sometimes these facts are used to falsify hypotheses, and at other times, many such revealed facts can serve as the genesis of a conjecture intended to explain them. Multiple comparison procedures attempt to control family-wise Type I error rates across a number of comparisons, which comprise a family of comparisons in the sense that a false rejection of any one of them would lead a researcher to claim as false a statement at a higher conceptual level, and which one would not like to do in error at a rate higher than the adopted alpha. But the LSD method does just that. If the F test is not significant, the experiment is stopped. If it is significant in error, it holds the error rate at the appropriate level in falsifying the higher-level proposition, and any contrasts examined afterward and found significant erroneously do not contribute to the overall error rate, because it is already wrong at an acceptable rate. If the F is correctly significant, one cannot make an error in declaring the higher-level statement false, and one is thus in fact-generating mode for the next attempt at an improved explanation. So Fisher was right after all.

References


Inferences about the Population Mean: Empirical Likelihood versus Bootstrap-t

Rand R. Wilcox
University of Southern California

The problem of making inferences about the population mean, \( \mu \), is considered. Known theoretical results suggest that a Bartlett corrected empirical likelihood method is preferable to two basic bootstrap techniques: a symmetric two-sided bootstrap-t and an equal-tailed bootstrap-t. However, simulations in this study indicate that, when the sample size is small, these two bootstrap methods are generally better in terms of Type I errors and probability coverage. As the sample size increases, situations are found where the Bartlett corrected empirical likelihood method performs better than the equal-tailed bootstrap-t, but the symmetric bootstrap-t gives the best results. None of the four methods considered are always satisfactory in terms of probability coverage or Type I errors, particularly when dealing with skewed distributions where the expected proportion of points flagged as outliers is somewhat high. If this proportion is 0.14, for example, all four methods can be unsatisfactory even with \( n = 300 \), but if sampling from a symmetric distribution or a skewed distribution with relatively light tails the results suggest using a symmetric two-sided bootstrap-t method.

Key words: Level robust methods, Bartlett correction, bootstrap-t.
As a result, numerous alternative methods have been proposed. One general approach is to use nonparametric techniques, which include empirical likelihood methods (Owen, 2001) as well as bootstrap methods (Efron & Tibshirani, 1993). Asymptotic results suggest that a Bartlett corrected empirical likelihood approach is superior to using a bootstrap-t method (DiCiccio, Hall & Romano, 1991). However, with small to moderate sample sizes, it appears that little or nothing is known regarding how these two approaches compare. Moreover, simulation results on the empirical likelihood technique are limited to a rather narrow range of situations.

This study compared two basic variations of the bootstrap-t method to two variations of the empirical likelihood method. A minor result is that the simulations support extant results that the Bartlett corrected empirical likelihood method is preferable to the basic empirical likelihood technique. A practical issue, however, is whether a Bartlett corrected empirical likelihood method provides better control over the Type I error probability, versus a bootstrap-t method, when dealing with small to moderate sample sizes. Yet another issue is the extent to which a Bartlett corrected empirical likelihood method gives improved results when sampling from a heavy-tailed distribution, particularly when the distribution is also skewed.

With n = 20, none of the methods compared are satisfactory among all of the distributions considered; none of the methods are satisfactory when sampling from a skewed, heavy-tailed distribution with n ≤ 300. With a small sample size, the simulations indicate that the bootstrap-t methods are generally better than the empirical likelihood methods. As the sample size gets large, situations are found where the Bartlett corrected empirical likelihood method performs better than the equal-tailed bootstrap-t, but all indications point to the symmetric bootstrap-t as best for general use.

Let X₁, ..., Xₙ be a random sample from a distribution with mean μ. Note that Rosenblum and van der Laan (2009) described a method for computing a confidence interval for the mean. Their method is based on Hoeffding’s inequality (Hoeffding, 1963), which guarantees probability coverage at least 1 − α if W can be specified such that with probability 1, |Xᵢ| ≤ W. For the special case 1 − α = .95, the resulting 0.95 confidence interval is

\[ (\bar{X} - 2.72W / \sqrt{n}, \bar{X} + 2.72W / \sqrt{n}) \]

A simple way of implementing this approach is to take W to be the maximum of the observed |Xᵢ| values, but a possible concern from a hypothesis testing point of view is that it is too conservative in terms of Type I errors. In the simulations herein, this approach was considered when sampling from various distributions, including a normal distribution, and based on 5,000 replications, the hypothesis \( H₀ : \mu = \mu₀ \), where \( \mu₀ \) is the true population mean, was never rejected with sample sizes n = 20 and n = 200. Consequently, this approach was eliminated from consideration.

Methods for Comparison: Descriptions

Equal-Tailed Bootstrap-t

The idea behind the bootstrap-t method is to use the observed data to approximate the distribution of

\[ T = \frac{\bar{X} - \mu}{s / \sqrt{n}}, \]

where \( \bar{X} \) and s are the usual sample mean and sample standard deviation, respectively. The strategy begins by generating a bootstrap sample of size n; that is, randomly sample with replacement n values from X₁, ..., Xₙ yielding X₁*, ..., Xₙ*. Let \( \bar{X} \) and s be the mean and standard deviation based on this bootstrap sample, and let

\[ T^* = \frac{\bar{X}^* - \bar{X}}{s^* / \sqrt{n}}. \]  \( (1) \)

Repeat this process B times yielding T₁*, ..., T_B* and let \( T_{(1)} ≤ ... ≤ T_{(B)}^* \) be the B bootstrap T* values written in ascending order. Let \( \ell = \alpha B \), rounded to the nearest integer, and \( u = B - \ell \), in which case an estimate of the \( \alpha / 2 \) and 1-
\( \alpha / 2 \) quantiles of the distribution of \( T \) are \( T_{(\ell+1)}^* \) and \( T_{(u)}^* \), respectively. The resulting equal-tailed \( 1 - \alpha \) confidence interval for \( \mu \) is

\[
(\bar{X} - T_{(u)}^* \frac{s}{\sqrt{n}}, \bar{X} - T_{(\ell+1)}^* \frac{s}{\sqrt{n}})
\]

(2)

It might seem that \( T_{(u)}^* \) should be used to compute the upper end of the confidence interval, not the lower end, but it can be shown that this is not the case. Also, \( T_{(\ell+1)}^* \) is negative, which helps explain why \( T_{(\ell+1)}^* s / \sqrt{n} \) is subtracted from \( \bar{X} \).

Symmetric Bootstrap-t

In contrast to the equal-tailed bootstrap-t is the symmetric confidence interval

\[
\bar{X} \pm T_{(c)}^* \frac{s}{\sqrt{n}},
\]

where \( c = (1 - \alpha)B \) rounded to the nearest integer and the absolute value of the right side of (1) is used to define \( T^* \). This symmetric two-sided confidence interval enjoys some theoretical (asymptotic) advantages over the equal-tailed confidence interval (Hall, 1988a, 1988b), but it is known that - for small sample sizes - situations arise where an equal-tailed confidence interval is more satisfactory (Wilcox, 2005).

Empirical Likelihood

The empirical likelihood method can be used to construct a confidence interval for \( \mu \), but for simplicity it is described in terms of testing \( H_0 : \mu = \mu_0 \). Consider distributions \( F_\rho \), \( \rho = (p_1, \ldots, p_n) \), supported on the sample \( X_1, \ldots, X_n \), where \( X_i \) is assigned mass \( p_i \). For a specified value of \( \mu \), the empirical likelihood \( L(\mu) \) is defined to be the maximum value of \( \Pi p_i \) over all such distributions that satisfy \( \sum X_i p_i = \mu \). Because \( \Pi p_i \) attains its overall maximum when \( p_i = 1 / n \), it follows that the empirical likelihood is maximized when \( \mu = \bar{X} \).

The empirical likelihood ratio for testing \( H_0 \) is

\[
W = -2 \log \left\{ \frac{L(\mu_0)}{L(\bar{X})} \right\}.
\]

When the null hypothesis is true, \( W \) has approximately a Chi-squared distribution with 1 degree of freedom. In particular, \( H_0 \) will be rejected at the \( \alpha \) level if \( W \geq c \), where \( c \) is the \( 1 - \alpha \) quantile of a Chi-squared distribution with 1 degree of freedom.

Bartlett Corrected Empirical Likelihood

The Bartlett corrected empirical likelihood method is applied as follows. Let

\[
\hat{\mu}_j = \sum (X_i - \bar{X})^j / n
\]

and

\[
a = \frac{1}{2} \hat{\mu}_4 \hat{\mu}_2^{-2} - \frac{1}{3} \hat{\mu}_3^2 \hat{\mu}_2^{-3};
\]

the null hypothesis is rejected if

\[
W(1 - an^{-1}) \geq c.
\]

Comments on Designing a Simulation Study

Presumably there are situations where sampling is from a relatively light-tailed, symmetric distribution and outliers are relatively rare, but in various situations it is known that the reverse is true. In a review of 440 large-sample psychological studies, Micceri (1989) reported that 97% (35 of 36 studies) “of those distributions exhibiting kurtosis beyond the double exponential (3.00) also showed extreme or exponential asymmetry” (p. 161). Moreover, 72% (36 of 50) of distributions that exhibited skewness greater than two also had tail weights that were heavier than the double exponential.

In a sexual attitude study by Pedersen, Miller, Putcha-Bhagavatula and Yang (2002), skewness and kurtosis, based on 105 participants, was estimated to be 15.9 and 256.3, respectively. In a related study based on 16,288 participants, the ten variables had estimated skewness that ranged between 52.1 and 115.5, and kurtosis that ranged between 3,290 and 13,357. Based on a boxplot, the proportion of points flagged as outliers ranged between 0.12 and 0.39. Consequently, there are some practical reasons for considering heavy-tailed distributions in simulation studies as well as
distributions that have a fairly high degree of skewness.

An important point is that extant simulation studies regarding empirical likelihood methods do not consider a very wide range of distributions. For example, DiCiccio, et al. (1991) considered a Student’s t distribution with 5 degrees of freedom, which has a median proportion of outliers (over many studies) approximately equal to 0.03 based on the boxplot rule in Frigge, Hoaglin and Iglewicz (1989). In addition to a normal distribution, they also considered a Chi-squared distribution with 1 degree of freedom for which the median proportion of outliers is approximately 0.07. Their simulations reveal unsatisfactory control over the probability of a Type I error with \( n = 20 \), but with \( n = 40 \) the Bartlett corrected version was found to perform reasonably well. This study describes situations where it performs poorly with \( n = 300 \).

Results

Simulations were used to study the actual Type I error probability when testing \( H_0 : \mu = \mu_0 \). The distributions used were standard normal, Chi-squared with 1 degree of freedom, Student’s t with 5 degrees of freedom, lognormal, contaminated normal, and three g-and-h distributions. For convenience these distributions are labeled distributions 1-8, respectively.

The family of contaminated (or mixed) normal distributions used is defined as follows. Let \( X \) be a standard normal random variable having the distribution \( \Phi(x) = P(X \leq x) \). Let \( \varepsilon \) be any constant, \( 0 \leq \varepsilon \leq 1 \) and let \( K \) be any positive constant. The contaminated normal distribution is

\[
H(x) = (1 - \varepsilon)\Phi(x) + \varepsilon\Phi(x / K).
\]

Following Tukey (1960), \( K = 10 \) and \( \varepsilon = .1 \) are used resulting in a symmetric, heavy-tailed distribution, with the median proportion of points declared outliers approximately equal to 0.08. The first three distributions were chosen to illustrate how the bootstrap-t compares to the empirical likelihood methods for the same distributions used by DiCiccio, et al. (1991).

The g-and-h distributions (Hoaglin, 1985) arise as follows. If \( Z \) has a standard normal distribution, then

\[
W = \frac{\exp(gZ) - 1}{g}\exp(hZ^2 / 2),
\]

\( g > 0 \), has a g-and-h distribution where \( g \) and \( h \) are parameters that determine the first four moments. When \( g = 0 \),

\[
W = Z\exp(hZ^2 / 2).
\]

The three g-and-h distributions used were \( g = h = 0.2 \) and 0.5, and \( (g, h) = (0.2, 0) \). Table 1 shows the skewness (\( \gamma_1 \)) and kurtosis (\( \gamma_2 \)) for each of the g-and-h distributions considered. When \( g > 0 \) and \( h > 1/k \), \( E(W^k) \) is not defined and the corresponding entry in Table 1 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

<table>
<thead>
<tr>
<th>g</th>
<th>h</th>
<th>( \gamma_1 )</th>
<th>( \gamma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.0</td>
<td>0.61</td>
<td>3.68</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>2.81</td>
<td>155.98</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Some Properties of the g-and-h Distribution

To add perspective, note that the median proportion of outliers generated, when dealing with \( g = h = 0.5 \), is approximately 0.11 when \( n = 100 \), based on the variation of the boxplot rule recommended by Frigge, Hoaglin & Iglewicz (1989). For \( g = h = 0.2 \) it is 0.05 and for \( (g, h) = (0.2, 0) \) it is 0.01. For a Chi-squared distribution with 1 degree of freedom, \( \epsilon_5 \), the lognormal and the contaminated normal, the median proportion of outliers is approximately 0.07, 0.03, 0.08 and 0.08, respectively. (These results are based on simulations with 5,000 replications.)

Table 2 shows the estimated Type I error probabilities. First consider \( n = 20 \), and note that
the Bartlett corrected empirical likelihood method always improves on the uncorrected approach. Both bootstrap methods have estimated Type I error probabilities less than the estimates using the empirical likelihood methods. Although the seriousness of a Type I error depends on the situation, Bradley (1978) has suggested that generally, at a minimum, the actual Type I error probability should be between 0.025 and 0.075. Based on this criterion, none of the methods are satisfactory. However, for skewed distributions for which the median proportion of outliers does not exceed 0.05, the symmetric bootstrap method gives satisfactory results.

The symmetric bootstrap method can be too conservative when sampling from a symmetric heavy-tailed distribution, but this might be judged to be less serious than having an actual Type I error greater than 0.075, as is the case when using the empirical likelihood methods. Note that with n = 20, the symmetric bootstrap method has a Type I error probability of 0.08 when sampling from a Chi-squared distribution with 1 degree of freedom.

Increasing the sample size to n = 25, the estimate drops to 0.065, and for n = 30 it is 0.059.

For n = 50, the empirical likelihood methods compete better with the bootstrap-t methods, but the symmetric bootstrap-t performs well in situations where the empirical likelihood methods are unsatisfactory based on Bradley’s criterion. Again, a criticism of the symmetric bootstrap-t is that for a symmetric heavy-tailed distribution (the contaminated normal), the Type I error probability drops below 0.025, but the other three methods have estimates greater than 0.12. Thus, for general use, the symmetric bootstrap-t seems best.

Additional simulations were conducted with n = 100 and it was found that the empirical likelihood methods continue to perform poorly when sampling from the heavy-tailed distributions considered here. With n = 200 they perform well when sampling from the contaminated normal but estimates exceed 0.15 when sampling from the g-and-h distribution with g = h = 0.5.

<table>
<thead>
<tr>
<th>n</th>
<th>Distribution</th>
<th>Empirical Likelihood (EL)</th>
<th>Bartlett Corrected Empirical Likelihood (BCEL)</th>
<th>Bootstrap-t, Equal-Tailed (BEQ)</th>
<th>Bootstrap-t, Symmetric (BSYM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1</td>
<td>0.074</td>
<td>0.064</td>
<td>0.058</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.117</td>
<td>0.103</td>
<td>0.068</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.075</td>
<td>0.059</td>
<td>0.067</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.137</td>
<td>0.120</td>
<td>0.099</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.169</td>
<td>0.138</td>
<td>0.116</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.090</td>
<td>0.072</td>
<td>0.083</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.094</td>
<td>0.080</td>
<td>0.083</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.270</td>
<td>0.241</td>
<td>0.231</td>
<td>0.186</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.052</td>
<td>0.050</td>
<td>0.055</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.074</td>
<td>0.069</td>
<td>0.055</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.062</td>
<td>0.058</td>
<td>0.072</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.068</td>
<td>0.062</td>
<td>0.058</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.137</td>
<td>0.125</td>
<td>0.145</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.061</td>
<td>0.057</td>
<td>0.073</td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.074</td>
<td>0.066</td>
<td>0.080</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.215</td>
<td>0.203</td>
<td>0.207</td>
<td>0.194</td>
</tr>
</tbody>
</table>
In terms of controlling the probability of a Type I error, the most difficult situation seems to occur when sampling from an asymmetric distribution with heavy-tails. Even using \( n = 300 \) none of the methods considered are satisfactory. In particular, for the g-and-h distribution with \( g = h = 0.5 \), all four methods estimated Type I error probabilities exceeding 0.14. One of the main points is that - for symmetric distributions with heavy tails - the symmetric bootstrap-t avoids Type I errors well above the nominal level even with \( n = 20 \) (albeit with small sample sizes the actual level can drop below 0.025). By contrast, the Bartlett corrected empirical likelihood method has an actual level of approximately 0.09 with \( n = 100 \), and with \( n = 200 \) the level drops to 0.063. Consequently, it seems that the symmetric bootstrap-t is best for general use. Except for skewed heavy-tailed distributions, it performs reasonably well with \( n \geq 50 \).

References

Type I error and power of the standard independent samples t-test were compared with the trimmed and Winsorized t-test with respect to continuous distributions and various discrete distributions known to occur in applied data. The continuous and discrete distributions were generated with similar levels of skew and kurtosis but the discrete distributions had a variety of structural features not reflected in the continuous distributions. The results showed that the Type I error rates of the t-tests were not seriously affected, but the power rate of the trimmed and Winsorized t-test varied greatly across the considered distributions.

Key words: Nonnormality, independent samples t-test, trimming, Winsorizing.

Introduction
Monte Carlo simulation studies are commonly used to assess the performance of statistical strategies under defined and controlled conditions. Often the question of interest involves the performance of one or more strategies under violations of the assumptions associated with the mathematical model on which a procedure is based. While simulation studies are informative, their conditions and results may be generated in ways that are not relevant for applied research settings. Of particular concern is the accuracy of simulation studies’ recommendations about the impact of assumption violations in continuous and unbounded distributions for applied distributions that are primarily discrete and bounded.

A number of traditional statistical procedures assume a normal distribution for the underlying population from which scores were drawn (e.g., t-test, ANOVA). In simulation studies that evaluate the robustness of statistical significance tests of mean differences, nonnormality is usually created in smooth, continuous and theoretically unbounded distributions. Several methods exist for transforming normally distributed random numbers into nonnormal distributions, including Hoaglin’s (1985) g and h method, Fleishman’s (1978) power method, and the use of Chi-square distributions with varying degrees of freedom. The nonnormality generated with these methods can primarily be defined in terms of skew and kurtosis. In contrast to simulated data, applied distributions of psychometric tests and achievement tests are usually discrete with bounded score ranges and are noted to have features such as lumps, bimodalities, or popular, unpopular or impossible scores (Holland & Thayer, 2000; Micceri, 1989). While these discrete distributions can be described in terms of their skew and kurtosis, a complete description would require more attention to their structural features. Continuous and discrete
distributions with similar skew and kurtosis can reflect very different shapes.

Simulation studies that have evaluated significance tests of mean differences for nonnormal continuous distributions have produced different recommendations than simulation studies that consider nonnormal discrete distributions. Studies based on nonnormal continuous distributions have recommended that standard tests of mean differences be abandoned in favor of robust tests of trimmed mean differences (Keselman, Othman, Wilcox & Fradette, 2004; Lix & Keselman, 1998). In contrast, Sawilowsky and Blair (1992) used a variety of discrete distributions as population distributions and found that the standard t-test’s Type I error rate was relatively unaffected by their populations.

The interest of this study is to investigate how the data generation method and population distributions used in a simulation study influence the results and recommendations of statistical strategies. Data were generated from the continuous distributions commonly considered in simulation studies and from various discrete and bounded distributions noted to occur in applied data (Holland & Thayer, 2000; Micceri, 1989; Sawilowsky, & Blair, 1992). The continuous and discrete distributions were generated with similar levels of skew and kurtosis but the discrete distributions had structural features not reflected in the continuous distributions.

Type I error and power were assessed in the standard independent samples t-test and one of its most recommended alternatives for nonnormal data, Yuen’s (1974) trimmed and Winsorized t-test (Keselman, et al., 2004). In addition, this article considers the relevance of simulation studies’ recommendations of statistical strategies for applied data.

Methodology
The objective of this study was to compare the Type I error and power rates for the standard t-test and the trimmed and Winsorized t-test when used to compare means in discrete distributions noted to occur in applied data and in continuous distributions of equal skew and kurtosis typically considered in simulation studies. The Type I error and power rates were computed from 10,000 replications where in each replication two random samples of size 30 were drawn from one of nine population distributions and the groups’ means were compared using the standard t-test and the trimmed and Winsorized t-test. The nine population distributions included one continuous distribution and three discrete distributions of symmetric shape and one continuous distribution and four discrete distributions of asymmetric shape.

Population Distributions
The population distributions reflected two basic shapes, asymmetric and symmetric. The two shapes were modeled with bounded and discrete distributions and one accompanying continuous distribution. The asymmetric shape is skewed (approximately −1.75) and leptokurtic (kurtosis approximately 3.75). The asymmetric continuous and unbounded population distribution is shown in Figure 1. One of the asymmetric discrete distributions is smooth (Figure 2), and the others have structures such as teeth (Figure 3), a lump at score zero (Figure 4) and favorite scores (Figure 5). The means, standard deviations, skews and kurtosis of these five distributions are summarized in Table 1.

The symmetric distributions included three discrete and bounded distributions and one continuous and unbounded distribution (Table 2, Figures 6-9). All four symmetric distributions have skews of 0. The symmetric continuous distribution is shown in Figure 6. One of the symmetric discrete distributions is smooth (Figure 7); the others have peaks (Figure 8) and bimodality (Figure 9).

Data Generation Methods
The first data generation method produced data (i.e., Y scores for two groups) that reflected the discreteness and shapes of the discrete distributions where only the integer scores in defined score ranges were possible and where each possible score had a corresponding population probability (Figures 2-5 & 7-9). Samples of 30 scores were randomly drawn from these population distributions with the scores’ population probabilities defining the probabilities of those scores appearing in the sample datasets.
Table 1: Summary Statistics for Four Negatively Skewed Discrete Distributions and One Continuous Distribution

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>15.00</td>
<td>4.00</td>
<td>-1.75</td>
<td>3.75</td>
</tr>
<tr>
<td>Smooth &amp; Discrete</td>
<td>15.73</td>
<td>2.90</td>
<td>-1.85</td>
<td>3.88</td>
</tr>
<tr>
<td>Teeth</td>
<td>14.46</td>
<td>3.45</td>
<td>-1.81</td>
<td>3.94</td>
</tr>
<tr>
<td>Lump at Zero</td>
<td>12.08</td>
<td>3.79</td>
<td>-1.97</td>
<td>3.85</td>
</tr>
<tr>
<td>Favorite Scores</td>
<td>17.36</td>
<td>4.13</td>
<td>-1.92</td>
<td>3.73</td>
</tr>
</tbody>
</table>

Table 2: Summary Statistics for Three Symmetric Discrete Distributions and One Continuous Distribution

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>15.00</td>
<td>4.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Smooth &amp; Discrete</td>
<td>15.00</td>
<td>4.00</td>
<td>0.00</td>
<td>-0.15</td>
</tr>
<tr>
<td>7 Peaks</td>
<td>10.50</td>
<td>4.88</td>
<td>0.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Bimodal</td>
<td>15.00</td>
<td>6.42</td>
<td>0.00</td>
<td>-1.18</td>
</tr>
</tbody>
</table>

Figure 1: Asymmetric Continuous Distribution

These probabilities are not exactly zero

Figure 2: Asymmetric Smooth & Discrete Distribution
Figure 3: Asymmetric Teeth Distribution

Figure 4: Asymmetric Lump at Zero Distribution

Figure 5: Asymmetric Favorite Scores Distribution
Figure 6: Symmetric Continuous Distribution

Figure 7: Symmetric Smooth & Discrete Distribution

Figure 8: Symmetric 7 Peaks Distribution

These probabilities are not exactly zero

These probabilities are not exactly zero
The second data generation method was a continuous data generation method. The continuous data generation method used in this study is known as Fleishman’s (1978) power method. Sample datasets of 30 standard normal deviates ($Z$) were generated and these normal deviates were transformed into samples from the desired population distributions,

$$Y = \mu + \sigma (a + bZ + cZ^2 + dZ^3).$$  \hspace{1cm} (1)

Sets of $\mu$, $\sigma$, $a$, $b$, $c$, and $d$ values were used to produce $Y$ values that had means, standard deviations, skews and kurtoses that reflected the symmetric and asymmetric discrete distributions.

For the Asymmetric Continuous distribution (Figure 1), $\mu$ and $\sigma$ were 15 and 4, respectively, and constants of $a$, $b$, $c$, and $d$ values of 0.3995, 0.9297, $\text{−}0.3995$ and $\text{−}0.0365$ were used to achieve the asymmetry and non-normality (skew = $\text{−}1.75$; kurtosis =3.75). For the Symmetric Continuous distribution (Figure 6), $\mu$ and $\sigma$ were 15 and 4, and $a$, $b$, $c$, and $d$ values of 0, 1, 0 and 0 were used to achieve the symmetry and normality (skew = 0; kurtosis = 0).

Statistical Strategies for Testing Mean Differences

Two statistical tests were considered for evaluating the mean differences in $Y$ for groups $j = 1$ and 2. The standard independent samples $t$-test assuming homogeneous variances is defined as,

$$t_{\text{Standard}} = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}},$$  \hspace{1cm} (2)

where $\overline{Y}_1$ and $\overline{Y}_2$ are the groups’ sample means,

$$\frac{1}{n_j} \sum_{i} Y_{i,j},$$  \hspace{1cm} (3)

$s_1^2$ and $s_2^2$ are the groups’ sample variances,

$$\frac{1}{n_j - 1} \sum_{i} (Y_{i,j} - \overline{Y}_j)^2.$$  \hspace{1cm} (4)

used to compute the pooled variance, $s^2$,

$$\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$  \hspace{1cm} (5)

The statistical significance of $t_{\text{Standard}}$ is determined by computing its percentile on a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Yuen’s (1974) trimmed and Winsorized t-test was also considered. First the $Y$ scores are ordered within each treatment group,
where \( g_j = \gamma n_j \) is then defined as the proportion of individuals trimmed in each tail of the distribution (\( \gamma = 0.1 \) & 0.2 in this study) and the effective sample size for group \( j \) is \( h_j = n_j - 2g_j \). The trimmed mean for group \( j \) is computed as,

\[
\bar{Y}_{t,j} = \frac{1}{n_j} \sum_{i \geq g_j+1} Y_{i,j}.
\]

(6)

The data for group \( j \) are Winsorized as,

\[
X_{ij} = \begin{cases} 
Y_g & \text{if } Y_g \leq Y_{g+j,1}, \\
Y_g & \text{if } Y_j < Y_{g+j,1} \leq Y_g < Y_{n_j-g_j,1}, \\
Y_{n_j-g_j,1} & \text{if } Y_j \geq Y_{n_j-g_j,1} 
\end{cases}
\]

(7)

and the Winsorized data are used to compute group \( j \)'s Winsorized mean,

\[
\bar{X}_{w,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{i,j},
\]

(8)

and variance,

\[
s^2_{w,j} = \frac{1}{n_j-1} \sum_{i=1}^{n_j} (X_{i,j} - \bar{X}_{w,j})^2.
\]

(9)

Finally, the t-test for comparing groups' trimmed means is computed as,

\[
t_{\text{trimWinsorized}} = \frac{\bar{Y}_{t,1} - \bar{Y}_{t,2}}{\sqrt{d_1 + d_2}},
\]

(10)

where

\[
d_j = \frac{(n_j - 1)s^2_{w,j}}{h_j(h_j - 1)}.
\]

The statistical significance of the \( t_{\text{trimWinsorized}} \) statistic is determined by computing its percentile on a t distribution with

\[
\frac{(d_1 + d_2)^2}{d_1/(h_1 - 1) + d_2/(h_2 - 1)}
\]

degrees of freedom.

Both the standard and the trimmed and Winsorized t-tests were implemented as two-tailed significance tests with nominal Type I error rates of 0.05. The trimmed and Winsorized t-test was based on symmetric trimming and Winsorizing of 10% and 20% of the most extreme lowest and highest observations of the two groups’ \( Y \) distributions.

Type I Error and Power Evaluations

The standard and trimmed and Winsorized t-tests were used to evaluate the statistical significance of the differences in means of two groups whose scores were generated as samples from one of the nine population distributions. The t-tests were evaluated with respect to their Type I error (where the population difference in groups’ means was zero) and power (where the population difference in groups’ means was not zero).

All t-tests’ Type I error and power rates were rates at which the t-tests indicated that the groups’ mean differences were statistically significant across 10,000 replications (i.e., 10,000 statistical significance tests of groups’ mean differences). The t-tests’ Type I error rates were computed in conditions where the sample datasets for the two groups were drawn from one population distribution and were not altered prior to their analyses with the t-tests. The robustness of the t-tests’ Type I error rates were considered with respect to two criteria, the Type I error range defined as ±2 standard errors of the nominal 0.05 rate for a simulation study based on 10,000 replications (i.e., 0.05 ± 2 × standard errors of the nominal 0.05 rate for a simulation study based on 10,000 replications = 0.0456 to 0.0544), and a wider robustness criterion proposed by Bradley (0.025 to 0.075, 1978). The t-tests’ power rates were computed in the simulated conditions where the sample datasets for the two groups were drawn from one population distribution and then 1/2 of the population
distribution’s standard deviation was added to one of the groups’ scores.

Results

Type I Error

Table 3 presents the t-tests’ Type I error rates across this study’s nine population distributions. Comparisons of the standard and trimmed and Winsorized t-tests for the two continuous distributions pertain to the t-test evaluations of interest in most simulation studies. Comparisons of the t-tests across the discrete distributions are unconsidered in most simulation studies.

The Type I error rates of the three t-tests across all population distributions were within the 0.025 to 0.075 range defined by Bradley’s (1978) criterion, but several fell outside of the ±2 standard error range (0.0456 to 0.0544). The nonrobust Type I error rates were conservative (less than 0.05) rather than the liberal (greater than 0.05) Type I error rates that would prompt the greatest concern of the t-tests’ robustness. The trimmed and Winsorized t-test had more nonrobust, conservative Type I error rates than the standard t-test across the continuous and discrete distributions.

The extent of trimming had distribution-dependent influences on Type I error, where 20% trimming versus 10% trimming reduced Type I error for some distributions (i.e., the Asymmetric Continuous, Asymmetric Smooth & Discrete, and the Symmetric 7 Peaks distributions) and increased Type I error for other distributions (i.e., the Asymmetric Favorite Scores, Asymmetric Lump at Zero, Asymmetric Teeth, Symmetric Continuous, Symmetric Smooth & Discrete and the Symmetric Bimodal distributions).

Power

Table 4 presents the t-tests’ power rates across this study’s nine population distributions. The t-tests’ power rates were most clearly affected by whether the distributions were symmetric or asymmetric. For the asymmetric distributions, the trimmed and Winsorized t-test was more powerful than the standard t-test. The greater power of the trimmed and Winsorized t-test held across the asymmetric continuous and asymmetric discrete distributions, and was especially apparent in the Asymmetric Teeth and Asymmetric Lump at Zero distributions. For the Asymmetric Teeth and Asymmetric Lump at Zero distributions, 20% trimming resulted in increased power relative to 10% trimming. For most of the symmetric distributions, the trimmed and Winsorized t-test was less powerful than the standard t-test. For all but the Symmetric 7 Peaks distribution, 20% trimming reduced power relative to 10% trimming.

Conclusion

In simulation research considerable attention has been devoted to the effects of nonnormality on the accuracy of statistical significance tests for groups’ mean differences (Glass, Peckham & Saunders, 1972; Keselman, et al., 2004; Lix, & Keselman, 1998; Lix, Keselman & Keselman, 1996). In this research nonnormality is predominantly characterized in terms of the level of skew and kurtosis of continuous and theoretically unbounded distributions.

Recent results and proposals from simulation research have suggested that standard significance tests should be abandoned in favor of alternative significance tests that are designed to be robust to nonnormality (Lix, Keselman & Keselman, 1996; Wilcox, 1995). However, a somewhat unique simulation study found that the standard t-test can be quite robust with respect to the types of nonnormality noted to occur in real world distributions of psychometric and achievement tests, where score ranges are discrete and bounded and where nonnormality cannot be completely characterized with respect to skew and kurtosis (Sawilowsky & Blair, 1992). This study was designed to reconsider the Type I error and power of standard and trimmed and Winsorized t-tests of mean differences with respect to the types of distributions considered in the majority of simulation studies and the types of distributions noted to occur in applied psychometric and achievement test data.

In terms of Type I error, the results show that the standard and trimmed and Winsorized t-tests did not exhibit extreme lack of robustness for any of the considered distributions. Type I error rates obtained for the continuous distributions considered in simulation studies were reasonably representative of the Type I error rates obtained
### Table 3: Type I Error Results

<table>
<thead>
<tr>
<th>Symmetry</th>
<th>Distribution</th>
<th>Standard t-test</th>
<th>Trimmed &amp; Winsorized t-test (10% trimming)</th>
<th>Trimmed &amp; Winsorized t-test (20% trimming)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetric</td>
<td>Continuous</td>
<td>0.0424*</td>
<td>0.0431*</td>
<td>0.0393*</td>
</tr>
<tr>
<td></td>
<td>Favorite Scores</td>
<td>0.0454*</td>
<td>0.0360*</td>
<td>0.0502</td>
</tr>
<tr>
<td></td>
<td>Lump at Zero</td>
<td>0.0476</td>
<td>0.0333*</td>
<td>0.0460</td>
</tr>
<tr>
<td></td>
<td>Smooth &amp; Discrete</td>
<td>0.0471</td>
<td>0.0435*</td>
<td>0.0431*</td>
</tr>
<tr>
<td></td>
<td>Teeth</td>
<td>0.0473</td>
<td>0.0364*</td>
<td>0.0455*</td>
</tr>
<tr>
<td>Symmetric</td>
<td>Continuous</td>
<td>0.0447*</td>
<td>0.0450*</td>
<td>0.0452*</td>
</tr>
<tr>
<td></td>
<td>7 Peaks</td>
<td>0.0493</td>
<td>0.0451*</td>
<td>0.0379*</td>
</tr>
<tr>
<td></td>
<td>Smooth &amp; Discrete</td>
<td>0.0494</td>
<td>0.0469</td>
<td>0.0498</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>0.0478</td>
<td>0.0477</td>
<td>0.0495</td>
</tr>
</tbody>
</table>

*The Type I error rate is outside of the +/- 2 standard error range (0.0456 to 0.0544)*

### Table 4: Power Results

<table>
<thead>
<tr>
<th>Symmetry</th>
<th>Distribution</th>
<th>Standard t-test</th>
<th>Trimmed &amp; Winsorized t-test (10% trimming)</th>
<th>Trimmed &amp; Winsorized t-test (20% trimming)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetric</td>
<td>Continuous</td>
<td>0.4910</td>
<td>0.5241</td>
<td>0.5135</td>
</tr>
<tr>
<td></td>
<td>Favorite Scores</td>
<td>0.5001</td>
<td>0.6144</td>
<td>0.5012</td>
</tr>
<tr>
<td></td>
<td>Lump at Zero</td>
<td>0.4980</td>
<td>0.6698</td>
<td>0.7437</td>
</tr>
<tr>
<td></td>
<td>Smooth &amp; Discrete</td>
<td>0.5014</td>
<td>0.5352</td>
<td>0.5254</td>
</tr>
<tr>
<td></td>
<td>Teeth</td>
<td>0.5030</td>
<td>0.6511</td>
<td>0.7543</td>
</tr>
<tr>
<td>Symmetric</td>
<td>Continuous</td>
<td>0.4810</td>
<td>0.4527</td>
<td>0.4213</td>
</tr>
<tr>
<td></td>
<td>7 Peaks</td>
<td>0.4756</td>
<td>0.4590</td>
<td>0.5849</td>
</tr>
<tr>
<td></td>
<td>Smooth &amp; Discrete</td>
<td>0.4813</td>
<td>0.4391</td>
<td>0.4104</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>0.4746</td>
<td>0.3805</td>
<td>0.2813</td>
</tr>
</tbody>
</table>
from different types of discrete distributions. The Type I error rates of the t-tests were more likely to be slightly conservative rather than liberal. The trimmed and Winsorized t-test had a Type I error that was usually more conservative than that of the standard t-test.

This study’s power results were more extreme than the Type I error results, and varied by the type of t-test, by whether the population distribution was symmetric or asymmetric, and by the specific features of the population distribution. To assess the power results in more detail, this study’s power simulations were re-run and analyzed with respect to issues such as the expected mean differences in the samples, the standard error of the mean differences in the samples, and the accuracy of the estimated standard error of the mean differences. To simplify the analyses, all of the simulated data were transformed so that all population standard deviations were four, all population mean differences were two and the standard errors of these population untrimmed mean differences were about 1.03 (given the group sample sizes of 30). The score transformations had negligible effects on the power rates reported in Table 4 and no effect on the discreteness and structures of the distributions.

The results of the re-run power analyses are presented in Table 5, where the 27 power rates corresponding to the nine population distributions and three t-tests are sorted from highest to lowest. Along with the power rates, the standard errors of the mean differences are shown (i.e., the standard deviation of the differences in the means evaluated by the t-tests across the 10,000 replications of the simulations). These 27 standard errors correlated −0.97 with the 27 power rates and provide a useful basis for understanding how power was affected by the population distributions and t-tests considered in this study. The major power results can be described as follows:

- Power was highest for the distributions and t-tests where the standard error of mean differences was lowest. Power was lowest for the distributions and t-tests where the standard error of mean differences was highest.
- The trimmed and Winsorized t-test had high power and a low standard error when used with all of the asymmetric distributions. The trimmed and Winsorized t-test had low power and a high standard error when used with all of the symmetric distributions except for the Symmetric 7 Peaks distribution.
- The extent of trimming had mixed results, in that for some distributions increased trimming resulted in increased power and decreased standard errors while for other distributions increased trimming resulted in decreased power and increased standard errors.
- The issue of continuous and discrete distributions had an influence on the power of the trimmed and Winsorized t-test such that power rates were less extreme for the continuous distributions of comparable levels of skew. That is, the power for the Asymmetric Continuous distribution was lower than the power for the asymmetric discrete distributions while the power for the Symmetric Continuous distribution was greater than the power for the symmetric discrete distributions.
- The standard t-test’s power and standard errors were less influenced than the trimmed and Winsorized t-test across the distributions, being less powerful than the trimmed and Winsorized t-test for the asymmetric distributions and more powerful than the trimmed and Winsorized t-test for the symmetric distributions. In contrast to the trimmed and Winsorized t-test, the standard t-test was slightly less powerful for the symmetric distributions than for the asymmetric distributions.

Implications for Practice

This study’s findings regarding how a data generation method affects the relative power of different t-tests have implications for practice. The trimmed and Winsorized t-test is more complexly affected by the type of distribution than the standard t-test. Some of the power issues with the trimmed and Winsorized t-test could be anticipated with careful examination of the data at hand. Specifically, for
datasets that have structures and asymmetry resulting in only a small number of the possible scores being observed (i.e., the Asymmetric Teeth and Asymmetric Lump at Zero distributions), trimming and Winsorizing of these observed scores will produce a dataset with even fewer unique scores, a standard error of the trimmed mean that is relatively small, and a power rate that may be large relative to the standard t-test.

For datasets where many of the possible scores are observed (i.e., the Symmetric Bimodal distribution), trimming and Winsorizing of the observed scores will produce a dataset with a large range of unique scores, a standard error of the trimmed mean that is relatively large, and a power rate that is small relative to the standard t-test. If the data at hand are skewed and/or are based on a sample size that is extremely small, trimming and Winsorizing could remove all of the scores from the data and make a significance test of mean differences impossible.
Note that this study focused on creating distributions that reflect structures that have been observed in psychometric and achievement test data (Holland & Thayer, 2000; Micceri, 1989). While the discrete distributions considered in this study may be more realistic than the continuous distributions typically created in simulation studies, these discrete distributions clearly do not reflect all of the possible distributions encountered in applied data.

Important distributions that were not considered in this study are distributions of counted variables, such as individuals’ income, individuals’ total of social connections to other individuals, or websites’ numbers of hits. Extreme observations are more likely in distributions of unbounded counted variables than in distributions of psychometric and achievement test scores. Simulations based on distributions where extreme observations are likely may show that the standard t-test has a nonrobust Type I error rate whereas the trimmed and Winsorized t-test is robust.

Implications for Simulation Research

This study’s findings of how the data generation method affected the relative Type I error and power rates of different t-tests have implications for simulation research. One issue that could be reconsidered is how assumptions are violated in simulation studies. For example, in simulation studies’ continuous and unbounded distributions, the levels of skew and kurtosis can be much greater than are possible to create in discrete and bounded distributions, such as skew values of 120 and kurtosis values ranging from 8.9 to beyond 18,000 (Keselman, et al., 2004; Wilcox, 1994). The current study suggests that simulation studies’ results based on extreme levels of assumption violations do not always generalize to situations where levels of assumption violations are more limited. In particular, this study suggests that for the relatively limited levels of assumption violations that can occur in bounded distributions, the robustness of standard tests of mean differences is not likely to be as serious a concern as implied when robust strategies are promoted. This study also showed that more can be learned about robust statistical procedures proposed as replacements for standard statistical tests. Additional simulation studies that consider the distributions and assumption violations likely to be encountered in applied research are encouraged.

References


The Small-Sample Efficiency of Some Recently Proposed Multivariate Measures of Location

Marie Ng  Rand R. Wilcox
University of Hong Kong  University of Southern California

Numerous multivariate robust measures of location have been proposed and many have been found to be unsatisfactory in terms of their small-sample efficiency. Several new measures of location have recently been derived, however, nothing is known about their small-sample efficiency or how they compare to the sample mean under normality. This research compared the efficiency for \( p = 2, 5, \) and \( 8 \) with sample sizes \( n = 20 \) and \( 50 \) for \( p \)-variate data. Although previous studies indicate that so-called skipped estimators are efficient, this study found that variations of this approach can perform poorly when \( n \) is small and \( p \) exceeds 5. One of the best estimators was found to be a skipped estimator where outliers detected by a projection method are eliminated. The TBS, OGK and RMBA estimators were included and; in some cases, they performed well, however, serious exceptions were identified suggesting that a skipped estimator based on a projection-type outlier detection method is preferable based on efficiency.

Key words: Robust methods, OGK estimator, TBS estimator, median ball algorithm, minimum generalized variance technique, projection methods, skipped estimators of location.

Introduction

A fundamental goal of this research is estimating some appropriate measure of location based on a random sample from some \( p \)-variate distribution. From basic principles, the sample mean has various optimal properties under normality; however, slight departures from normality can render it highly atypical and relatively inefficient. This has led to a variety of robust estimators, many of which are known to have relatively poor small-sample efficiency (Masse & Plante, 2003); thus study expands on Masse and Plante in several ways. First, recently proposed estimators are considered, next so-called skipped estimators are included, and lastly the present study is not limited to the bivariate case. In particular, the small-sample efficiency of the OGK estimator proposed by Maronna and Zamar (2002), the TBS (translated biweight) estimator derived by Rocke (1996) and the RMBA (median ball algorithm) suggested by Olive (2004, 2007) are examined. Skipped estimators simply mean that some appropriate multivariate outlier detection method is applied, any outliers found are removed and the mean of the remaining values is used as a measure of location.

This study considered two types of outlier detection methods. The first is based on a robust analog of Mahalanobis distance where the usual mean and covariance matrix are replaced by some robust measure of location and scale, respectively; in this case, the OGK, TBS and RMBA are considered. The second type does not use the Mahalanobis distance. One of the alternative strategies is based on a particular set of data projections in which a point is declared an outlier if it is flagged as an outlier by any projection. The other method, called the MGV method, belongs to this second class of techniques and assigns a measure of depth to points based in part on generalized variances of subsets of the data.

Marie Ng is an Assistant Professor in the Faculty of Education. Email: marieng@uw.edu. Rand R. Wilcox is a Professor of Psychology. Email: rwilcox@usc.edu.
location estimators. Some basic concerns and results about multivariate outlier detection techniques are reviewed, and a description of the methods used in this research is provided. (See Wilcox (2008) for a more detailed comparison of the outlier detection methods.)

When choosing a multivariate outlier detection technique method at least two fundamental properties are of interest. The first is the outside rate per observation, which is the expected proportion of outliers among a sample of size \(n\), for example, \(p_n\). When sampling from a multivariate normal distribution, it is generally desirable to have a reasonably small \(n p\), for example 0.05; often methods are tuned to achieve this goal, at least when \(n\) is large (Rousseeuw & van Zomeren, 1990).

A second fundamental goal is to avoid masking. Roughly, a method is said to suffer from masking if the very presence of outliers causes them to be missed. Let \(M\) be some multivariate measure of location based on data randomly sampled from some \(p\)-variate distribution and let \(C\) be some measure of scatter. If \(M\) is the usual sample mean and \(C\) the usual covariance matrix based on \(X_1, \ldots, X_n\), then a classic approach is to use the Mahalanobis distance

\[
D_i = \sqrt{(X_i - M)C^{-1}(X_i - M)}
\]

and declare \(X_i\) an outlier if \(D_i\) is sufficiently large. In particular, if the goal is to have \(p_n = \alpha\), then \(X_i\) is declared an outlier if

\[
D_i \geq \sqrt{X_i^2 - \alpha/2} ,
\]

the square root of the \(1 - \alpha/2\) quantile of a Chi-squared distribution with \(p\) degrees of freedom. It is known, however, that this method suffers from masking (Rousseeuw & Leroy, 1987), roughly because the usual sample mean and covariance matrix are not robust, that is, outliers can greatly influence their values thus causing \(D_i\) to be small even when \(X_i\) is highly atypical.

A seemingly natural approach to avoid masking is to take \(M\) and \(C\) to be some robust measure of location and scatter in equation (1) and then use equation (2). Campbell (1980) proposed using a particular M-estimator. The M-estimator Campbell used has a rather unsatisfactory breakdown point, however; the breakdown point of an estimator is the smallest proportion of points that must be altered to make it arbitrarily large or small. The M-estimator has a breakdown point of only \(1/(p+1)\): this means that masking can be a problem - particularly as \(p\) gets large. Consequently, Rousseeuw and van Zomeren (1990) suggested using the minimum volume ellipsoid (MVE) estimator introduced by Rousseeuw (1985) and discussed in detail by Rousseeuw and Leroy (1987).

It appears that this method performs well in terms of achieving \(p_n \approx 0.05\) (Wilcox, 2005); however, serious concerns have been expressed by Olive (2004) and Hawkins and Olive (2002). In addition, Fung (1993) described conditions where MVE can declare too many points outliers. Rousseeuw and van Driessen (1999) suggested replacing the MVE estimator with the fast minimum covariance determinant (FMCD) estimator, but with small to moderate sample sizes \(p_n\) becomes unstable and might exceed 0.05 by an unacceptable amount (Wilcox, 2005). At least three alternatives to the MVE and FMCD estimators exist and might be used instead.

The OGK Estimator

In its general form, the orthogonal Gnanadesikan-Kettenring (OGK) estimator, derived by Maronna and Zamar (2002), is applied as follows. Let \(\sigma(X)\) and \(\mu(X)\) be any measures of dispersion and location, respectively. The method proposed by Gnanadesikan and Kettenring (1972) begins with the robust covariance between any two variables, for example \(X\) and \(Y\), is:

\[
\text{cov}(X, Y) = \frac{1}{4}(\sigma(X + Y)^2 - \sigma(X - Y)^2)
\]

(3)
When \( \sigma(X) \) and \( \mu(X) \) are the usual standard deviation and mean, the usual covariance between \( X \) and \( Y \) results. Following Maronna and Zamar (2002), \( \sigma(X) \) is taken to be the tau scale of Yohai and Zamar (1988). Let

\[
W_c(x) = \left(1 - \left(\frac{x}{c}\right)^2\right)^2 I(|x| \leq c)
\]

and

\[
\rho_c(x) = \min(x^2, c^2),
\]

where the indicator function \( I(|x| \leq c) = 1 \) if \( |x| \leq c \) and 0 otherwise. For the univariate sample \( X_1, \ldots, X_n \), let \( \text{MAD}(X) \) be the median of \( |X_1 - M_x|, \ldots, |X_n - M_x| \), where \( M_x \) is the usual median of \( X_1, \ldots, X_n \), and let

\[
w_i = W_{4.5} \left(\frac{X_i - M_x}{\text{MAD}(X)}\right).
\]

Again, following Maronna and Zamar (2002) the location and scale statistics are defined as

\[
\mu(X) = \frac{\sum w_i X_i}{\sum w_i}
\]

and

\[
\sigma(X)^2 = \frac{\text{MAD}(X)}{n} \sum \rho_3 \left(\frac{X_i - \mu(X)}{\text{MAD}(X)}\right)
\]

Using the measure of scale in (3), the resulting measure of covariance will be denoted by \( \nu(X, Y) \).

Following the notation in Maronna and Zamar (2002), let \( x_i \) be the \( i \)th row of the \( n \times p \) matrix \( X \), a scatter matrix \( V(X) \) and a location vector \( t(X) \) are defined as follows:

1. Let \( D = \text{diag}(\sigma(X_1), \ldots, \sigma(X_p)) \) and \( y_i = D^{-1} x_i \), \( i = 1, \ldots, n \).

2. Compute \( U = (U_{jk}) \) by applying \( \nu \) to the \( Y \) columns.

3. Compute the eigenvectors \( e_j \) of \( U \) and let \( E \) be the matrix whose columns are the \( e_j \)’s.

4. Let \( A = DE \), \( z_i = A^{-1} x_i \), in which case \( V = A \Gamma A' \) and \( t(X) = A \nu \), where \( \Gamma = \text{diag}(\sigma(Z_1)^2, \ldots, \sigma(Z_p)) \) and \( \nu = (\mu(Z_1), \ldots, \mu(Z_p)) \).

Maronna and Zamar (2002) noted that the above procedure can be iterated and they report results suggesting that a single iteration be used. More precisely, compute \( V \) and \( t \) for \( Z \) (the matrix corresponding to \( z_i \) computed in step 4) and express them in the original coordinate system, namely, \( V_2 = AV(Z)A' \) and \( t_2(X) = At(Z) \).

Maronna and Zamar showed that the estimate can be improved by a reweighting step. Let

\[
d_i = \sum \left(\frac{z_{ij} - \mu(Z_j)}{\sigma(Z_j)}\right),
\]

\[
w_i = I(d_i \leq d_0),
\]

and

\[
d_0 = \frac{\chi_{p, \beta}^2 \text{med}(d_1, \ldots, d_n)}{\chi_{p, 5}^2},
\]

where \( \chi_{p, \beta}^2 \) is the \( \beta \) quantile of the Chi-squared distribution with \( p \) degrees of freedom and \( \text{med} \) denotes the median. The measure of location is now estimated to be

\[
t_w = \frac{\sum w_i x_i}{\sum w_i},
\]

and the measure of scatter is

\[
V_w = \frac{\sum w_i (x_i - t_w)(x_i - t_w)'}{\sum w_i}.
\]
When using the OGK estimator to check for outliers in this study (2) was used. Results reported by Maronna and Zamar (2002) suggest using \( \beta = .9 \), but Wilcox (2008) found that this can result in \( p_n \) exceeding 0.05 by a considerable amount when \( n \) is small, moreover, \( p_n \) is unstable as a function of \( n \). Thus,

\[
\beta = \max(.95, \min(.99, 1/n + .94)),
\]

was found to be more satisfactory and was therefore used in this research.

The TBS Estimator

Rocke (1996) proposed an estimator known as the translated-biweight S (TBS) estimator. Generally, S-estimators of multivariate location and scatter are values for \( \hat{\Theta} \) and \( \hat{S} \) that minimize \( \|S\| \), the determinant of \( S \), subject to

\[
\frac{1}{n} \sum \xi((X_i - \hat{\Theta})'S^{-1}(X_i - \hat{\Theta}))^{1/2} = b_0, \tag{4}
\]

where \( b_0 \) is some constant, and \( \xi \) is a non-decreasing function. However, Rocke (1996) showed that S-estimators can be sensitive to outliers even if the breakdown point is close to 0.5. He suggested an alternative approach where the function \( \xi(d) \) is defined as follows: let \( m \) and \( c \) be values to be determined, then when \( m \leq d \leq m + c \),

\[
\xi(d) = \frac{m^2}{2} - \frac{m^2(m^4 - 5m^2c^2 + 15c^4)}{30c^4} + d^2 \left( \frac{5m^4 - m^2c^2}{2c^4} - \frac{m^3}{c^2} \right) + d^3 \left( \frac{4m^4 - 4m^3}{3c^2} - \frac{4m^3}{3c^4} \right) + d^4 \left( \frac{3m^2}{2c^4} - \frac{1}{2c^2} \right) \frac{4md^2}{5c^4} + \frac{d^6}{6c^4}
\]

for \( 0 \leq d < m \)

\[
\xi(d) = \frac{d^2}{2};
\]

and for \( d > m + c \),

\[
\xi(d) = \frac{m^2}{2} + \frac{c(5c + 16m)}{30}.
\]

The values for \( m \) and \( c \) can be chosen to achieve both the desired breakdown point and the asymptotic rejection probability, roughly referring to the probability that a point will get zero weight when the sample size is large. If the asymptotic rejection probability is \( \gamma \), for example, then \( m \) and \( c \) are determined by

\[
E_{\epsilon_\alpha}(\xi(d)) = b_0
\]

and

\[
m + c = \sqrt{\chi^2}_{p, 1-\gamma}.
\]

An iterative estimation method was used to compute the measures of location and scatter (Rocke & Woodruff, 1993) which requires an initial estimate of location and scatter. Here the initial estimate is the FMCD estimator which was computed with the R function cov.mcd, but some results on using an alternative initial estimate are also mentioned herein. As with the OGK estimator, when using TBS checks for outliers are based on (2).

Median Ball Algorithm

Following Olive (2004, 2007), the median ball algorithm (RMBA) begins with two initial estimates of location and scatter, both of which are based on an iterative algorithm. The strategy is as follows. For the \( j \)th estimator (\( j = 1, 2 \)), let \( (T_{0,j}, C_{0,j}) \) be some starting value. Compute all \( n \) Mahalanobis distances \( D_j(T_{0,j}, C_{0,j}) \) based on this measure location and scatter. Next estimate the usual mean and covariance matrix based on the \( c_n \approx n/2 \) cases corresponding to the smallest distances, this yields \( (T_{1,j}, C_{1,j}) \). Repeating this process, which is based on \( D_j(T_{1,j}, C_{1,j}) \), yields an updated measure of location and scatter, \( (T_{2,j}, C_{2,j}) \); following Olive (2005, 2007) \( (T_{5,j}, C_{5,j}) \) was used.
The first of the two starting values used by Olive takes \((T_1, C_{0,1})\) to be the usual mean and covariance matrix. The other starting value, \((T_{0,2}, C_{0,2})\), is the usual mean and covariance based on the \(c_n\) cases that are closest to the coordinatewise median in Euclidean distance. Let \((T_{\delta,i}, C_{\delta,i}) = (T_{5,i}, C_{5,i})\), where \(i = 1\) if the determinant \(|C_{5,1}| \leq |C_{5,2}|\), otherwise \(i = 2\). The MBA estimator of location is \(T_A\) and the measure of scatter is

\[
C_{\text{MBA}} = \frac{\text{MED}(D_i^2(T_{\delta,i}, C_{\delta,i}))}{\chi^2_{p,5}} C_A
\]

To compute the RMBA estimate, first compute \(D_i^2(T_{\text{MBA}}, C_{\text{MBA}})\), then

1. Compute the classical estimator \((T, C)\) for the cases with \(D_i^2 \leq \chi^2_{p,975}\).
2. Scale for normality: let \(T_1 = T\) and

\[
C_1 = \frac{\text{MED}(D_i^2(T, C))}{\chi^2_{p,5}}
\]

Repeat steps 1 and 2 to obtain \((T_{\text{RMBA}}, C_{\text{RMBA}})\).

(The R function rmba available at www.math.siu.edu/olive/rpack.txt, computes the RMBA estimate of location and scatter and was used in the simulations.)

Wilcox (2008) found that if the Mahalanobis distance is computed using the RMBA estimator, and points are declared outliers using (2) with \(\alpha = 0.975\), the outside rate per observation is reasonably close to 0.05 under normality, provided that \(n/p \geq 10\), at least for \(2 \leq p \leq 12\); otherwise the outside rate per observation can be very unsatisfactory. For example, with \(n = 20\) and \(p = 5\) it was estimated to exceed 0.24 regardless of the correlation among the variables.

Thus, this approach is not as satisfactory compared to the OGK and TBS methods, but it was included for two reasons. First, the efficiency of the RMBA estimate of location, relative to the other methods considered, is unknown. Second, when applying the MGV method, an initial estimate of the center of a data cloud is required, and using RMBA appears to have a practical advantage in terms of controlling the outside rate per observation.

The Minimum Generalized Variance Method

From basic multivariate techniques, the generalized variance is the determinant of the usual covariance matrix; it reflects how tightly a cloud of points is clustered together. The minimum generalized variance (MGV) method is based on the fact that the generalized variance is not robust; a single unusual point can greatly inflate its value. The MGV method is applied as follows:

1. Initially, all \(n\) points are described as belonging to set \(A\).
2. Find the \(p\) points that are most centrally located (many options exist to accomplish this). Based on results in Wilcox (2008), the approach used here takes the \(p\) most centrally located points to be the \(p\) points having the smallest Mahalanobis distance based on the RMBA estimators, \(T_A\) and \(C_{\text{RMBA}}\).
3. Remove the \(p\) centrally located points from set \(A\) and put them into set \(B\). At this step, the generalized variance of the points in set \(B\) is zero.
4. If among the points remaining in set \(A\), the \(i^{th}\) point is placed in set \(B\), then the generalized variance of the points in set \(B\) will be changed to some value labeled \(s^2_{g,i}\), that is associated with every point remaining in \(A\). The value \(s^2_{g,i}\), is the resulting generalized variance when it - and it only - is placed in set \(B\). Compute \(s^2_{g,i}\) for every point in \(A\).
5. Among the \(s^2_{g,i}\) values computed in the previous step, permanently remove the point associated with the smallest \(s^2_{g,i}\) value from set \(A\) and put it in set \(B\). That is, find the point in set \(A\) which is most tightly clustered.
together with the points in set B; after this point is identified, permanently remove it from A and place it in B.

6. Repeat steps 4 and 5 until all points are now in set B.

The first p points removed from set A have a generalized variance of zero, this is labeled \( s_{g(1)}^2 = \cdots = s_{g(p)}^2 = 0 \). When each point is removed from A and put into B (using steps 3 and 4), the resulting generalized variance of set B is labeled \( s_{g(p+1)}^2 \), as this process continues each point has associated with it some generalized variance when it is put into set B. Based on this process, the \( i \)th point has associated with it one of the generalized variances computed. For convenience, this generalized variance associated with the \( i \)th point, \( s_{g(i)}^2 \), is labeled \( C_i \).

The p deepest points have \( C \) values of zero. Points located at the edges of a scatterplot have the highest \( C \) values meaning that they are relatively far from the center of the cloud of points. A strategy for detecting outliers is simply applying some good univariate outlier rule to the \( C_i \) values. Note that a point would be declared only if an outlier \( C_i \) is large.

In terms of maintaining an outside rate per observation that is both stable as a function of \( n \) and \( p \), and approximately equal to 0.05 under normality, a boxplot rule for detecting outliers seems best when \( p = 2 \), and for \( p > 2 \) a slight generalization of Carling’s (2002) modification of the boxplot rule appears to perform well. In particular, if \( p = 2 \), then the \( i \)th point is declared an outlier if

\[
C_i > q_2 + 1.5(q_2 - q_1),
\]

where \( q_1 \) and \( q_2 \) are the ideal fourths based on the \( C_i \) values. For \( p > 2 \) variables, the \( i \)th point is declared an outlier if

\[
C_i > M_C + \sqrt{\chi_{0.975,p}^2(q_2 - q_1)},
\]

where \( M_C \) is the usual median of the \( C_i \) values. (Thus, the inverse of a covariance matrix and Mahalanobis distance do not play a role when checking for outliers.)

A criticism, when detecting outliers among the \( C_i \) values, is that the interquartile range has a breakdown point of 0.25. Ideally, a univariate outlier detection method would have a breakdown point of 0.5, the highest possible value. This can be achieved with a commonly used MAD-median rule. When \( p = 2 \), for example, it means that a point \( X_i \) is declared an outlier if

\[
\frac{|C_i - M_C|}{MAD_C} > 2.24,
\]

A Projection Method

Consider any projection of data onto a straight line. A projection-type method for detecting outliers among multivariate data is based on the idea that, if a point is an outlier, then it should be an outlier for some projection of the \( n \) points. Thus, if we were possible to consider all possible projections and, if for some projection a point is an outlier, then the point is declared an outlier. Not all projections can be considered, hence, following Wilcox (2005), the strategy is to orthogonally project the data onto all \( n \) lines formed by the center of the data cloud, as represented by \( \hat{\xi} \), and each \( X_i \). Here, \( \hat{\xi} \) was taken to be the RMBA measure of location. (Checks suggest that other choices for \( \hat{\xi} \) have no practical value for the problem considered herein.)

The computational details are as follows. Fix \( i \), and for the point \( X_i \), orthogonally project all \( n \) points onto the line connecting \( \hat{\xi} \) and \( X_i \), and let \( D_{ij} \) be the
distance between \( \hat{\xi} \) and \( X_j \) based on this projection. Let

\[
A_i = X_i - \hat{\xi},
\]

and

\[
B_j = X_j - \hat{\xi},
\]

where both \( A_i \) and \( B_j \) are column vectors having length \( p \). Next let

\[
C_j = \frac{A'B_j}{B_j},
\]

where \( j = 1, \ldots, n \). Then when projecting the points onto the line between \( X_i \) and \( \hat{\xi} \), the distance of the \( j \)th point from \( \hat{\xi} \) is

\[
D_{ij} = || C_j ||,
\]

where

\[
|| C || = \sqrt{C_1^2 + \cdots + C_p^2}.
\]

Here, an extension of Carling’s modification of the boxplot rule (similar to the modification used by the MGV method) is used to check for outliers among \( D_{ij} \) values. Let

\[
\ell = \left\lceil \frac{n}{4} + \frac{5}{12} \right\rceil,
\]

where \( \lceil \cdot \rceil \) is the greatest integer function and let

\[
h = \frac{n}{4} + \frac{5}{12} - \ell.
\]

For fixed \( i \), let \( D_{i(1)} \leq \cdots \leq D_{i(n)} \) be the \( n \) distances written in ascending order.

If the ideal fourths associated with the \( D_{ij} \) values are

\[
q_1 = (1-h)D_{i(\ell)} + hD_{i(\ell+1)}
\]

and

\[
q_2 = (1-h)D_{i(k)} + hD_{i(k-1)},
\]

where \( k = n - j + 1 \), then the \( j \)th point is declared an outlier if

\[
D_{ij} > M_D + \sqrt{\frac{\chi^2_{0.95, p}}{n}} (q_2 - q_1),
\]

where \( M_D \) is the usual sample median based on \( D_1, \ldots, D_n \).

The process described is for a single projection; for fixed \( i \), points are projected onto the line connecting \( X_i \) to \( \hat{\xi} \). Repeating this process for each \( i, i = 1, \ldots, n \), a point is declared an outlier if for any of these projections, it satisfies equation (8). This will be called method OP, which has certain similarities with a projection method suggested by Pena and Prieto (2001). One important difference is that the method used Pena and Prieto is based on the usual sample mean, which is not robust and could result in masking.

As was the case with the MGV method, a simple and seemingly desirable modification of the method described is to replace the interquartile range with the median absolute deviation (MAD) measure of scale based on the values \( D_1, \ldots, D_n \). Thus, if MAD is the median of the values \( |D_1 - M_D|, \ldots, |D_n - M_D| \), which is denoted by \( MAD_D \), then the \( j \)th point is declared an outlier if for any \( i \),

\[
D_{ij} > M_D + \sqrt{\frac{\chi^2_{0.95, p}}{n}} \cdot 0.6745 \cdot MAD_D
\]

(Similar to the MGV method, equation (2) is not used when checking for outliers.) Equation (9) represents an approximation of the method given by Donoho and Gasko (1992).

An appealing feature of MAD is that it has a higher finite sample breakdown point than the interquartile range; however, a negative feature of equation (9) is that the outside rate per observation appears to be less stable as a function of \( n \). In the bivariate case, for example, it is approximately 0.09 with \( n = 10 \) and drops below 0.02 as \( n \) increases. For the same situations, the outside rate per observation using equation (9) ranges, approximately, between 0.043 and 0.038.
Summary of the Estimators

In summary, eight alternatives to the sample mean were considered. The first three were RMBA, OGK and TBS. The remaining five are skipped estimators where outliers are removed after which the mean of the remaining data is computed. Three of these five estimators use (2) in conjunction with MVE, MCD and TBS; the other two use the MGV and OP outlier detection methods with the initial measure of location given by RMBA. For convenience, the estimators RMBA, OGK, MCD(S), OP, MVE(S), MGV and TBS(S) are labeled $\hat{\eta}_1, \ldots, \hat{\eta}_7$, respectively. The usual sample mean is labeled $\hat{\eta}_8$.

Results

Simulations were used to compare the efficiency of the sample mean to the eight alternative estimators. The efficiency of the $j$th estimator ($j = 1, \ldots, 8$) was measured with

$$E = \frac{V(\hat{\eta}_j)}{V(\hat{\eta}_8)},$$

where $V(\hat{\eta}_j)$ is the generalized variance associated with the sampling distribution of $\hat{\eta}_j$.

All simulations were conducted using the software R. Methods OP and MGV were applied with software from Wilcox (2005) that was downloaded from http://psychology.usc.edu/faculty\_homepage.php?id=43. (The R function smean in Wilcox (2005) defaults to method OP. The R code for all estimators is available from the author upon request.)

To describe how data were generated, first consider the univariate case. An observation $X$ from a $g$-and-$h$ distribution (Hoaglin, 1985) is generated by first generating a value from a standard normal distribution yielding $Z$, for example, and computing

$$X = \frac{\exp(gZ)-1}{g}\exp(hZ^2/2)$$

where $g$ and $h$ are parameters that determine the third and fourth moments. When $g = 0$, this last equation is taken to be

$$X = Z \exp(hZ^2/2)$$

For the multivariate case, data were generated from a multivariate normal distribution having a common correlation, $\rho$, and the values of the marginal distributions were transformed to a $g$-and-$h$ distribution. The four (marginal) $g$-and-$h$ distributions used were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($g = 0$, $h = 0.2$), an asymmetric distribution with relatively light tails ($g = 0.5$, $h = 0$), and an asymmetric distribution with heavy tails ($g = 0.5$, $h = 0.2$). (For details about these distributions, see Hoaglin, 1985.)

The values for $\rho$ were taken to be 0, 0.5 and 0.8.

Tables 1-6 show the estimated efficiency of the eight estimators based on 1,000 replications. One method to condense the results in a useful way is to determine which robust estimator has the best efficiency among each of the 72 conditions studied. The OP estimator was best for 56 conditions and it was among the top two for 62 conditions. Another perspective considers which estimator competes best with the mean under normality; with two exceptions, this is method OP. The two exceptions occur when $\rho = 0$ and $p = 5$ or $p = 8$, in which case MGV is best.

With $p = 5$ the advantage of OP over MGV is not striking but with $p=8$ (and if $\rho =0$), MGV may have a worthwhile advantage. MGV is often among the two best estimators however, when sampling from a heavy-tailed distribution the mean can have better efficiency - sometimes strikingly so - even when other estimators beat the mean by a considerable amount. Although, RMBA, OGK and TBS do not compete well with OP in general, they can offer an advantage when $p = 8$, $\rho = 0.5$ or $\rho = 0.8$ and sampling is from a skewed, heavy-tailed distribution.

Conclusion

The success of the OP method is not surprising considering the results in detecting outliers...
recently summarized in Wilcox (2008). Also based on results from Wilcox (2008), there was some anticipation that MGV would compete effectively with OP. Under some conditions it is a reasonable alternative, but it seems that, in terms of efficiency, the skipped estimator based on the OP outlier detection method is generally preferable, sometimes by a substantial amount. The poor performance of MGV when \( p = 8 \) and sampling is from a skewed, heavy-tailed distribution, was not expected. The OGK, TBS and RMBA estimators compete well with OP, particularly when sampling from a skewed, heavy-tailed distribution and \( p \geq 5 \), but for routine use, OP seems preferable and - for a variety of situations - it offers a distinct advantage.

<table>
<thead>
<tr>
<th>n</th>
<th>g</th>
<th>h</th>
<th>p</th>
<th>RMBA</th>
<th>OGK</th>
<th>TBS</th>
<th>MCD(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.84</td>
<td>1.84</td>
<td>2.84</td>
<td>2.93</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.47</td>
<td>1.98</td>
<td>2.87</td>
<td>2.7</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>10.06</td>
<td>4.2</td>
<td>8.42</td>
<td>11.08</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2.73</td>
<td>3.73</td>
<td>3.25</td>
<td>10.13</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>112.25</td>
<td>9.14</td>
<td>33.97</td>
<td>33.97</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>13.2</td>
<td>7.19</td>
<td>5.99</td>
<td>57.52</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.61</td>
<td>0.71</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.64</td>
<td>0.74</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>1.03</td>
<td>0.87</td>
<td>1.19</td>
<td>1.36</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.52</td>
<td>0.6</td>
<td>0.53</td>
<td>0.77</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>2.65</td>
<td>0.83</td>
<td>1.91</td>
<td>1.91</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>1.18</td>
<td>0.87</td>
<td>0.71</td>
<td>2.02</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.49</td>
<td>1.58</td>
<td>1.57</td>
<td>1.94</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.39</td>
<td>1.33</td>
<td>1.31</td>
<td>2.04</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>5.15</td>
<td>3.68</td>
<td>5.7</td>
<td>6.65</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>3.38</td>
<td>2.86</td>
<td>3.44</td>
<td>6.92</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>16.34</td>
<td>12.19</td>
<td>19.63</td>
<td>19.58</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>17.32</td>
<td>11.71</td>
<td>13.98</td>
<td>54.34</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.27</td>
<td>0.36</td>
<td>0.27</td>
<td>0.33</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.13</td>
<td>0.15</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.08</td>
<td>0.16</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.06</td>
<td>0.07</td>
<td>0.68</td>
<td>0.1</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>0.09</td>
<td>0.15</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Table 2: Estimated Efficiency for First Four Estimators, $\rho = .5$

<table>
<thead>
<tr>
<th>n</th>
<th>g</th>
<th>h</th>
<th>p</th>
<th>RMBA</th>
<th>OGK</th>
<th>TBS</th>
<th>MCD(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.76</td>
<td>1.99</td>
<td>2.85</td>
<td>2.82</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.39</td>
<td>1.88</td>
<td>2.51</td>
<td>2.38</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>8.46</td>
<td>3.7</td>
<td>8.32</td>
<td>10.53</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>3.15</td>
<td>4.05</td>
<td>3.42</td>
<td>12.06</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>126.95</td>
<td>11.09</td>
<td>41.53</td>
<td>1.49</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>13.73</td>
<td>7.89</td>
<td>6.5</td>
<td>65.62</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.62</td>
<td>0.73</td>
<td>0.7</td>
<td>0.76</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.56</td>
<td>0.69</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.67</td>
<td>0.51</td>
<td>0.64</td>
<td>0.69</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.31</td>
<td>0.38</td>
<td>0.3</td>
<td>0.41</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>1.88</td>
<td>0.57</td>
<td>1.53</td>
<td>1.52</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.49</td>
<td>0.34</td>
<td>0.28</td>
<td>0.75</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.13</td>
<td>1.14</td>
<td>1.11</td>
<td>1.42</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.17</td>
<td>1.1</td>
<td>0.94</td>
<td>1.54</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>1.51</td>
<td>1.43</td>
<td>1.95</td>
<td>2.05</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>1.91</td>
<td>1.39</td>
<td>1.63</td>
<td>2.8</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>2.5</td>
<td>1.57</td>
<td>3.78</td>
<td>3.8</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>1.72</td>
<td>0.97</td>
<td>1.34</td>
<td>2.42</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.18</td>
<td>0.24</td>
<td>0.18</td>
<td>0.21</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.18</td>
<td>0.2</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>
Table 3: Estimated Efficiency for First Four Estimators, $\rho = .8$

<table>
<thead>
<tr>
<th>n</th>
<th>g</th>
<th>h</th>
<th>p</th>
<th>RMBA</th>
<th>OGK</th>
<th>TBS</th>
<th>MCD(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.94</td>
<td>2.02</td>
<td>2.88</td>
<td>2.97</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.42</td>
<td>2.14</td>
<td>2.73</td>
<td>2.46</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>11.47</td>
<td>4.18</td>
<td>9.33</td>
<td>12.15</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2.64</td>
<td>3.74</td>
<td>3.03</td>
<td>10.33</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>119.43</td>
<td>9.11</td>
<td>36.34</td>
<td>36.31</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>11.69</td>
<td>6.39</td>
<td>6.07</td>
<td>63.91</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.52</td>
<td>0.69</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.54</td>
<td>0.69</td>
<td>0.56</td>
<td>0.61</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.5</td>
<td>0.36</td>
<td>0.45</td>
<td>0.51</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.22</td>
<td>0.3</td>
<td>0.2</td>
<td>0.29</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.56</td>
<td>0.2</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.17</td>
<td>0.13</td>
<td>0.1</td>
<td>0.27</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.18</td>
<td>1.29</td>
<td>1.16</td>
<td>1.33</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.18</td>
<td>1.28</td>
<td>0.96</td>
<td>1.53</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.87</td>
<td>0.85</td>
<td>1.14</td>
<td>1.17</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.74</td>
<td>0.55</td>
<td>0.77</td>
<td>1.08</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.61</td>
<td>0.39</td>
<td>0.79</td>
<td>0.7</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.54</td>
<td>0.23</td>
<td>0.52</td>
<td>0.7</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.09</td>
<td>0.14</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.11</td>
<td>0.15</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>
Table 4: Estimated Efficiency for Four Skipped Estimators, $\rho = 0$

<table>
<thead>
<tr>
<th>n</th>
<th>g</th>
<th>h</th>
<th>p</th>
<th>OP</th>
<th>MVE(S)</th>
<th>MGV</th>
<th>TBS(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.36</td>
<td>2.48</td>
<td>1.47</td>
<td>1.92</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.32</td>
<td>1.91</td>
<td>1.5</td>
<td>1.53</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2.22</td>
<td>6.98</td>
<td>1.49</td>
<td>8.03</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1.86</td>
<td>3.48</td>
<td>1.36</td>
<td>2.39</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>3.54</td>
<td>12.64</td>
<td>2.21</td>
<td>34.25</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>2.74</td>
<td>9.55</td>
<td>1.93</td>
<td>4.63</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.56</td>
<td>0.64</td>
<td>0.61</td>
<td>0.59</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.56</td>
<td>0.67</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.45</td>
<td>0.84</td>
<td>0.91</td>
<td>0.84</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.49</td>
<td>0.71</td>
<td>0.93</td>
<td>0.58</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.52</td>
<td>2.55</td>
<td>2.09</td>
<td>3.28</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.39</td>
<td>0.87</td>
<td>1.18</td>
<td>0.58</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.02</td>
<td>1.86</td>
<td>1.12</td>
<td>1.46</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>1.21</td>
<td>2.22</td>
<td>1.35</td>
<td>1.55</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>1.56</td>
<td>4.36</td>
<td>1.66</td>
<td>5.49</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.45</td>
<td>0.66</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.51</td>
<td>2.27</td>
<td>1.87</td>
<td>2.95</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.39</td>
<td>0.7</td>
<td>1.05</td>
<td>0.57</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.2</td>
<td>0.25</td>
<td>0.26</td>
<td>0.22</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.21</td>
<td>0.25</td>
<td>0.24</td>
<td>0.23</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.06</td>
<td>0.19</td>
<td>1.21</td>
<td>0.18</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.05</td>
<td>0.11</td>
<td>1.18</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Table 5: Estimated Efficiency for Four Skipped Estimators, $\rho = .5$

<table>
<thead>
<tr>
<th>n</th>
<th>g</th>
<th>h</th>
<th>p</th>
<th>OP M</th>
<th>VE(S)</th>
<th>MGV</th>
<th>TBS(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.31</td>
<td>2.55</td>
<td>1.36</td>
<td>0.174</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.32</td>
<td>2.1</td>
<td>1.63</td>
<td>1.56</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1.43</td>
<td>6.12</td>
<td>1.52</td>
<td>8.28</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1.33</td>
<td>3.27</td>
<td>1.42</td>
<td>2.27</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1.3</td>
<td>11.32</td>
<td>2.5</td>
<td>38.73</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1.18</td>
<td>7.73</td>
<td>1.97</td>
<td>3.77</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.52</td>
<td>0.65</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.47</td>
<td>0.52</td>
<td>0.5</td>
<td>0.47</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.33</td>
<td>0.69</td>
<td>1.07</td>
<td>0.71</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.26</td>
<td>0.43</td>
<td>0.79</td>
<td>0.3</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.26</td>
<td>1.16</td>
<td>1.83</td>
<td>1.15</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.22</td>
<td>0.44</td>
<td>1.12</td>
<td>0.29</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>0.98</td>
<td>1.46</td>
<td>1.09</td>
<td>1.3</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>0.98</td>
<td>1.5</td>
<td>1</td>
<td>1.22</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.69</td>
<td>1.55</td>
<td>0.142</td>
<td>1.69</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.74</td>
<td>1.76</td>
<td>1.56</td>
<td>1.45</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.84</td>
<td>3.31</td>
<td>1.86</td>
<td>3.65</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.7</td>
<td>1.84</td>
<td>1.93</td>
<td>2.18</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.16</td>
<td>0.2</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.12</td>
<td>0.13</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.02</td>
<td>0.04</td>
<td>1.19</td>
<td>0.03</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;0.01</td>
<td>0.03</td>
<td>2.2</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;0.01</td>
<td>0.02</td>
<td>1.36</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
Table 6: Estimated Efficiency for Four Skipped Estimators, \( \rho = .8 \)

<table>
<thead>
<tr>
<th>n</th>
<th>g</th>
<th>h</th>
<th>p</th>
<th>OP</th>
<th>MVE(S)</th>
<th>MGV</th>
<th>TBS(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.21</td>
<td>2.48</td>
<td>0.14</td>
<td>1.8</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.3</td>
<td>2.23</td>
<td>1.62</td>
<td>1.59</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1.16</td>
<td>7.26</td>
<td>1.72</td>
<td>8.7</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1.21</td>
<td>3.3</td>
<td>1.48</td>
<td>2.4</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>11.11</td>
<td>14.55</td>
<td>2.68</td>
<td>39.5</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1.04</td>
<td>8.88</td>
<td>2.31</td>
<td>4.77</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.49</td>
<td>0.64</td>
<td>0.53</td>
<td>0.54</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>2</td>
<td>0.52</td>
<td>0.63</td>
<td>0.58</td>
<td>0.54</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.27</td>
<td>0.45</td>
<td>0.94</td>
<td>0.46</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>5</td>
<td>0.2</td>
<td>0.24</td>
<td>0.75</td>
<td>0.19</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.21</td>
<td>0.9</td>
<td>1.93</td>
<td>0.62</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0.2</td>
<td>8</td>
<td>0.14</td>
<td>0.17</td>
<td>1.03</td>
<td>0.11</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>0.94</td>
<td>1.34</td>
<td>1.01</td>
<td>1.18</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>2</td>
<td>0.98</td>
<td>1.41</td>
<td>1.06</td>
<td>1.15</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.61</td>
<td>1.1</td>
<td>1.35</td>
<td>1.02</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>5</td>
<td>0.51</td>
<td>0.87</td>
<td>1.03</td>
<td>0.69</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.43</td>
<td>1.64</td>
<td>1.98</td>
<td>1.32</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0</td>
<td>8</td>
<td>0.35</td>
<td>0.45</td>
<td>1.46</td>
<td>0.54</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>2</td>
<td>0.12</td>
<td>0.13</td>
<td>0.01</td>
<td>0.12</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.02</td>
<td>0.01</td>
<td>0.79</td>
<td>0.01</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>5</td>
<td>0.01</td>
<td>0.01</td>
<td>0.82</td>
<td>0.01</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>2.11</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>50</td>
<td>0.5</td>
<td>0.2</td>
<td>8</td>
<td>&lt;.01</td>
<td>&lt;.01</td>
<td>1.37</td>
<td>&lt;.01</td>
</tr>
</tbody>
</table>
References


The Effectiveness of Stepwise Discriminant Analysis as a Post Hoc Procedure to a Significant MANOVA

Erik L. Heiny  
Utah Valley University

Daniel J. Mundfrom  
University of Northern Colorado

The effectiveness of SWDA as a post hoc procedure in a two-way MANOVA was examined using various numbers of dependent variables, sample sizes, effect sizes, correlation structures, and significance levels. The procedure did not work well in general except with small numbers of variables, larger samples and low correlations between variables.

Key words: Stepwise discriminant analysis, MANOVA, post hoc procedures.

Introduction

One common type of research question in multivariate analysis involves searching for differences between multiple groups on several different response variables. Considering response variables as a vector of dependent variables, a one-way MANOVA can be used to test the hypothesis that the mean vectors are the same across groups. However, if a significant MANOVA has been found, how does the researcher determine which of the response variables contribute to group differences?

Currently, most researchers use either multiple univariate F-tests, which are simply inappropriate, or descriptive discriminant analysis (DDA), which has been shown to lack power through simulation studies. Hawkins (1976) proposed the use of a stepwise MANOVA procedure, similar to stepwise regression, for selecting the best subset of variables to use in the MANOVA analysis. Hawkins further advocated for a Bonferroni adjustment to the α-level used at each step in the stepwise selection to control the overall Type I error rate, neither of these suggestions, however, seem to be in much use today. Another approach used by some researchers is stepwise discriminant analysis (SWDA). Criticisms of stepwise methods in general have been well-documented in the literature, most notably by Thompson (1995), which would appear to also apply to Hawkins’ stepwise MANOVA procedure.

Essentially the criticisms center on stepwise methods being biased towards finding significance. Although this is a legitimate concern, it should be less prevalent in the context of this study; in this study real group differences exist on the dependent variables, therefore SWDA is not just fishing for differences that do not exist. Considering that some researchers are currently using SWDA in this context and that univariate F-tests and DDA are poor alternatives, empirical evidence is needed regarding the viability of SWDA as a post hoc procedure to a significant MANOVA.

The purpose of this research is to investigate the effectiveness of SWDA in distinguishing between significant and non-significant dependent variables when the MANOVA null hypothesis has been rejected. Specifically, it examines what the percentage of MANOVA dependent variables with means that differ between groups that are correctly identified as significantly different in a two-group SWDA (i.e., the power), and the percentage of MANOVA dependent variables with means that are the same in both groups that are incorrectly identified as significantly different.
different in a two-group SWDA (i.e., the Type I error). The effect of sample size, n, the number of dependent variables in the MANOVA, p, the correlation structure among the dependent variables, ρ, the effect size, d, and the significance level used in the stepwise selection, α, were also investigated.

Rencher and Larson (1980) performed a Monte Carlo simulation to examine the bias in Wilk’s lambda in SWDA. In SWDA, an F-statistic can be used to test the significance of the reduction in Wilk’s lambda when an additional variable is added to the model. The larger the reduction in Wilk’s lambda due to the additional variable, the larger the F-statistic will become. Rencher and Larson note that if an arbitrary variable is considered for entry, the F-statistic follows a true F-distribution. However, in SWDA several variables are considered for entry at each step and the maximum F-statistic from these variables is compared to the F-critical value. Because the F-statistic is maximized at each step, it does not follow an F-distribution and the procedure becomes biased towards selecting variables that do not contain discriminatory information. Rencher and Larson conclude that the bias becomes most pronounced when there are a large number of variables under consideration and a relatively small sample size. They write, “In the author’s experience, such cases are fairly common. Habbema and Hermans (1977, p. 492) note that ‘sample sizes of say 10-40 are not unusual, with a number of variables ranging from 10-200.’” (p. 350). The most drastic case in this study will be sample sizes of 50 with the number of variables equal to 8.

In addition, Rencher and Larson (1980) write, “we have restricted out attention to the null case of no difference between groups so as to provide some indication of the levels Wilk’s lambda may reach when there is no real separation from group to group” (p. 351). In this study, SWDA was used when the null hypothesis is false, that is, real separation exists from group to group. Therefore, the bias in Wilk’s lambda was not expected to be as severe in this study, but Type I errors in excess of alpha were likely and were watched closely.

Methodology

A Monte Carlo simulation was run using SAS PROC Interactive Matrix Language (IML). Two p-dimensional multivariate normal populations were created with characteristics that varied according to pre-set levels of the number of MANOVA dependent variables, p, which varied across the values, 2, 3, 4, 5, 6, 7, and 8, and a correlation structure among the p variables. In one population, the mean vector contained all zeros, whereas in the other population mean vector had half of the values set at 0 while the other half differed from 0 by an effect size, d, that varied across 0.2 (small), 0.5 (medium), and 0.8 (large). When the value of p was odd, the mean of the extra variable was set at 0; for example, with p = 5 and a small effect size, the two mean vectors were:

\[
\mu_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 0 \\ 0 \\ 0.2 \\ 0.2 \end{bmatrix}.
\]

Both populations were generated with the same correlation matrix, ρ.

Six different correlation structures were examined. In each structure, variables were divided into set A, those that had the same mean in both groups, and set B, those that had means that differed between the groups. The within-set correlations, those between pairs of variables in set A (and between pairs of variables in set B), were varied across the values 0.20, 0.40 and 0.60. Initially, the across-set correlations, those between pairs of variables in which one variable came from set A and the other came from set B, was set at 0.20. For example, with p = 5, the three correlation matrices used were:

\[
\rho_1 = \begin{bmatrix} 1 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1 \end{bmatrix}
\]
Because many of the scenarios examined with these correlation structures had large Type I error rates, the across-set correlations were reduced to 0.10 in order to see how this change would affect the results. Again for the \( p = 5 \) case, the three additional correlation matrices were:

\[
\rho_4 = \begin{bmatrix}
1 & .2 & .2 & .1 & .1 \\
.2 & 1 & .2 & .1 & .1 \\
.2 & .2 & 1 & .1 & .1 \\
.1 & .1 & .1 & 1 & .2 \\
.1 & .1 & .1 & .2 & 1 \\
\end{bmatrix}
\]

and

\[
\rho_5 = \begin{bmatrix}
1 & .4 & .4 & .1 & .1 \\
.4 & 1 & .4 & .1 & .1 \\
.4 & .4 & 1 & .1 & .1 \\
.1 & .1 & .1 & 1 & .4 \\
.1 & .1 & .1 & .4 & 1 \\
\end{bmatrix}
\]

Additionally, sample sizes were varied across 50, 100, 250, and 500 (although \( n = 500 \) was the only sample size used for the last three correlation structures), and the significance level used for variable selection in the SWDA was varied across 0.01, 0.05 and 0.10.

For each of 945 scenarios determined by the values of \( p, d, n, \rho \) and \( \alpha \), 5,000 replications were performed. Each replication consisted of selecting a random sample of size \( n \) from each population described above, which led to two sample mean vectors. A SWDA was performed on each sample using SAS PROC STEPDISC with the stepwise selection method and the F-test criterion for a chosen level of \( \alpha \). The percentage of correctly identified significant variables (power) and the percentage of non-significant variables incorrectly identified as significant (Type I error) were computed for each sample. Averaging these values across the 5,000 replications produced power and Type I error estimates for each scenario. Successful results were defined to be those situations for which power was maintained at 0.80 or higher and the Type I error rate did not exceed 0.10.

Results

Scenarios with Correlation Structure One, Two or Three

For correlation structures one, two and three, SWDA was only successful for certain situations when \( p \) was small, 2 or 3. As long as \( p \) was not larger than 3, varying the correlation structure between levels one, two and three had almost no effect on the results. The larger \( p \) became, however, the more the results changed for different correlation structures (see Tables 1 and 2). For \( p = 2 \) or 3 and a small sample size, \( n = 50 \), SWDA worked well for large effect sizes, \( d = 0.8 \), and \( \alpha = 0.01 \) (Table 2). Type I errors were inflated above \( \alpha \) but only to 0.03, and power was above 0.90.

As \( n \) increased to 100, and \( p \) was set equal to 2 or 3, SWDA was still successful for large effect sizes, but only when \( \alpha \) was set to 0.01 (see Table 2). Power was over 0.99 and Type I error was 0.06. Additionally, for the same levels of \( n \) and \( p \), SWDA worked well for medium effect sizes, \( d = 0.5 \), as long as \( \alpha \) was set to 0.05 or 0.01 (see Table 1). For \( \alpha = 0.01 \), power was around 0.82 and Type I error was near 0.025. For \( \alpha = 0.05 \), power was 0.94 and Type I error was around 0.09.

\[
\rho_6 = \begin{bmatrix}
1 & .6 & .6 & .1 & .1 \\
.6 & 1 & .6 & .1 & .1 \\
.6 & .6 & 1 & .1 & .1 \\
.1 & .1 & .1 & 1 & .6 \\
.1 & .1 & .1 & .6 & 1 \\
\end{bmatrix}
\]
EFFECTIVENESS OF SWDA AS A POST HOC SIGNIFICANT MANOVA PROCEDURE

Table 1: Power and Type I Error for $\alpha = 0.01$, $d = 0.5$, Across-Set $\rho = 0.2$

<table>
<thead>
<tr>
<th>p</th>
<th>Within-Set $\rho = 0.2$</th>
<th>Within-Set $\rho = 0.4$</th>
<th>Within-Set $\rho = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=50</td>
<td>n=100</td>
<td>n=250</td>
</tr>
<tr>
<td>2</td>
<td>0.4554 0.8214 0.9988 1.0000</td>
<td>0.4666 0.8268 0.9980 1.0000</td>
<td>0.4520 0.8190 0.9990 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0126 0.0266 0.0662 0.1616</td>
<td>0.0132 0.0282 0.0710 0.1560</td>
<td>0.0156 0.0246 0.0588 0.1218</td>
</tr>
<tr>
<td>3</td>
<td>0.4510 0.8326 0.9978 1.0000</td>
<td>0.4550 0.8226 0.9980 1.0000</td>
<td>0.4466 0.8280 0.9986 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0153 0.0256 0.0665 0.1447</td>
<td>0.0140 0.0245 0.0653 0.1315</td>
<td>0.0141 0.0242 0.0588 0.1218</td>
</tr>
<tr>
<td>4</td>
<td>0.3615 0.6393 0.9715 0.9999</td>
<td>0.3236 0.5100 0.8472 0.9947</td>
<td>0.3012 0.4613 0.6063 0.9083</td>
</tr>
<tr>
<td></td>
<td>0.0143 0.0457 0.1990 0.4206</td>
<td>0.0160 0.0312 0.1094 0.2720</td>
<td>0.0148 0.0275 0.0754 0.1804</td>
</tr>
<tr>
<td>5</td>
<td>0.3653 0.6388 0.9705 0.9999</td>
<td>0.3196 0.5088 0.8522 0.9953</td>
<td>0.2953 0.4601 0.6106 0.9077</td>
</tr>
<tr>
<td></td>
<td>0.0203 0.0462 0.1805 0.3631</td>
<td>0.0157 0.0327 0.1039 0.2196</td>
<td>0.0124 0.0275 0.0656 0.1430</td>
</tr>
<tr>
<td>6</td>
<td>0.3039 0.5242 0.9134 0.9990</td>
<td>0.2501 0.3790 0.6610 0.9062</td>
<td>0.2292 0.3215 0.4527 0.6637</td>
</tr>
<tr>
<td></td>
<td>0.0196 0.0664 0.2840 0.5677</td>
<td>0.0179 0.0373 0.1203 0.2646</td>
<td>0.0176 0.0282 0.0696 0.1586</td>
</tr>
<tr>
<td>7</td>
<td>0.3059 0.5315 0.9194 0.9996</td>
<td>0.2521 0.3763 0.6618 0.9135</td>
<td>0.2302 0.3193 0.4558 0.6677</td>
</tr>
<tr>
<td></td>
<td>0.0239 0.0618 0.2589 0.5055</td>
<td>0.0177 0.0330 0.1070 0.2200</td>
<td>0.0145 0.0248 0.0615 0.1361</td>
</tr>
<tr>
<td>8</td>
<td>0.2622 0.4622 0.8529 0.9970</td>
<td>0.2055 0.3072 0.5421 0.7793</td>
<td>0.1839 0.2458 0.3654 0.5203</td>
</tr>
<tr>
<td></td>
<td>0.0260 0.0844 0.3364 0.6519</td>
<td>0.0190 0.0374 0.1184 0.2383</td>
<td>0.0155 0.0275 0.0641 0.1388</td>
</tr>
</tbody>
</table>

Table 2: Power and Type I Error for $\alpha = 0.01$, $d = 0.8$, Across-Set $\rho = 0.2$

<table>
<thead>
<tr>
<th>p</th>
<th>Within-Set $\rho = 0.2$</th>
<th>Within-Set $\rho = 0.4$</th>
<th>Within-Set $\rho = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=50</td>
<td>n=100</td>
<td>n=250</td>
</tr>
<tr>
<td>2</td>
<td>0.9146 0.9994 1.0000 1.0000</td>
<td>0.9158 0.9990 1.0000 1.0000</td>
<td>0.9074 0.9992 1.0000 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0328 0.0662 0.1836 0.4286</td>
<td>0.0294 0.0602 0.1906 0.4310</td>
<td>0.0312 0.0600 0.1900 0.4278</td>
</tr>
<tr>
<td>3</td>
<td>0.9180 0.9984 1.0000 1.0000</td>
<td>0.9160 0.9992 1.0000 1.0000</td>
<td>0.9162 0.9986 1.0000 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0321 0.0622 0.1656 0.3547</td>
<td>0.0304 0.0587 0.1555 0.3119</td>
<td>0.0262 0.0542 0.1455 0.2880</td>
</tr>
<tr>
<td>4</td>
<td>0.6908 0.9575 1.0000 1.0000</td>
<td>0.5509 0.8083 0.9885 1.0000</td>
<td>0.4894 0.5855 0.9495 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0529 0.1698 0.4612 0.7833</td>
<td>0.0361 0.0979 0.3072 0.5063</td>
<td>0.0312 0.0645 0.2189 0.3919</td>
</tr>
<tr>
<td>5</td>
<td>0.6909 0.9612 1.0000 1.0000</td>
<td>0.5444 0.8162 0.9984 1.0000</td>
<td>0.4872 0.5874 0.9511 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0565 0.1530 0.4060 0.6827</td>
<td>0.0371 0.0867 0.2470 0.3790</td>
<td>0.0281 0.0563 0.1693 0.2840</td>
</tr>
<tr>
<td>6</td>
<td>0.5708 0.8580 0.9997 1.0000</td>
<td>0.4020 0.6227 0.9427 0.9995</td>
<td>0.3351 0.4275 0.6946 0.9328</td>
</tr>
<tr>
<td></td>
<td>0.0753 0.2309 0.5992 0.8669</td>
<td>0.0397 0.1097 0.2889 0.4639</td>
<td>0.0296 0.0602 0.1801 0.3013</td>
</tr>
<tr>
<td>7</td>
<td>0.5711 0.8704 0.9998 1.0000</td>
<td>0.4089 0.6258 0.9447 0.9996</td>
<td>0.3347 0.4313 0.6957 0.9379</td>
</tr>
<tr>
<td></td>
<td>0.0713 0.2092 0.5294 0.8123</td>
<td>0.0410 0.0972 0.2391 0.3765</td>
<td>0.0267 0.0562 0.1533 0.2334</td>
</tr>
<tr>
<td>8</td>
<td>0.4906 0.7726 0.9978 1.0000</td>
<td>0.3318 0.5055 0.8111 0.9866</td>
<td>0.2545 0.3454 0.5533 0.7617</td>
</tr>
<tr>
<td></td>
<td>0.0850 0.2610 0.6603 0.9231</td>
<td>0.0430 0.1033 0.2565 0.4191</td>
<td>0.0299 0.0595 0.1593 0.2369</td>
</tr>
</tbody>
</table>
As $n$ increased to 250 while $p$ remained equal to 2 or 3, the procedure was successful for medium effect sizes and $\alpha = 0.01$ (see Table 1). Power was over 0.99 and Type I error was less than 0.07. The procedure became too aggressive for large effect sizes, with observed Type I error going as high as 0.50 in some situations.

When $n$ increased to 500 while $p$ was still limited to 2 or 3, SWDA was only successful for small effect sizes, $d = 0.20$, and $\alpha = 0.05$. Power was approximately 0.89 and Type I error was near 0.09. When $\alpha$ was lowered to 0.01, Type I error dropped to 0.02 but power went down to 0.72. When $\alpha$ was increased to 0.10, power increased to 0.94 but Type I error was high, 0.15. Due to the aggressive nature of SWDA, the procedure did not work well for medium or large effect sizes when $n = 500$. The power was very high, but Type I error increased well above 0.10.

Scenarios with Correlation Structure Four, Five or Six

As noted, correlation structures four, five and six were simulated with $n = 500$ to investigate the increase in Type I error which accompanied any increase in effect size. Recall that in correlation structures one, two and three, the across-set correlations were kept constant at 0.20. In correlation structures four, five and six, these correlations were reduced to 0.10. In comparison, SWDA was much more successful under correlation structures four, five and six. The procedure worked well for many scenarios when $p$ was equal to 2 or 3, and it also worked well under certain conditions for $p$ as high as 7 (see Table 3). When $p$ was 2 or 3, alternating between correlation structures four, five and six produced almost identical results (see Table 3).

When $p$ was equal to 2 or 3, SWDA worked well for small effect sizes, $d = 0.20$ and $\alpha = 0.05$. Power was equal to 0.89 and Type I error was 0.06. For $\alpha = 0.01$ power decreased to 0.71, and for $\alpha = 0.10$ Type I error increased to 0.11. For medium and large effect sizes, $d = 0.50$ and $d = 0.80$ respectively, SWDA worked well if $\alpha = 0.01$. Power was equal to 1.00 in both cases, and Type I error was 0.04 and 0.08 respectively (see Table 3).

### Table 3: Power and Type I Error for $\alpha = 0.01$, $n = 500$, Across-Set $\rho = 0.1$

<table>
<thead>
<tr>
<th>$p$</th>
<th>Within-Set $\rho = 0.2$</th>
<th>Within-Set $\rho = 0.4$</th>
<th>Within-Set $\rho = 0.6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d$</td>
<td>$d$</td>
<td>$d$</td>
</tr>
<tr>
<td></td>
<td>0.2 0.5 0.8</td>
<td>0.2 0.5 0.8</td>
<td>0.2 0.5 0.8</td>
</tr>
<tr>
<td>2</td>
<td>0.7158 1.0000 1.0000</td>
<td>0.7234 1.0000 1.0000</td>
<td>0.7198 1.0000 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0116 0.0354 0.0804</td>
<td>0.0122 0.0354 0.0824</td>
<td>0.0122 0.0392 0.0786</td>
</tr>
<tr>
<td>3</td>
<td>0.7128 1.0000 1.0000</td>
<td>0.7292 1.0000 1.0000</td>
<td>0.7222 1.0000 1.0000</td>
</tr>
<tr>
<td></td>
<td>0.0143 0.0353 0.0737</td>
<td>0.0123 0.0331 0.0702</td>
<td>0.0110 0.0308 0.0624</td>
</tr>
<tr>
<td>4</td>
<td>0.5581 0.9999 1.0000</td>
<td>0.4551 0.9921 1.0000</td>
<td>0.4197 0.8966 0.9995</td>
</tr>
<tr>
<td></td>
<td>0.0159 0.0857 0.2103</td>
<td>0.0126 0.0662 0.1443</td>
<td>0.0108 0.0444 0.0972</td>
</tr>
<tr>
<td>5</td>
<td>0.5621 0.9999 1.0000</td>
<td>0.4563 0.9916 1.0000</td>
<td>0.4519 0.8933 0.9995</td>
</tr>
<tr>
<td></td>
<td>0.0153 0.0796 0.1884</td>
<td>0.0138 0.0578 0.1194</td>
<td>0.0124 0.0380 0.0840</td>
</tr>
<tr>
<td>6</td>
<td>0.4601 0.9963 1.0000</td>
<td>0.3390 0.8811 0.9991</td>
<td>0.3002 0.6499 0.9062</td>
</tr>
<tr>
<td></td>
<td>0.0179 0.1391 0.2933</td>
<td>0.0135 0.0689 0.1622</td>
<td>0.0138 0.0402 0.0986</td>
</tr>
<tr>
<td>7</td>
<td>0.4641 0.9961 1.0000</td>
<td>0.3424 0.8791 0.9987</td>
<td>0.2991 0.6527 0.9120</td>
</tr>
<tr>
<td></td>
<td>0.0188 0.1229 0.2564</td>
<td>0.0134 0.0654 0.1413</td>
<td>0.0118 0.0365 0.0813</td>
</tr>
<tr>
<td>8</td>
<td>0.4025 0.9746 0.9999</td>
<td>0.2757 0.7325 0.9664</td>
<td>0.2305 0.5070 0.7395</td>
</tr>
<tr>
<td></td>
<td>0.0222 0.1646 0.3248</td>
<td>0.0146 0.0683 0.1553</td>
<td>0.0128 0.0377 0.0869</td>
</tr>
</tbody>
</table>
EFFECTIVENESS OF SWDA AS A POST HOC SIGNIFICANT MANOVA PROCEDURE

For values of $p$ greater than 3, alternating between correlation structures four, five and six begins to make a difference. For correlation structure four, the within-set correlations were set equal to 0.20. For correlation structures five and six, these correlations were increased to 0.40 and 0.60 respectively. SWDA worked well for $p = 4$ or $5$ when $d = 0.50$ and $\alpha = 0.01$ (see Table 3). Power and Type I error values were very similar for both $p = 4$ or $5$, but were different for different correlation structures. Under correlation structures four, five and six, power was equal to 0.9999, 0.9920, and 0.8950 respectively, and Type I error was equal to 0.08, 0.06, and 0.04 respectively.

For values of $p$ greater than 5, SWDA was effective in a couple of scenarios: for $p = 6$ or $7$, the procedure worked well for medium and large effect sizes when $\alpha = .01$. Power was around 0.88 and 0.90 respectively, and Type I error was around 0.07 and 0.09 respectively. Lowering the across-set correlations from 0.20 to 0.10 appeared to improve the effectiveness of SWDA, specifically with respect to Type I error. However, even with the across-set correlations reduced, SWDA still appeared to enjoy limited success when values of $p$ increased above 3.

Effect of Independent Variables on Power and Type I Error

$p$ – The Number of MANOVA Dependent Variables

SWDA appeared to become less effective as the number of MANOVA dependent variables increased. Generally, as $p$ increased, the power decreased and Type I error increased. Power and Type I error tended to be very similar when results are grouped by $p = 2$ or $3$, then by $p = 4$ or $5$, by $p = 6$ or $7$ and finally by $p = 8$. It should be noted that for each of these groupings, the number of variables with means that differed between the two groups is the same. Satisfactory results were usually obtained for only $p = 2$ or $3$, this may be largely due to having only one variable whose mean is different between the two groups. Satisfactory results might still be obtained for values of $p$ greater than 3, as long as only one of the variables has a mean that differs between the groups.

When the sample size was large, especially if the within-set correlation was low, SWDA became too aggressive resulting in Type I errors that were too high. This problem was exacerbated as $p$ increased. In some cases, Type I error increased from 0.30 to 0.80 as $p$ increased from 2 to 8. These results support the claim by Thompson (1995) that stepwise methods tend to increase the likelihood of Type I errors, especially for larger values of $p$. Thompson suggests that because several variables are considered for entry at each step, more degrees of freedom should be charged to the numerator from the denominator of the $F$-statistic. This technique will produce a smaller value for the $F$-statistic, making Type I errors less likely. However, Thompson mentions as a caveat that this outcome is less likely to be an issue when the number of dependent variables is small.

Less favorable results regarding power were also observed when $p$ increased. The $F$-statistic used in SWDA is described by Klecka (1980) as “the $F$-to-enter is a partial multivariate $F$-statistic which tests the additional discrimination introduced by the variable being considered after taking into account the discrimination achieved by the other variables already entered (Dixon, 1973, p. 241)” (p. 57). For certain variables, when only the additional contribution to discrimination is considered, problems can arise if these variables share information with other variables that are already in the model. “…two or more of the variables may share the same discriminating information even though individually they are good discriminators. When some of these are employed in the analysis, the remainder are redundant” (Klecka, p. 52).

For this study, if multiple variables differ between the two groups the power can be reduced if SWDA considers one or more of these variables as redundant, thus, when $p$ increased, the number of variables that differed between the two groups also increased. With respect to power in SWDA, it could be that increasing $p$ by itself does not reduce power, but increasing the number of variables whose means differ between the two groups does reduce power because SWDA may consider some of these to be redundant. This effect can be observed in Tables 1, 2 and 3.
n – Sample Size

Results based on sample size were as expected: as n increased, both power and Type I error increased as well. Unfortunately, SWDA appears to be too aggressive when the sample size gets large. Under correlation structures one, two or three, when n was 250 or 500, Type I error was too high except under certain conditions. High enough power was not an issue when n got large, but in order to keep Type I error below 0.10 the effect size needed to be small (d = 0.20) and α = 0.01. When the across-set correlation was reduced from 0.20 to 0.10, the Type I error rate was controlled much better (see Table 3). For correlation structures four, five and six, there were situations for medium and large effect sizes, as well as small effect sizes, where the Type I error stayed below 0.10. Lower levels of across-set correlation enables SWDA to perform more efficiently but caution should be used by the researcher when using SWDA with large sample sizes; at the very least, small levels of α should be used in this situation.

d – Effect Size

As expected, when effect size increased, power increased. This pattern was observed regardless of sample size, but was more apparent with smaller values of n. When the sample size became large the power of SWDA was high even for small effect sizes. Discrepancy in power for different effect sizes can be observed in Table 3.

Surprisingly, Type I error increased as well as power when effect size increased. It was believed that with higher effect sizes it would be easier for SWDA to distinguish between variables with means that differed between the groups and variables with means that were the same in both groups. However, this outcome was not the case and the pattern became even more apparent as n and p increased. This pattern is shown when comparing Tables 1 and 2. In some cases, for large n, large p and large d, Type I errors in excess of 0.90 were observed - SWDA becomes more aggressive as effect size increases.

The only connection between variables whose means are different in the two groups and variables whose means are the same in the two groups is the across-set correlation. Increasing the effect size did nothing to change the across-set correlation, but when the effect size became larger, a variable with the same mean in both groups was now correlated with a variable whose mean had an even larger difference between the two groups. This relationship appeared to increase the likelihood of the variable with the same mean in both groups, being incorrectly identified by SWDA.

To examine this relationship further, additional simulations were run at n = 500 and with the across-set correlation reduced to 0.10. Results for these scenarios (see Table 3) show that the same pattern was still observed. As effect size increased, the likelihood of Type I error increased as well. However, the Type I error rate was reduced significantly under correlation structures four, five and six. With the across-set correlation reduced from 0.20 to 0.10, a variable with the same mean in both groups now had a smaller correlation with a variable whose mean differed between the two groups. When the effect size was increased, therefore, the variable with the same mean in both groups was less likely to be incorrectly identified by SWDA.

It is difficult to explain why this happens in SWDA, but it appears that the across-set correlation is the key. Apparently, when a variable with the same mean in both groups is correlated to a degree with a variable with a high level of discriminatory power, SWDA has a tendency to select both variables. There appears to be a guilty-by-association factor present. The likelihood of incorrectly selecting the variable with the same mean in both groups increases as the correlation between the two variables increases.

ρ – Correlation

Within-set correlations varied among levels 0.20 (correlation structures one and four), 0.40 (correlation structures two and five) and 0.60 (correlation structures three and six). With all other independent variables held constant, as the within-set correlations increased, power and Type I error both decreased. This result indicates that SWDA becomes more conservative as correlations among MANOVA dependent variables increases; this pattern became more apparent as p increased. When one variable with
EFFECTIVENESS OF SWDA AS A POST HOC SIGNIFICANT MANOVA PROCEDURE

means that differed between the groups had been correctly selected by SWDA, the likelihood of selecting another variable with means that differed between the groups went down as the correlation between these two variables increased. The higher the correlation between these two variables, the less unique discriminatory information was offered by the second variable.

The same pattern was observed among variables with the same mean in both groups. Once one of these variables had been incorrectly selected by SWDA, the likelihood of incorrectly selecting a second variable went down as the correlation between the two variables increased. Again with higher correlation between these two variables, any imagined discriminatory information detected by SWDA, appeared to be redundant for the second variable.

Across-set correlations varied among levels 0.20 (correlation structures one, two and three) and 0.10 (correlation structures four, five and six). As across-set correlations increased, the likelihood of Type I error also increased. For a variable with the same mean in both groups, any correlation it shared with a variable with means that differed between the groups, made it more likely to be incorrectly selected by SWDA (this outcome is the same guilty-by-association factor previously mentioned).

A final observation was made on the effect of correlations among MANOVA dependent variables on SWDA due to a programming error early in the simulation process. The error in the simulations produced correlation matrices that were identity matrices so that all MANOVA dependent variables were statistically independent. The results for power and Type I error were very good using SWDA in this context. It should be noted that complete statistical independence between all dependent variables is not a realistic correlation structure, but it gives a little more insight into the effectiveness of SWDA as a post hoc procedure to MANOVA. For situations in which there is little correlation among the MANOVA dependent variables, SWDA may be an effective post hoc procedure to a significant MANOVA. The sample correlation matrix can help researchers estimate the level of correlations among the dependent variables.

\( \alpha \) – Level of Significance

As expected, when \( \alpha \) increased, power and Type I error increased as well. For small \( n \) and small \( d \), observed values of Type I errors were very close to the set level of \( \alpha \). This relationship was consistent regardless of \( p \) or the level of correlation among the MANOVA dependent variables. However, as \( n \) and/or \( d \) increased, the observed value of Type I error tended to increase to well above the set level of \( \alpha \). In some extreme cases the observed Type I error exceeded 0.90 and Type I error values in the 0.40 to 0.50 range were commonplace for large values of \( n \) or \( d \).

Inflated Type I error levels were expected in this study but the actual inflation in the Type I error rates were much larger than expected. Rencher and Larson (1980) observed that the F-statistic used in SWDA is biased towards including variables that should not be selected. However, Rencher and Larson only considered the case where the MANOVA null hypothesis was true. In this study, the MANOVA null hypothesis was false, therefore it was expected that Type I errors would not be drastically inflated since SWDA wasn’t fishing for significant results. Inflated Type I errors were observed, however, suggesting that researchers using SWDA should set \( \alpha \) to lower than desired values of Type I error, especially for larger sample sizes (\( n = 250 \) or 500).

Conclusion

Although SWDA appears to be a very powerful procedure, it seems to be too aggressive in general. The biggest issue in this study was inflated Type I error; researchers who are using SWDA need to be aware of this problem. However, researchers may be able to use the procedure quite successfully under certain conditions. First, researchers should keep the number of dependent variables small, probably no more than three or four according to this study. Secondly, SWDA will be most successful when the correlations among the dependent variables are small. This condition is very important and researchers should check the sample correlation matrix before using SWDA. Finally, researchers may be able to interpret the order in which variables are selected, albeit with some caution. Although there is no empirical
In this study, it was observed that when SWDA became too aggressive and selected too many variables, the variables with means that differed between the groups were generally selected first. If researchers are aware of this pattern, they can compare sample mean vectors on the variables that were selected later by SWDA, and make some tentative conclusions on the discriminatory power of these variables.

Inflation of Type I error was a serious issue in this study when sample size increased. Because the order in which the variables were selected was generally correct, future researchers should look for ways to make SWDA stop in time, especially for larger sample sizes. One possible solution would be to use the squared partial correlation criterion, rather than the F-test criterion used in this study. The squared partial correlation criterion and the F-test criterion select variables in the same order, but the F-test criterion tends to select more variables as the sample size increases (SAS Institute Inc., 2004). Future researchers can also conduct simulations using Thompson’s (1995) adjustment for degrees of freedom to determine how well this method controls Type I error.

Another possibility could be to make a Bonferroni-type adjustment to the $\alpha$-level that is used to select the significant variables, similar to what Hawkins (1976) advocated with his stepwise MANOVA procedure. When SWDA is used in this context, it can be reasonably viewed as a multiple comparison-type procedure, similar to how the Scheffe’ and Bonferroni procedures are used as a follow-up to a significant ANOVA. In that context, it is common practice to adjust the significance level for each of the multiple follow-up tests to control the family-wise error rate.

Because SWDA is also performing multiple tests on several variables at each step of the selection process, using some type of adjustment for each test at each step would seem like a reasonable step to take. This study did not address the utility of making a Bonferroni-type adjustment, so further research would be needed in order to determine the effectiveness of doing so, as well as how much of an adjustment to the $\alpha$-level for each test would be needed to control the overall Type I error rate at the nominal level.

References


Assessing Classification Bias in Latent Class Analysis: Comparing Resubstitution and Leave-One-Out Methods

Marc H. Kroopnick  Jinsong Chen  Jaehwa Choi  C. Mitchell Dayton
Association of American Medical Colleges  The George Washington University  University of Maryland, College Park

This Monte Carlo simulation study assessed the degree of classification success associated with resubstitution methods in latent class analysis (LCA) and compared those results to those of the leave-one-out (L-O-O) method for computing classification success. Specifically, this study considered a latent class model with two classes, dichotomous manifest variables, restricted conditional probabilities for each latent class and relatively small sample sizes. The performance of resubstitution and L-O-O methods on the lambda classification index was assessed by examining the degree of bias.

Key words: Resubstitution methods, multivariate classification, latent class analysis, leave-one-out, lambda classification index.

Introduction
Classifying individuals into groups is a popular multivariate technique, methods for which include: logistic regression analysis and discriminant function analysis with manifest group membership and cluster analysis and latent class analysis (LCA) with latent group membership (Everitt, Landau & Leese, 2001). Measures of classification success, however, can be biased in the positive direction because the data used for model estimation are also used to evaluate the success of classification (Hand, 1986). Measures of classification success based on the same data used to fit the model are referred to as resubstitution measures (Huberty, 1994; Clancy, 1997). The leave-one-out method (L-O-O), initially proposed by Lachenbruch (1967) to obtain approximately unbiased classification success measures, may be a viable alternative to the resubstitution method. Huberty (1994) also provides an illustration of the L-O-O method compared to other methods in the context of discriminant function analysis.

Two common measures for classification success in LCA are proportion correctly classified, $P_c$, and the statistic, $\lambda$ (lambda), which adjusts $P_c$ for chance level classification into the largest latent class (Goodman & Kruskall, 1954). Investigation of this bias in small samples sizes was suggested in Dayton (1998) but has yet to be widely addressed in the latent class literature. In order to assess the degree of bias, the traditional resubstitution computation of $\lambda$ and the $\lambda$ computed using the L-O-O method were compared to a theoretical value for $\lambda$.

Latent Class Analysis
Latent Class Analysis (LCA) is a statistical technique for multivariate categorical data that is used to discover subtypes of
individuals or to confirm hypothesized subtypes of individuals (see Dayton, 1998, for more latent class model details). LCA is useful for: (1) estimating latent class proportions (class sizes) for two or more latent classes and conditional probabilities for the manifest variables; and (2) assigning individuals to the latent classes using Bayes’ theorem. An example of LCA is locating distinctive cognitive diagnostic categories from examinees’ answers to achievement test items in an educational context. Subsequently, Bayes’ theorem can be used to assign examinees to the diagnostic categories that are most likely based on their observed responses.

Theoretical Framework

Successful classification of individuals into latent classes is a fundamental component to LCA. Following Dayton (1998), Bayes’ theorem is used to determine the posterior probability of membership in each latent class, \( t \), given a specific response vector, \( y_s \):

\[
P(t \mid y_s) = \frac{P(t \mid y_s) \times \pi_t^X}{\sum P(t \mid y_s) \times \pi_t^X}
\]

(1)

where \( \pi_t \) is the latent class proportion, \( X \) is the latent variable with levels (classes) \( t \in T \), and \( \sum P(t \mid y_s) \times \pi_t^X \) is the unconditional (across all latent classes) probability for the response vector \( y_s \). All individuals with the same response pattern are classified into the latent class, \( t \), with the largest posterior probability corresponding to its response vector, \( y_s \). The following formula expresses the proportion correctly classified, \( P_c \):

\[
P_c = \frac{\sum n_s \times \max P(t \mid y_s)}{N}
\]

(2)

where \( \max P(t \mid y_s) \) is the largest posterior probability for response \( y_s \) across all latent classes \( T \), \( n_s \) is the number of cases corresponding to the response vector \( y_s \), and \( N \) is the total number of cases. Note that the number of possible response vectors is \( 2^v \), where \( v \) is the number of manifest variables; thus, \( 2^v \) elements would be in the summation at the population level and, for sample based analyses, up to \( 2^v \) elements.

Chance level of correct classification, which is maximized by classifying all cases into the largest latent class, is not accounted for in \( P_c \). Goodman and Kruskall (1954) developed the \( \lambda \) (lambda) statistic as an adjusted value of \( P_c \).

\[
\lambda = \frac{P_c - \pi_M^X}{1 - \pi_M^X},
\]

(3)

where \( \pi_M^X \) represents the largest latent class proportion.

Considering that the parameter estimation and classification success for the latent class model are based on the same data (i.e., resubstitution), Dayton (1998) noted that values for \( P_c \) and \( \lambda \) tend to be biased upward (more so with small sample sizes) and that research investigating the magnitude and methods to correct for this have yet to be studied in great detail; thus, this provided the motivation for this study. Work by Dias and Vermut (2006), however, used bootstrapping techniques to assess classification uncertainty in LCA. Their research brought to light the risk of using traditional resubstitution methods, especially at the individual response vector level.

The Leave-One-Out Method

A so-called jackknife method for determining an unbiased estimate for classification accuracy was developed by Lachenbruch (1967). His study focused on discriminant analysis and his method has been named the leave-one-out (L-O-O) method (Huberty, 1994). This method involves two basic steps. First, the model is estimated in the sample with one observation deleted, and then the resulting parameter estimates are used to classify the single deleted observation. This process was carried out \( N \) times so that each observation was deleted and classified. Consequently, the measure of successful classification is the proportion of times that the deleted observation was correctly classified (Huberty, 1994).

In order to investigate the bias reduction property of the L-O-O method, Lachenbruch
(1967) conducted a small Monte Carlo simulation study with 300 replications for a two group discriminant analysis. The proportions of correct classifications according to both the resubstitution and L-O-O methods were calculated and empirical 95% confidence intervals (CIs) were obtained for those proportions. The CIs for the L-O-O method contained the true population value 93.3% of the time and the resubstitution method contained the true value 84.7% of the time. These results suggested the appropriateness and usefulness of Lachenbruch’s L-O-O technique. Lachenbruch’s procedure, with modifications, was employed in this LCA study, which involved a greater number of replications.

Methodology
Simulation Conditions
This study considered a latent class model with two classes, dichotomous manifest variables, restricted conditional probabilities for each latent class and relatively small sample sizes. The number of manifest variables considered was 4 and 6; this was purposefully small due to the small sample size focus of the study and the computation complexity associated with additional variables. Sample size varied in three ways based on the number of manifest variables.

Simulation sample sizes were 3, 5, or 7 times the number of possible response vectors. For example, applying the first weight, 3, to the four variable case yields a sample size of \(3 \times 2^4 = 48\). The latent class proportions and conditional probabilities for responses to the manifest variables followed a structure similar to that used in Holt and Macready (1989). The first set of latent class proportions had no discrepancy (.5, .5), and the second set had a large discrepancy, (.8, .2).

Three sets of conditional probabilities were tested; the first set had a small disparity (.7, .4), the second set had a moderate disparity (.8, .3) and the last had larger disparity (.9, .05). The first number in the set corresponded to the conditional probability of a positive response to all items for the larger latent class (if there was one) and the second number applied to the smaller latent class (if there was one). Thus, the conditional probabilities were homogeneous across manifest variables within each latent class. In sum, this simulation included the following number of cells: 2 (number of variables) * 3 (sample size cases) * 2 (latent class proportions) * 3 (conditional probability sets) for a total of 36 simulation conditions.

Data Generation and LCA Parameter Estimation
Monte Carlo simulation methods were used to generate data consistent with the parameters described above. MATLAB (The MathWorks Inc., 2007) was used to conduct the simulation. Following guidelines in Holt & Macready (1989), there were 500 replications per cell. The flexible Expectation-Maximization (EM) (Dempster, Laird & Rubin, 1977; McLachlan & Krishnan, 1997) algorithm was programmed in MATALB to provide the maximum-likelihood estimates (MLE) of the parameters in the latent class model. The iterative EM algorithm is a popular parameter estimation technique in LCA because there is no closed form formulation for their MLE computation (Dayton, 1998). It is the default estimation method in LEM (Vermut, 1997) or Mplus (Muthén & Muthén, 2004) and, typically, LEM or MPlus would be the program of choice, but MATLAB offers more advanced and useful data manipulation options. The accuracy of the custom MATLAB code was compared the estimates obtained in Mplus.

Resubstitution and L-O-O Methods for Lambda Computation
The performance of resubstitution and L-O-O methods on the lambda (\(\lambda\)) classification index was assessed by examining the degree of bias. Thus, for each replication in each simulation cell, the L-O-O and resubstitution lambda was computed and compared to the theoretical \(\lambda\) value. The calculation of the sample based resubstitution \(P_1\) and \(\lambda^*\) followed equations (2) and (3), respectively, but used the MLE parameter estimates obtained from the LCA estimation from the sample data associated with each replication in each cell.

The L-O-O method calculation was conducted in a similar fashion to that of the Lachenbruch (1967) simulation study, but was modified for LCA. A description of this
procedure is: For each response vector from the generated sample data, each unique response vector was deleted and the parameters re-estimated. The max $P'(t|y_s)$ for the deleted response vector, $y_s$, was determined according to equation (1), but based on the re-estimated parameters from the $N – 1$ cases. The deleted response vector was placed back in the data set and the process was repeated for the next unique response vector.

After this process, each of the (up to) $2^v$ max $P'(t|y_s)$ values was weighted by the appropriate $n_s$, summed, and divided by $N$ (equation 2); essentially this is a jackknifed $P'_c$, which will be called $P'_c^*$. Alternately the equivalent procedure (described above) could be conducted by deleting each case instead of each unique response vector and equally weighting the max $P'(t|y_s)$ associated with each deleted case. The latter was performed for this study. Note that the L-O-O method based estimate for this index requires $N$ estimations and the possibility exists for not getting a converged solution during each of the $N$ estimations. If the estimation associated with a given deleted case failed to converge, the case was eliminated from the analysis and $N$ was adjusted accordingly.

This value appeared in the numerator of the L-O-O method lambda, which will be called $\lambda^*$. The maximum latent class proportion estimate used to compute $\lambda'$ was also used to compute $\lambda^*$. This provided a means by which to be able to directly compare the degree of classification success above the chance success of classifying all simulees in the largest estimated latent class proportion based on the entire dataset, $\pi^X_M$. The formula for $\lambda^*$ is:

$$\lambda^* = \frac{(P'_c - \pi^X_M)}{(1 - \pi^X_M)}.$$  \hspace{1cm} (4)

Simulation Study Outcomes

The two outcome measures evaluated were the degree of bias and the performance of 95% confidence intervals based on $\lambda'$ and $\lambda^*$ in capturing the true value, $\lambda$. The true value, $\lambda$, was computed by applying the true population generating parameters to equations (1), (2) and (3). First, to evaluate the bias of $\lambda'$ and $\lambda^*$, the mean of the estimates, $M$, was computed and compared to the theoretical value for lambda. The percent difference between each mean and corresponding $\lambda$ was reported.

Second, within each cell, up to 500 (depending on the number of converged solutions) 95% CIs were computed for each $\lambda'$ and $\lambda^*$. As noted, for the L-O-O method, an estimate of $\lambda'$ is treated as a converged solution unless the $N$ estimations do not converge while there is only one estimation required to obtain $\lambda^*$, the resubstitution value. The method for CI construction was based on the method for computing proportion CIs developed by Wilson (1927) and further described by Newcombe (1998). The computation of the interval is as follows:

$$\frac{2np + z^2 \pm \sqrt{z^2 + 4npq}}{2(n + z^2)},$$  \hspace{1cm} (6)

where $p$ is the lambda value, $q$ is 1-$p$, $n$ is the sample size for the given cell, and $z$ is 1.96. The degree of bias was measured by subtracting the proportion of times the two types (resubstitution and L-O-O) of CIs contained the theoretical $\lambda$ from 95%. Note that both of these measures are reasonable methods, but not necessarily the only ways, to assess the performance of the two methods in terms of bias (i.e., comparing the observed to statistic to truth).

Results

The simulation outcome measures described above are summarized in Tables 1 and 2 for the 4 and 6 variables cases, respectively. Note that, except for the confidence interval coverage for one cell of the study, the difference between both simulation outcome measures associated with resubstitution and L-O-O methods was very small; i.e., less than .02 in absolute value. Figures 1 and 3 provide a graphical display of the outcome measures for the 4 variable case and Figures 2 and 4 provide a graphical display for the six variable case. While the results for the resubstitution and L-O-O methods mirrored each other, trends emerged from the various factors manipulated.
Table 1: Simulation Results when $v = 4$

<table>
<thead>
<tr>
<th>$N$</th>
<th>LC Max</th>
<th>Cond. Prob.</th>
<th>%RE</th>
<th>%LOO</th>
<th>.95-%RE</th>
<th>.95-%LOO</th>
<th>$M_{RE} - \lambda$</th>
<th>$M_{LOO} - \lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>0.500</td>
<td>(.7,.4)</td>
<td>0.330</td>
<td>0.274</td>
<td>0.620</td>
<td>0.676</td>
<td>0.204</td>
<td>0.200</td>
</tr>
<tr>
<td>48</td>
<td>0.500</td>
<td>(.8,.3)</td>
<td>0.682</td>
<td>0.680</td>
<td>0.268</td>
<td>0.270</td>
<td>0.066</td>
<td>0.068</td>
</tr>
<tr>
<td>48</td>
<td>0.500</td>
<td>(.9,.05)</td>
<td>0.992</td>
<td>0.994</td>
<td>-0.042</td>
<td>-0.044</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>48</td>
<td>0.800</td>
<td>(.7,.4)</td>
<td>0.010</td>
<td>0.018</td>
<td>0.940</td>
<td>0.932</td>
<td>0.531</td>
<td>0.526</td>
</tr>
<tr>
<td>48</td>
<td>0.800</td>
<td>(.8,.3)</td>
<td>0.164</td>
<td>0.154</td>
<td>0.786</td>
<td>0.796</td>
<td>0.242</td>
<td>0.233</td>
</tr>
<tr>
<td>48</td>
<td>0.800</td>
<td>(.9,.05)</td>
<td>0.964</td>
<td>0.962</td>
<td>-0.014</td>
<td>-0.012</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>80</td>
<td>0.500</td>
<td>(.7,.4)</td>
<td>0.356</td>
<td>0.344</td>
<td>0.594</td>
<td>0.606</td>
<td>0.091</td>
<td>0.094</td>
</tr>
<tr>
<td>80</td>
<td>0.500</td>
<td>(.8,.3)</td>
<td>0.728</td>
<td>0.736</td>
<td>0.222</td>
<td>0.214</td>
<td>0.027</td>
<td>0.028</td>
</tr>
<tr>
<td>80</td>
<td>0.500</td>
<td>(.9,.05)</td>
<td>0.986</td>
<td>0.986</td>
<td>-0.036</td>
<td>-0.036</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>80</td>
<td>0.800</td>
<td>(.7,.4)</td>
<td>0.032</td>
<td>0.036</td>
<td>0.918</td>
<td>0.914</td>
<td>0.419</td>
<td>0.413</td>
</tr>
<tr>
<td>80</td>
<td>0.800</td>
<td>(.8,.3)</td>
<td>0.262</td>
<td>0.248</td>
<td>0.688</td>
<td>0.702</td>
<td>0.164</td>
<td>0.163</td>
</tr>
<tr>
<td>80</td>
<td>0.800</td>
<td>(.9,.05)</td>
<td>0.930</td>
<td>0.930</td>
<td>0.020</td>
<td>0.020</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>112</td>
<td>0.500</td>
<td>(.7,.4)</td>
<td>0.360</td>
<td>0.344</td>
<td>0.590</td>
<td>0.606</td>
<td>0.016</td>
<td>0.016</td>
</tr>
<tr>
<td>112</td>
<td>0.500</td>
<td>(.8,.3)</td>
<td>0.756</td>
<td>0.758</td>
<td>0.194</td>
<td>0.192</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>112</td>
<td>0.500</td>
<td>(.9,.05)</td>
<td>0.982</td>
<td>0.982</td>
<td>-0.032</td>
<td>-0.032</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>112</td>
<td>0.800</td>
<td>(.7,.4)</td>
<td>0.042</td>
<td>0.050</td>
<td>0.908</td>
<td>0.900</td>
<td>0.317</td>
<td>0.312</td>
</tr>
<tr>
<td>112</td>
<td>0.800</td>
<td>(.8,.3)</td>
<td>0.356</td>
<td>0.356</td>
<td>0.594</td>
<td>0.594</td>
<td>0.117</td>
<td>0.116</td>
</tr>
<tr>
<td>112</td>
<td>0.800</td>
<td>(.9,.05)</td>
<td>0.928</td>
<td>0.926</td>
<td>0.022</td>
<td>0.024</td>
<td>0.003</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Note: LC MAX is the first latent class population proportion; Cond. Prob. is the population conditional probability for all responses; %RE is the percentage of the resubstitution method CIs containing $\lambda$; %RE is the percentage of the resubstitution method CIs containing $\lambda$; $M_{RE} - \lambda$ is the mean of the $\lambda$ estimates based on the resubstitution method; $M_{RE} - \lambda$ is the mean of the $\lambda$ estimates based on the L-O-O method.
Table 2: Simulation Results when $\nu = 6$

<table>
<thead>
<tr>
<th>$N$</th>
<th>LC Max</th>
<th>Cond. Prob.</th>
<th>%RE</th>
<th>%LOO</th>
<th>.95-%RE</th>
<th>.95-%LOO</th>
<th>$M_{RE} - \lambda$</th>
<th>$M_{LOO} - \lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>192</td>
<td>0.500</td>
<td>(.7,.4)</td>
<td>0.554</td>
<td>0.558</td>
<td>0.396</td>
<td>0.392</td>
<td>-0.021</td>
<td>-0.024</td>
</tr>
<tr>
<td>192</td>
<td>0.500</td>
<td>(.8,.3)</td>
<td>0.878</td>
<td>0.878</td>
<td>0.072</td>
<td>0.072</td>
<td>-0.011</td>
<td>-0.011</td>
</tr>
<tr>
<td>192</td>
<td>0.500</td>
<td>(.9,.05)</td>
<td>0.990</td>
<td>0.990</td>
<td>-0.040</td>
<td>-0.040</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>192</td>
<td>0.800</td>
<td>(.7,.4)</td>
<td>0.046</td>
<td>0.050</td>
<td>0.904</td>
<td>0.900</td>
<td>0.247</td>
<td>0.240</td>
</tr>
<tr>
<td>192</td>
<td>0.800</td>
<td>(.8,.3)</td>
<td>0.538</td>
<td>0.542</td>
<td>0.412</td>
<td>0.408</td>
<td>0.047</td>
<td>0.047</td>
</tr>
<tr>
<td>192</td>
<td>0.800</td>
<td>(.9,.05)</td>
<td>0.972</td>
<td>0.972</td>
<td>-0.022</td>
<td>-0.022</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>320</td>
<td>0.500</td>
<td>(.7,.4)</td>
<td>0.494</td>
<td>0.496</td>
<td>0.456</td>
<td>0.454</td>
<td>-0.054</td>
<td>-0.054</td>
</tr>
<tr>
<td>320</td>
<td>0.500</td>
<td>(.8,.3)</td>
<td>0.836</td>
<td>0.836</td>
<td>0.114</td>
<td>0.114</td>
<td>-0.011</td>
<td>-0.011</td>
</tr>
<tr>
<td>320</td>
<td>0.500</td>
<td>(.9,.05)</td>
<td>0.986</td>
<td>0.986</td>
<td>-0.036</td>
<td>-0.036</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>320</td>
<td>0.800</td>
<td>(.7,.4)</td>
<td>0.092</td>
<td>0.086</td>
<td>0.858</td>
<td>0.864</td>
<td>0.130</td>
<td>0.128</td>
</tr>
<tr>
<td>320</td>
<td>0.800</td>
<td>(.8,.3)</td>
<td>0.636</td>
<td>0.636</td>
<td>0.314</td>
<td>0.314</td>
<td>0.020</td>
<td>0.020</td>
</tr>
<tr>
<td>320</td>
<td>0.800</td>
<td>(.9,.05)</td>
<td>0.968</td>
<td>0.968</td>
<td>-0.018</td>
<td>-0.018</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>448</td>
<td>0.500</td>
<td>(.7,.4)</td>
<td>0.446</td>
<td>0.446</td>
<td>0.504</td>
<td>0.504</td>
<td>-0.053</td>
<td>-0.054</td>
</tr>
<tr>
<td>448</td>
<td>0.500</td>
<td>(.8,.3)</td>
<td>0.850</td>
<td>0.850</td>
<td>0.100</td>
<td>0.100</td>
<td>-0.010</td>
<td>-0.010</td>
</tr>
<tr>
<td>448</td>
<td>0.500</td>
<td>(.9,.05)</td>
<td>0.994</td>
<td>0.994</td>
<td>-0.044</td>
<td>-0.044</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>448</td>
<td>0.800</td>
<td>(.7,.4)</td>
<td>0.176</td>
<td>0.178</td>
<td>0.774</td>
<td>0.772</td>
<td>0.099</td>
<td>0.096</td>
</tr>
<tr>
<td>448</td>
<td>0.800</td>
<td>(.8,.3)</td>
<td>0.588</td>
<td>0.588</td>
<td>0.362</td>
<td>0.362</td>
<td>0.009</td>
<td>0.010</td>
</tr>
<tr>
<td>448</td>
<td>0.800</td>
<td>(.9,.05)</td>
<td>0.974</td>
<td>0.974</td>
<td>-0.024</td>
<td>-0.024</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: LC MAX is the first latent class population proportion; Cond. Prob. is the population conditional probability for all responses; %RE is the percentage of the resubstitution method CIs containing $\lambda$; %RE is the percentage of the resubstitution method CIs containing $\lambda$; $M_{RE}$ is the mean of the $\lambda$ estimates based on the resubstitution method; $M_{LOO}$ is the mean of the $\lambda$ estimates based on the L-O-O method.
Figure 1: 95%RE and 95%LOO over Conditional Probabilities when $\nu = 4$
Figure 2: .95 - %RE and 95 - %LOO over Conditional Probabilities when $v = 6$

![Graph showing .95 - %RE and 95 - %LOO over Conditional Probabilities](image-url)
Figure 3: RE and LOO BIAS over Conditional Probabilities when $\nu = 4$

- RE BIAS
  - Sample Size = 48 (.5,.5)
  - Sample Size = 48 (.8,.2)
  - Sample Size = 80 (.5,.5)
  - Sample Size = 80 (.8,.2)
  - Sample Size = 112 (.5,.5)
  - Sample Size = 112 (.8,.2)

- LOO BIAS
  - Sample Size = 48 (.5,.5)
  - Sample Size = 48 (.8,.2)
  - Sample Size = 80 (.5,.5)
  - Sample Size = 80 (.8,.2)
  - Sample Size = 112 (.5,.5)
  - Sample Size = 112 (.8,.2)
Figure 4: RE and LOO BIAS over Conditional Probabilities when \( v = 6 \)
Overall, results were largely consistent with expectations: Assessing classification accuracy improves with increasing samples size, larger numbers of variables, more discrepant conditional probabilities, and equal (i.e., less discrepant) latent class proportions. In terms of absolute numbers, the outcome measures from the simulation strongly suggested the best results across all other conditions occurred when the conditional probabilities were the most discrepant. In sum:

- Overall, more bias and less confidence interval coverage for the (.8, .2) latent class proportions resulted compared to the (.5, .5) latent class proportions.
- Overall, more bias and less confidence interval coverage for the 4 variable case was observed compared to the 6 variable case.
- For any given pair of latent class proportions, bias decreased and confidence interval coverage increased as sample size increased.
- For any given pair of latent class proportions, the variability of bias across sample sizes decreased as the discrepancy of conditional probabilities increased.
- For any given pair of latent class proportions, as the discrepancy of the conditional probabilities increased, the bias decreased and the confidence interval coverage increased.

Conclusion

The primary purpose of the study was to illustrate differences between L-O-O and resubstitution methods for assessing classification accuracy in latent class analysis. Simulation results indicated very little difference in the methods based on outcome measures. However, the accuracy measures did vary over the factors manipulated in this study and should provide researchers with a guide regarding what to expect in their studies. It is important to note that when the conditional probabilities were very discrepant, other factors had little influence and accuracy was high.

Generalizing beyond the factors and the scope of this study should be approached cautiously. As noted earlier, only a two class latent class model with restricted conditional probabilities and relatively small sample sizes was considered. Research comparing and evaluating these classification accuracy measures applied to more complicated latent class models, larger sample sizes and an increased number of variables is warranted. This research provides a baseline of possible outcomes when those future studies are conducted.

References


The Performance of Multiple Imputation for Likert-type Items with Missing Data

Walter Leite  
University of Florida  

S. Natasha Beretvas  
The University of Texas at Austin

The performance of multiple imputation (MI) for missing data in Likert-type items assuming multivariate normality was assessed using simulation methods. MI was robust to violations of continuity and normality. With 30% of missing data, MAR conditions resulted in negatively biased correlations. With 50% missingness, all results were negatively biased.

Key words: Multiple imputation, missing data, missing at random, Likert-type items

Introduction

Missing values introduce several problems to statistical analyses and researchers have tried many methods to ameliorate these problems. A few popular methods have become popular and have been implemented in statistical software, which has boosted their usage. These methods include listwise deletion, pairwise deletion, mean substitution, regression imputation, maximum-likelihood methods and multiple imputation. Among these procedures, multiple imputation (MI), together with maximum likelihood estimation, is becoming one of the preferred techniques for dealing with missing data; due to its increasing popularity, this study focuses on the performance of MI.

MI was first proposed by Rubin (1987) as a way to handle missing data in public survey datasets. Research about MI in the statistical literature is abundant, however, only a handful of studies apply MI to missing data in datasets consisting of Likert-type items. This may be partially explained by the fact that MI depends on the extensive use of computer software and specialized software has only recently become easily accessible.

The MI method that best fits a set of data depends on the distribution assumed for the variables in the dataset. MI is most often performed under the assumption that the variables are multivariate normally distributed; cases exist, however, where this assumption may not be appropriate. In particular, surveys or scales used in organizational research frequently contain dichotomous or Likert-type items whose responses are not normally distributed. Very little research has been done concerning missing data in Likert-type scales and there are no studies evaluating the use of MI under a multivariate normal model with ordinal variables. Although Schafer (1997) argued that MI under the multivariate normal model is robust to departures from normality, extensive investigation of this issue does not currently exist in the literature. Thus, the objective of this study is to examine the performance of MI with datasets composed of Likert-type items.

Types of Missing Data

The existence of missing data in a dataset can result in loss of statistical power and biased parameter estimates. Causes of missing values in data are varied, for example: the refusal of some subjects to answers certain questions, data-entry errors and attrition (Little & Rubin, 1989). Missing data can be classified...
according to its pattern within the dataset. Little and Rubin (1989) adopted four classifications for patterns of missing data: general pattern of missingness, univariate missing data, univariate non-response and monotone missing data. A general pattern of missingness is characterized when values are missing in many variables without any special arrangement. If the data are missing in just one variable of the dataset, the missing data are univariate. Unit non-response is a pattern where a block of variables has missing values for the same set of cases, but data for those cases for all other variables is complete. Monotone missing data describe a pattern where complete cases in a variable that has \( X \) missing values will also be complete in a variable that has \( (X - 1) \) values.

Whether a procedure to deal with missing data will result in unbiased estimates of parameters depends on the relationships between the missing values, the incomplete variable and the other variables in the dataset. These relationships allow classification of missing data into three types (Rubin, 1976; Little & Rubin, 1987; Sinharay, Stern & Russell, 2001) commonly referred to as missing data mechanisms: data missing completely at random (MCAR), data missing at random (MAR) and data missing not at random (MNAR) or non-ignorable missingness.

Data are MCAR for a variable \( X \) when the missing values in this variable are independent of both the variable \( X \) and the other variables in the dataset. In this case, the observed variables can be considered a random sub-sample of the hypothetical complete data. Missing values for a variable are considered MAR when they depend on the other variables in the dataset, but not on the variable itself. MNAR or non-ignorable missingness occurs when the probability of the missing values for a variable \( X \) is dependent on the underlying value of \( X \) (Little & Rubin, 1987; Sinharay, Stern & Russell, 2001).

Multiple Imputation

The most common procedure to deal with missing data is deleting cases with incomplete data, called listwise deletion. However, listwise deletion results in unbiased parameter estimates only when (1) the data can be assumed MCAR, and (2) the fraction of missing data is very small (e.g., 5%) (Graham & Hofer, 2000). Other methods, such as person and item mean imputation, hot-deck imputation (Huisman, 2000), regression imputation and the expectation maximization algorithm (EM) can be used with MAR data, but they reduce the variability of the dataset and produce artificially small standard errors (McDonald, Thurston & Nelson, 2000) Among the many procedures that have been developed to cope with missing data, full-information maximum likelihood estimation and multiple imputation (MI) are the most sophisticated methods, and are also the ones likely to yield the least biased parameter estimates (Sinharay, Stern & Russell, 2001; Graham & Hofer, 2000).

The results of a missing-data procedure are affected by the type of missingness (MCAR, MAR or MNAR) and also by characteristics of the sample and variables being analyzed. These characteristics include sample size, scale of measurement of the variables, range of data points and distribution of the observed variables. In the case where the dataset contains scores of a psychometric scale, the reliability and validity of scores on the instrument are also important (Raaijmakers, 1999).

The MI method consists of creating a vector of possible values for every missing value in the database. It represents a step forward from regression-based single imputation and the EM algorithm because the multiply imputed values reflect the uncertainty of the imputed values. MI restores two sources of variability: the variability of each variable and the variability of the sample covariance matrix. The variability of each variable is restored because the imputed values do not fall exactly on the regression line. This is accomplished by adding error variability to the imputed missing values. These errors are sampled from the distribution of known errors. The variability of the sample covariance matrix is restored by sampling many covariance matrices from a simualted population. Due to the restoration of these sources of variability, the resulting imputed values will include a component of within-imputation and a between-imputation variance.

Schafer (1997) developed methods to execute MI by cycling through two steps. In the
MULTIPLE IMPUTATION FOR LIKERT-TYPE ITEMS WITH MISSING DATA

first step, missing values are imputed, and in the second step unknown parameters are estimated. After the second step, the estimated parameters are used to impute missing values and the cycle is repeated until reaching a criterion of convergence. The process begins with an initial estimate of the parameters given by the expectation maximization (EM) algorithm. Schafer (1997) calls the first step of MI the I (imputation) step. It consists of replacing missing data points by randomly drawn values from the conditional distribution of the missing data given the observed data and the parameter estimates provided by the EM algorithm. The second step is termed P (posterior) and consists of estimating parameters. The estimated parameters are then used in another I step, and this process is repeated until the distribution of covariance matrices stops changing substantially. The EM algorithm is used to calculate the initial parameter estimates for the first imputation step. After this initial estimate, missing values are imputed and parameters are estimated using the data augmentation method.

Data augmentation is an iterative procedure that imputes missing data under assumed values of the parameters and then draws new parameters from a posterior distribution based on the complete data (Schafer & Olsen, 1998). This process of imputing values and estimating parameters creates a Markov chain. When the Markov chain stabilizes, the data augmentation process has reached convergence. This state is characterized by a stable distribution of parameters. After convergence, multiple imputations are generated based on independent draws from this distribution. Any number of imputed data-sets can be obtained by repeating the data augmentation algorithm; consequently, each set of imputed values will be different from the others.

MI has been shown to depend on three assumptions to generate unbiased parameter estimates. The first assumption specifies what types of missing data can be addressed using MI. The other two assumptions are necessary due to the Bayesian nature of MI. The first assumption of MI is that the data are MCAR or MAR. This assumption is important because using MI with MNAR data may result in biased parameter estimates (Little & Rubin, 1989). The second assumption is the prior distribution; because MI is a Bayesian method, a prior distribution is used to represent the state of knowledge about the data before it is available. Usually a non-informative prior (Sinha Ray, Stern & Russell, 2001) is chosen corresponding to ignorance about the distribution of the data. Such a prior is ambiguous as to the location of the likelihood’s maximum, allowing a wide range of values. In some cases, it is adequate to specify an informative prior distribution. This distribution is chosen from a family of distributions and it is combined with the likelihood to generate a posterior distribution from the same family. These assumptions are essential because multiple imputations involve random draws from the posterior probability distribution of the unknown parameters given the observed values.

Finally, MI requires an assumption about the complete-data model. Each multiple imputation method uses a specific probability model to generate the imputed values. The distribution of the observed values should match this imputation model. MI software usually uses the multivariate normal model to impute numeric data and the loglinear model for categorical data. The multivariate normal is the most common model for multivariate statistical analysis. Schafer (1997) argues that the normal model is robust to departures of normality when the proportion of missing data is not large. The reason for this robustness is that the model only affects the missing data, leaving the observed values unchanged. In addition, Schafer & Olsen (1998) indicate that it is often acceptable to impute values of categorical variables under the normality assumption and round off the continuous imputed values to the nearest category.

MI allows the researcher to improve the quality of the imputed values by using information from variables that predict the missing values or correlate with the variables containing missing values. These variables may be of no interest for the data analysis itself, therefore, they can be included in the dataset during the multiple imputation procedure and then excluded in the data analysis. The variables that may help with the imputation process can be detected through an examination of correlations.
and contingency tables between these variables and the variables that have missing values. However, the inclusion of an exaggerated number of variables may result in multicollinearity problems and variance inflation (Wayman & Swaim, 2002).

Five to ten imputations are typically recommended because this number has been found to provide adequate estimates (Rubin, 1987; Collins, Schafer & Kam, 2001). After multiple imputed datasets are obtained the analysis of interest to the researcher should be conducted with each imputed data set. For example, a researcher might be interested in conducting a multiple regression analysis. Assume the researcher obtained ten multiply imputed datasets containing imputations replacing each missing value. The researcher would run the regression analysis using each data set, and the resulting parameter estimates (the regression coefficient estimates, for example) can then be combined across the \( m = 10 \) imputed datasets to obtain the single best estimate of the relevant parameter (Rubin, 1987). Specifically, the mean of the parameter estimates across the \( m \) imputed datasets, \( \bar{q} \), can be calculated as:

\[
\bar{q} = \frac{1}{m} \sum_{i=1}^{m} \hat{q}_i
\]  

(1)

where \( \hat{q}_i \) is the parameter estimate from the \( i^{th} \) imputed dataset and \( m \) represents the number of imputed datasets being combined.

To calculate the variance of each parameter estimate, two sources of variability should be combined (Schafer & Olsen, 1998): the variability within and between imputed datasets. The within-imputation variance, \( \bar{u} \), is the mean of the variance estimates from each imputed dataset:

\[
\bar{u} = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_i
\]  

(2)

where \( \hat{u}_i \) is the variance estimate for the relevant parameter estimated for imputed dataset \( i \). In the example described in which multiple regression analyses were conducted with each of the ten imputed data sets, the square of the standard error estimated for one of the predictor’s unstandardized regression coefficients for imputed dataset \( i \) would provide that imputation’s \( \hat{u}_i \).

The between-imputations variance, \( B \), is the variance of the parameter estimates across the set of imputations:

\[
B = \frac{1}{(m-1)} \sum_{i=1}^{m} (\hat{q}_i - \bar{q})^2
\]  

(3)

where \( \hat{q}_i \) represents the parameter estimate for imputation \( i \).

The total variance, \( T \), associated with the multiply imputed parameter estimate, \( \bar{q} \), is the sum of the within- and the between-imputations variances. This sum is corrected to account for the simulation error in \( \bar{q} \) (Schafer & Olsen, 1998) using the formula:

\[
T = \bar{u} + \left( 1 + \frac{1}{m} \right) B
\]  

(4)

This total variance provides the advantage of MI over other methods for dealing with missing data. The within-imputations variance component represents sampling variability while the between-imputations variance represents missing data uncertainty. These two components prevent the missing values from creating an artificial precision in the parameter estimates, resulting in negatively biased standard errors and associated test statistic \( p \)-values that are too low (Schafer, 1997).

Recently, many computer programs have become available to perform MI (e.g., NORM, S-Plus, R, SAS). NORM 2.02 is a stand-alone multiple imputation program developed by Schafer (1999) that executes MI under the multivariate normal model. The freely-available R software (R development core team, 2008) contains the norm library, which is an implementation of MI similar to the NORM software. Different implementations of MI in the R software can be found in the CAT, Mix,
Amelia and Mice packages. S-Plus (Insightful Corp., 2001) has a library that performs MI under the Gaussian, loglinear and conditional Gaussian models. The statistical package SAS Version 8.2 incorporated functions for MI but it has the disadvantage of allowing little control over the imputation model (Horton & Lipsitz, 2001).

Multiple Imputation of Likert-Type Items

Little research has been conducted concerning missing data in Likert-type scales. For example, Downey and King (1998) investigated missing data in Likert-type variables but only evaluated mean substitution methods (person mean and item mean). Roth, Switzer and Switzer (1999) investigated missing data in multiple item scales, but only examined listwise deletion, regression imputation, hot-deck imputation, person mean substitution and item mean substitution.

MI has been most frequently conducted under the assumption that the variables are multivariate-normally distributed. However, surveys and scales commonly contain non-normally distributed Likert-type items, whose distributions may only approximate normality. Although Schafer (1997) developed a MI method for categorical data based on the loglinear model, he argued that multivariate normal MI could be used for categorical variables. However, evaluation of this claim has yet to be conducted. If MI, under the assumption of normality, works sufficiently well with typically non-normal Likert-type (ordinal) variables/items, the analysis of this type of data would be simplified.

Methodology

The performance of MI was assessed using simulation methods assuming multivariate normality in the commonly occurring scenario in which some of the responses to Likert-type items are missing. The impact of the following factors on the performance of MI were assessed: the underlying distribution of the item responses (normal versus non-normal), the magnitude of the variables’ inter-correlations ($\rho = 0.2, \rho = 0.8$), the bluntness of the categorization of the data into discrete item scores (3, 5 and 7), the missing data mechanism (MCAR and MAR) and the degree of missingness (10%, 30% and 50%). Recovery of the true correlations will be used in the evaluation of MI’s performance.

Responses to a set of 10 items were generated to fit either multivariate normal or non-normal distributions with a known correlational structure. To simplify the generating correlation matrix, each variable was modeled to have the same correlation with each of the others (0.8 or 0.2). Next, each interval-scaled item score was discretized to match the Likert-scale format of relevance to the condition (3, 5 or 7), and the condition’s pattern (MCAR or MAR) and degree of missingness were built into the generated data. Three degrees of missingness were investigated (10%, 30%, 50%) and MI was used to impute missing data. For each iteration (and condition), the imputed datasets were summarized using Equations 1 - 4 to assess recovery of the generating correlation values. Due to their importance in methods such as multiple regression and factor analysis, correlations were the parameters of interest in this study.

Simulation of Item Data

The software, S-Plus (Insightful, 2001) was used to conduct the simulation. To represent items on a 10-item scale or survey, 10 continuous random variables were generated with normal and non-normal distributions. Each variable was sampled from a multivariate normal distribution with a mean of zero and standard deviation of one. The multivariate normal random values were created using the function RMVNORM of S-Plus, which generates pseudo-random numbers given a correlation matrix, vector of means and standard deviations and a random seed.

The skewness and kurtosis was introduced into the data using the method originated by Valle and Maurelli (1983), which produces multivariate non-normal distributions with a given value of skewness and kurtosis by combining Kaiser and Dickman’s method (1962) with one proposed by Fleishman (1978) to simulate univariate non-normal distributions with specified degrees of skewness and kurtosis. Fleishman’s method uses the transformation $y_{ij} = x_{ij} + \kappa_{ij} + \lambda_{ij} x_{ij}$, where $x_{ij}$ are the original values, $\kappa_{ij}$ and $\lambda_{ij}$ are parameters determined by Fleishman’s method, and $y_{ij}$ are the transformed values. The skewness $\kappa$ and kurtosis $\lambda$ are determined by solving the following equations:

$$\kappa = \frac{E[(X-\mu)^3]}{\sigma^3}$$

$$\lambda = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

where $E$ denotes the expected value, $\mu$ is the mean, $\sigma$ is the standard deviation, and $X$ is the variable.

Responses to the set of 10 items were generated to fit either multivariate normal or non-normal distributions with a known skewness and kurtosis. Each item was discretized to match the Likert-scale format of relevance to the condition (3, 5 or 7), and the condition’s pattern (MCAR or MAR) and degree of missingness were built into the generated data. Three degrees of missingness were investigated (10%, 30%, 50%) and MI was used to impute missing data. For each iteration (and condition), the imputed datasets were summarized using Equations 1 - 4 to assess recovery of the generating correlation values. Due to their importance in methods such as multiple regression and factor analysis, correlations were the parameters of interest in this study.
\[ Y = a + bX + cX^2 + dX^3 \]

where \( a, b, c, \) and \( d \) are constants, to convert variable \( X \) into variable \( Y \) with the desired degree of skewness and kurtosis. Fleishman (1978) provides equations and tables detailing values for these constants along with their associated skewness and kurtosis levels. When applied together, Kaiser and Dickman’s and Fleishman’s methods interact such that the correlations between the simulated non-normal variables differ from those specified in the population correlation matrix. Vale & Maurelli (1983) solved this problem by adjusting the values of the population correlations using the formula:

\[
\begin{align*}
\rho_{y_1y_2} &= \rho_{x_1x_2} (b_1 b_2 + 3b_1 d_2 + 3d_1 b_2 + 9d_1 d_2) \\
&+ \rho_{x_1x_2}^2 (2c_1 c_2) + \rho_{x_1x_2}^3 (6d_1 d_2)
\end{align*}
\]  

(5)

where \( \rho_{x_1x_2} \) is the population correlation between variables \( X_1 \) and \( X_2 \), \( \rho_{y_1y_2} \) is the adjusted correlation between the non-normal variables \( Y_1 \) and \( Y_2 \), and \( b_1, b_2, c_1, c_2, d_1, d_2 \), are Fleishman’s coefficients for \( Y_1 \) and \( Y_2 \). After adjusting the population correlations, non-normal random variables are obtained by first executing Kaiser and Dickman’s method and then using Fleishman’s method. The resulting variables will have the desired degrees of skewness, kurtosis and inter-correlations. However, with Likert-type variables this method has the limitation that the transformation of continuous variables into categorical variables results in a slight change of the degrees of skewness and kurtosis originally simulated.

For each of the conditions, 1,000 samples of 400 cases were generated and the variables were converted into Likert-type scores. Datasets with three types of Likert-type items were created (with scales ranging from 1 to 3, 1 to 5, and 1 to 7) by dividing the total range of the scores into \( k \) segments of equal size, where \( k \) is the desired number of categories. This resulted in discrete distributions that better approximated the shape of their continuous, generating distributions. The correlation matrices for each replication sample and condition were also calculated to allow an assessment of the change resulting from the categorization process and to serve as a baseline for later evaluations.

Simulation of Missing Data

Two types of missing data were introduced: MCAR and MAR. Three overall proportions of missing values were simulated (10%, 30% and 50%). MCAR missing data was obtained through random deletion of values from the datasets. To simulate the MAR condition, one variable in the dataset, \( Z \), was used to predict the missing values in the other nine variables. The predictor \( Z \) was the only variable in the dataset with no missing values. Data points were deleted according to the MAR-linear condition described by Collins, Schafer and Kam (2001). In the MAR-linear condition (perhaps better described as monotonically increasing rather than linear), the proportion of missing values is approximately linearly related to the value of \( Z \). To simulate this condition, the cases were grouped according to the value of \( Z \), and subgroups of cases with larger values of \( Z \) were assigned a higher probability of being missing.

Analyses

Values for the missing data were imputed assuming the multivariate normal model using the functions of the missing library (Schimert, et al., 2000) implemented in S-PLUS version 6.0 (Insightful, 2001). Ten imputations were created for each dataset and the correlation between each pair of variables was calculated for each imputed data set. When correlation estimates are the unit of analysis, Fisher’s (1928) normalizing and variance-stabilizing \( r \)-to-\( Z \) transformation is frequently used to correct the non-normality of the sampling distribution of \( r \). This transformation was used; specifically, each correlation was transformed to a \( Z \) using the formula:

\[
Z_r = (1/2) \ln \left[ \frac{1 + r}{1 - r} \right].
\]  

(6)
These 10 transformed correlations (one per imputation) for each pair of variables were combined using Rubin’s (1987) rules as outlined in Equation 1 to provide an overall transformed correlation estimate, $\bar{q}$ combined across imputations for each sample and condition. (This was repeated for each of the transformed correlations between the variables). In addition, the between-imputations variance, $B$, of the transformed correlation estimates (see Equation 3) was also calculated for each multiply imputed estimate, $\bar{q}$.

The criterion used to judge the performance of MI involved an assessment of the recovery of the correlations (conducted using the transformed correlations). While the original generating value for the correlations was either 0.2 or 0.8, this value applied only to the continuous distributions. It should be noted that, for the non-normal distributions, although data were transformed to have a slight degree of kurtosis and skewness, the transformations were chosen to maintain the generating correlation values. However, the categorization of the continuously scaled scores into ordinal-scaled data resulted in correlations between pairs of variables that differed from the original generating values. The values of the correlations were compared after categorization - but before missingness had been introduced - with the correlations estimated after MI had been used to compensate for the missingness. The correlations after categorization were transformed using Fisher’s r-to-$Z_r$ transformation to provide the average of the sampling distribution of $Z_r$s for categorized variables. For each dataset simulated, the $Z_r$ values calculated after MI were compared with the values describing the categorized distributions without missingness.

The comparisons were performed using relative bias averaged across replications. The relative bias (Hoogland & Boomsma, 1998) compares the average value of the parameter estimated $Z_r$ with the population value, $\zeta_\rho$, using the formula:

$$B(\hat{Z}_r) = \frac{\bar{Z}_r - \zeta_\rho}{\zeta_\rho}.$$

The relative bias of the parameter estimate was considered acceptable if its magnitude was less than 0.05 (Hoogland & Boomsma, 1998).

Because one of the benefits of using MI is that it provides better standard error estimates, this study also summarized the efficiency of the parameter estimates. Note that the variance associated with the multiply imputed parameter estimate, $\bar{q}$, is a function of the average within-imputation variance, $\bar{u}$, and the between-imputation variance, $B$ (see Equations 2, 3 and 4). When the parameter estimate of interest is the $Z_r$-transformed correlation, its within-imputation variance is solely a function of sample size ($\hat{u} = \frac{1}{n - 3}$). Because sample size was not varied in any of the conditions of this study, the average within-imputation variance, $\bar{u}$, was consistently equal to $\frac{1}{n - 3}$, regardless of condition and replication. However, the between-imputations variance associated with $Z_r$ did vary across conditions and provided the source of resulting differences in the total variance associated with $\bar{q}$. For this reason, the efficiency of the $Z_r$-transformed correlations was summarized by calculating the average between-imputation variances by condition.

Results

The relative biases of correlation estimates with normally and non-normally distributed data are presented in Table 1. This table shows that MI of Likert-type data assuming continuous multivariate-normal data can yield acceptable parameter estimates with different types of missing data (MCAR and MAR) if the percentage of missing data is approximately 10%. However, with 30% of missing data, only the MCAR conditions resulted in acceptable relative bias. With 50% of missing data, acceptable relative biases were not obtained in any of the conditions. MI, assuming continuous data, showed robustness to categorization. Only slight differences in relative biases were identified between the three types of Likert scales. MI was also found to be robust to violations of normality. The relative biases of the skewed and normal conditions were similar.
The magnitude of correlations between variables (i.e., 0.8 or 0.2) also did not affect the performance of MI. The biases of parameter estimates obtained with MI were found to be consistently negative across all conditions. This leads to the conclusion that the presence of imputed data in datasets results in systematic reduction of the values of correlation coefficient estimates.

With MI, the variance associated with the multiply imputed parameter estimate is a function of the variability between estimates from each multiply imputed dataset as well as the variance of each estimate. (see Equations 2 - 4). This accounts for the extra amount of error introduced by the imputation process. Table 2 shows the average between-imputations variance summarized across generating conditions. The proportion of missing data had the strongest effect on the between-imputation variance. More specifically, as the overall proportion of missingness increased so did the between-imputation variance. A smaller effect was also identified: With the exception of conditions with 10% of missing data, the between-imputation variances were larger with correlation equal to 0.8 than 0.2. Furthermore, the conditions with 50% of missing data and correlation of 0.8 produced somewhat higher between-imputation variances, which increased as the number of points in the Likert scale increased. It is possible that this is the result of a three-way interaction between percentage of missing data, correlation between variables and number of points of the Likert scale. Additional studies expanding the levels of these three conditions would be needed to confirm the interaction.

**Conclusion**

Study results show that multiple imputation is robust to violations of both continuity and normality. This supports the assertion by Schafer (1997) that multiple imputation assuming the normal model works well even with ordered categorical data. However, it seems that resulting statistical tests will be less powerful because the sampling variance of the correlation estimates tends to increase and the values of the

<table>
<thead>
<tr>
<th>Percentage of Missing Data</th>
<th>Type</th>
<th>Correlation = 0.8*</th>
<th>Correlation = 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likert Scale</td>
<td></td>
<td>( k = 3 )</td>
<td>( k = 5 )</td>
</tr>
<tr>
<td>Normally-Distributed Data</td>
<td></td>
<td>( k = 3 )</td>
<td>( k = 5 )</td>
</tr>
<tr>
<td>10%</td>
<td>MCAR</td>
<td>-0.002</td>
<td>-0.006</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>-0.003</td>
<td>-0.007</td>
</tr>
<tr>
<td>30%</td>
<td>MCAR</td>
<td>-0.032</td>
<td>-0.037</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>-0.041</td>
<td>-0.053</td>
</tr>
<tr>
<td>50%</td>
<td>MCAR</td>
<td>-0.118</td>
<td>-0.129</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>-0.163</td>
<td>-0.183</td>
</tr>
<tr>
<td>Non-Normally Distributed Data</td>
<td></td>
<td>( k = 3 )</td>
<td>( k = 5 )</td>
</tr>
<tr>
<td>10%</td>
<td>MCAR</td>
<td>-0.003</td>
<td>-0.007</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>-0.016</td>
<td>-0.012</td>
</tr>
<tr>
<td>30%</td>
<td>MCAR</td>
<td>-0.035</td>
<td>-0.038</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>-0.057</td>
<td>-0.064</td>
</tr>
<tr>
<td>50%</td>
<td>MCAR</td>
<td>-0.118</td>
<td>-0.130</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>-0.200</td>
<td>-0.212</td>
</tr>
</tbody>
</table>

*Bold numbers indicate unacceptable bias*
MULTIPLE IMPUTATION FOR LIKERT-TYPE ITEMS WITH MISSING DATA

correlations themselves tend to be negatively biased as the proportion of missing data increases. It should be noted that this decrease in power is a somewhat desirable feature of multiple imputation given that it adds a suitable degree of uncertainty to the resulting imputed datasets. Consequently, significance tests performed after MI will tend to be conservative compared with tests using complete data. Table 2 presents the average between-imputations variances for each condition in which missing data had been introduced. When no missingness exists, the between-imputations variance is zero and the resulting total variance for an estimate based on a dataset without missingness will be smaller with a concomitant increase in power.

For multiply imputed datasets, although the significance tests have less power, they will also meet the desired nominal \( \alpha \)-levels; this is not the case when other missing data procedures such as mean and/or regression imputation are used. While the power of associated statistical tests under mean or regression imputation is maintained at the level of a complete data set, inflated Type I error rates can occur. Alternative missing data procedures such as listwise and/or pairwise deletion, similar to MI, are also known to result in decreased power. In addition, these deletion procedures have also been known to result in biased estimates given large degrees of missingness and non-MCAR patterns of missingness (Roth, Stwitzer & Switzer, 1999).

Based on results of the many different conditions simulated herein, it is possible to conclude that MI can be safely used to estimate parameters if the overall proportion of missing data is small (i.e., approximately 10%). If the data is missing completely at random, it was observed that as much as 30% of missing data does not result in inadequate parameter estimates. However, the major difficulty for applied researchers dealing with missing data is that it is not possible to know with certainty whether the missing values in a dataset are missing completely at random.

Table 2: Average Between-Imputation Variance of the Z, Estimates For Normally Distributed Data

<table>
<thead>
<tr>
<th>Percentage of Missing Data Likert Scale</th>
<th>Type</th>
<th>Correlation = 0.8</th>
<th>Correlation = 0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( k = 3 )</td>
<td>( k = 5 )</td>
</tr>
<tr>
<td>Normally Distributed Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>MCAR</td>
<td>0.0004</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>30%</td>
<td>MCAR</td>
<td>0.0033</td>
<td>0.0050</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>0.0045</td>
<td>0.0081</td>
</tr>
<tr>
<td>50%</td>
<td>MCAR</td>
<td>0.0150</td>
<td>0.0235</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>0.0188</td>
<td>0.0311</td>
</tr>
<tr>
<td>Non-Normally Distributed Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>MCAR</td>
<td>0.0004</td>
<td>0.0004</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>0.0004</td>
<td>0.0004</td>
</tr>
<tr>
<td>30%</td>
<td>MCAR</td>
<td>0.0032</td>
<td>0.0051</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>0.0038</td>
<td>0.0050</td>
</tr>
<tr>
<td>50%</td>
<td>MCAR</td>
<td>0.0148</td>
<td>0.0237</td>
</tr>
<tr>
<td></td>
<td>MAR</td>
<td>0.0140</td>
<td>0.0237</td>
</tr>
</tbody>
</table>
For MAR conditions, this study did not omit the variable that caused the missing data (i.e., variable Z) from the datasets, which improves the performance of MI (Collins, Schafer & Kam, 2001). However, in real datasets, it is common that the researcher does not know or does not include the variables causing the missing data in the dataset. It can be expected that biases in the parameter estimates due to missing data would be larger if the variable causing missingness was omitted. A limitation of this study is that all datasets had a sample size of 400; different results might be obtained if smaller or larger sample sizes were used.

The datasets used in this study contained 10 inter-correlated variables. This type of dataset approximates a measurement situation where there is a scale or survey containing similar items. MI can benefit from the presence of inter-correlated variables, because the inter-correlations provide some of the missing information. The results of this study may have been different if uncorrelated variables were used; however, datasets containing uncorrelated variables are unlikely in measurement settings. Conversely, this study used some conditions where variable inter-correlations were probably weaker (i.e., 0.2) or stronger (i.e., 0.8) than those that would be found for responses to real scales or surveys. Items correlated at 0.2 would be realized in surveys, but would be somewhat lower that what would be expected for a psychometric scale measuring a single construct. These correlations were used in order to simulate distinct conditions.

Many unknowns exist regarding the ability of MI to generate acceptable estimates with large amounts of missing data. The question: What is the maximum amount of missing data that can be adequately imputed? has no easy solution, due to the interaction between the proportion of missing data and the pattern of correlations between variables in the dataset. Future research should address the effects of predictors included in the dataset to increase the accuracy of MI estimates in situations where the proportion of missing data is large. Another point deserving further investigation is the quality of correlation estimates when MI is used with a large imputation model containing several covariates. Sinharay, Stern and Russell (2001) found that MI of datasets with 20 covariates under the MAR assumption resulted in negatively biased correlation estimates. Additional research could address the effect of the covariates in MI of both continuous and categorical data.

Currently, MI together with full information maximum likelihood estimation are the frontrunners among missing data methods in terms of providing the most adequate estimates in the presence of MCAR and MAR missing data. Despite the fact that MI is available in many statistical programs, it has not become common practice in applied research. This may be due to the complex specification of the MI model that some software require (e.g., S-PLUS and R) or to the time consuming task of combining multiple imputed datasets. To promote an increase in use of MI among applied researchers, more automatic handling of imputed datasets by software is needed.

Although it was found that the appropriateness of MI to deal with missing data depends on whether data is MCAR or MAR as well as the proportion of missing data, Schafer and Olsen (1998) pointed out that it is misleading to classify the missing data in a dataset according to just one type of relationship between missing values and variables, because missing values can occur for many reasons within the same dataset. Furthermore, situations exist where neither the MCAR nor the MAR assumptions are plausible. Unfortunately, current missing data methods cannot handle MNAR data. Care should be taken to ensure that the procedure used to deal with missing data is appropriate for the missing data mechanism for a particular dataset.

References


Model Based vs. Model Independent Tests for Cross-Correlation

H. E. T. Holgersson  Peter S. Karlsson
Jönköping International Business School,
Sweden

This article discusses the issue of whether cross correlation should be tested by model dependent or model independent methods. Several different tests are proposed and their main properties are investigated analytically and with simulations. It is argued that model independent tests should be used in applied work.

Key words: Cross correlation, residuals, lag window, hypothesis tests.

Introduction
Statistical analysis frequently involves the problem of whether two variables are related to each other. One of the most popular approaches is correlation analysis, initially proposed by Galton (1888) and refined by Fisher (1915, 1921). Later on correlations became popular also in time series contexts. When estimated correlation coefficients are used to test formal hypotheses, a test statistic with a (asymptotically) known null distribution is needed. In the case of independently distributed data (i.i.d.) there are several known standard error formulas for the correlation coefficient (Stuart & Ord, 1994). If autocorrelation exists in the data, however, these null distributions are not valid because the variance of the test statistic will depend on the unknown autocorrelation. It is therefore important to develop tests that take this aspect into account.

Through the last three decades a number of articles have been concerned with this issue. Some important works include Haugh (1976) and McLeod (1979) both of whom dealt with the distributional properties of residual based cross-correlation coefficients, Koch and Yang (1986) extended these methods to include pattern in the cross-correlation function, and Hallin and Saidi (2001) extended these two methods to the general multivariate case. Hong (1996) proposed a different approach of using an AR(p) model where p is allowed to grow asymptotically with the sample size T, and Bouhaddioui and Roy (2006) further developed this idea in a more general VAR(p) context.

All of these studies share the property that they involve residual based tests, constructed by first pre-whitening the data. The rationale behind this method is that the variance of the cross-correlation coefficient is somewhat complicated for autocorrelated data, and becomes much easier to handle for variables without autocorrelation. Thus, as residuals are asymptotically uncorrelated and the main interest is in the possible cross-correlation - not in the autocorrelation - this approach is reasonable. However, there is also an option to use some linear function of the sample cross-correlations and to construct a model independent test.

Model based tests have the disadvantage that a misspecified model may lead to an inconsistent procedure but also have the potential of being more efficient than model independent tests because they are more parsimonious regarding the number of parameters. It may be questioned how model
dependent tests perform relative to model independent tests, or is the potential efficiency gain of model based methods worth the risk of using a misspecified model? The aim of this article is to examine the properties of five different, simple tests of cross-correlation of weakly stationary bivariate processes. These involve a test dependent on a known model plus known parameters, two tests dependent on a known model but not of known parameters and two model independent tests. The asymptotic properties of the tests are established analytically and the small sample properties are examined by Monte Carlo simulations.

Methodology
Some properties of the sample correlation coefficient calculated from two possibly autocorrelated variables are considered; in particular, the focus is on the variance of the correlation coefficient. A few relevant measures must first be defined. Let \( X_t \) and \( Y_t \) be two random sequences such that

\[
X_t = \mu_X + \sum_{i=0}^{\infty} \psi_{X,t} \varepsilon_{x,t-i},
\]

and

\[
Y_t = \mu_Y + \sum_{i=0}^{\infty} \psi_{Y,t} \varepsilon_{y,t-i},
\]

where \( \{\varepsilon_{x,t}\}_{t=1}^{\infty} \) and \( \{\varepsilon_{y,t}\}_{t=1}^{\infty} \) are two sequences of zero mean i.i.d. random variables and \( \psi_x \) and \( \psi_y \) are absolutely summable, that is, \( \sum_{i=0}^{\infty} |\psi_{x,t}| < \infty \) and \( \sum_{i=0}^{\infty} |\psi_{y,t}| < \infty \), \( V[\varepsilon_{X}] < \infty \) and \( V[\varepsilon_{Y}] < \infty \). Letting \( \sigma^2_X \) and \( \sigma^2_Y \) be the variance of \( X_t \) and \( Y_t \) respectively, the cross correlation coefficient is defined by

\[
\rho_{XY}(k) = \frac{E[X_t - \mu_X]E[Y_{t-k} - \mu_Y]}{\sqrt{\sigma^2_X \sigma^2_Y}} = \rho_{XY}(-k).
\]  

The main thrust of this article is the following hypothesis:

\[
H_0: \rho_{XY}(k) = 0 \ \forall k \in \mathbb{Z}
\]

\[
H_{\lambda}: \rho_{XY}(k) \neq 0 \ \exists k \in \mathbb{Z}
\]

where \( X \) and \( Y \) are covariance stationary but possibly autocorrelated. In order to test this hypothesis, a proper test statistic with an asymptotically known null distribution is needed. The population correlation \( \rho_{XY}(k) \) may be estimated by

\[
\hat{\rho}_{XY}(k) = \frac{\hat{\sigma}_{XY}(k)}{\sqrt{\hat{\sigma}^2_X \hat{\sigma}^2_Y}},
\]

where

\[
\hat{\sigma}^2_x = \frac{1}{T} \sum_{t=1}^{T} (Y_t - \bar{Y})^2,
\]

\[
\hat{\sigma}_{xy}(k) = \frac{1}{T} \sum_{t=1}^{T-k} (X_t - \bar{X})(Y_{t+k} - \bar{Y})
\]

and \( T \) is the number of observations. For identically independently distributed data it is well known that, if \( \rho_{XY}(k) = 0 \), then

\[
\hat{\rho}_{XY}(k) \xrightarrow{\ell} N(0,1/T)
\]

where \( \ell \) denotes convergence in law. An improvement of (2.3) is given by Fisher’s z-transformation (Fisher, 1921; Stuart & Ord, 1994). In cases when the data is not independent this variance is no longer valid. Using the well-known Bartlett approximation (for example, see Box, et al., 1994) the variance of the sample cross correlation is given by

\[
Var\left(\hat{\rho}_{XY}(k)\right) \approx
\]

\[
T^{-1} \sum_{j=-\infty}^{\infty} \begin{bmatrix}
\rho_{XX}(\tau) \rho_{YY}(\tau) \\
+ \rho_{XY}(k+\tau) \rho_{YY}(k-\tau) \\
-2 \rho_{XY}(k) \left( \rho_{XX}(\tau) \rho_{YY}(\tau+k) + \rho_{YY}(\tau+k) \rho_{XX}(\tau) \right) \\
- \rho_{XY}(k) \left( \rho_{YY}(\tau) + 0.5 \rho_{XX}^2(\tau) \right)
\end{bmatrix}
\]

\[
\]
where $\rho_{XY}(\tau)$ is the correlation between $X_t$ and $Y_{t-\tau}$. Equation (4) gives the variance of the sample cross-correlation coefficient between $X$ and $Y$ with a lag shift of $k$ steps. Hence, under the simple null hypothesis that $\rho_{XY}(k) = 0$ the equation (4) reduces to

$$\text{Var}\left[\hat{\rho}_{XY}(k) \bigg| \rho_{XY}(k) = 0\right] \approx T^{-1} \sum_{\tau=-\infty}^{\infty}\left[\rho_{XX}(\tau)\rho_{YY}(\tau) + \rho_{XY}(\tau+k)\rho_{YY}(\tau-k)\right]. \tag{5}$$

Furthermore, under the null hypothesis that all cross covariances are zero (as in (2)), results in

$$\lambda := \text{Var}\left[\hat{\rho}_{XY}(k) \bigg| \rho_{XY}(k) = 0 \forall k \in \mathbb{Z}\right] \approx T^{-1} \sum_{\tau=-\infty}^{\infty}\left[\rho_{XX}(\tau)\rho_{YY}(\tau)\right]. \tag{6}$$

Accordingly, if a consistent estimate of $\lambda$ can be obtained (for example, $\hat{\lambda}$), it follows that

$$\frac{\left(\hat{\rho}_{XY}(k) \bigg| \rho_{XY}(k) = 0 \forall k \in \mathbb{Z}\right)}{\sqrt{\hat{\lambda}}} \to N(0,1). \tag{7}$$

From these formulas it is apparent that several possible ways exist with which to test for zero cross-correlation.

Firstly, one may test if a particular cross-correlation at lag $k$ is zero while allowing for non-zero cross-correlations at other lags; then an estimate of (5) is sufficient to form a proper test statistic. Secondly, one might like to test whether there are any non-zero cross-correlations above a certain lag. Thirdly, one may test whether there are any non-zero cross-correlations at all. This is the hypothesis expressed in (2) and is the main issue here. The question is how to construct a test that is both consistent and also reasonably simple to perform. Observably, equation (7) can be used to form a consistent test if $\rho_{XY}(k) \neq 0$. However, the null hypothesis states that the cross-correlations are zero at all lags. The question is then what will happen if the cross-correlation at lag $k$ is zero but there is at least one non-zero coefficient at some other lag, e.g. if $\rho_{XY}(k) = 0$ but $\rho_{XY}(k+l) \neq 0$ for some $l \neq 0$.

To address this question, two things should be noted. First, the cross-correlation function is, in most cases, exponentially decaying so that even if the value of $k$ corresponding to the largest cross-correlation is not specified there will still be a non-zero cross-correlation at $k$. Thus, it is not likely that an inappropriately chosen $k$ is specified such that $\rho_{XY}(k) = 0$ under the alternative hypothesis. Second, in a comparison of equations (5) and (6), there will still be a sense in which the test is consistent as the test statistic will diverge from its null distribution. In other words, specifying a value $k$ that does not correspond exactly to the largest cross-correlation is merely a matter of optimality rather than consistency. There also exists a possibility to involve several $\rho_{XY}(k)$ explicitly in the test: one might use the sum of squared cross-correlations within a certain interval, for example, $\hat{\rho}_{XY}^2(-h)+...+\hat{\rho}_{XY}^2(h)$. Unfortunately such an approach will introduce additional complications as the sample cross-correlations will not be uncorrelated even under the null hypothesis (apart from the unlikely special case of independent data). Therefore, several authors, including Haugh (1976), McLeod (1979), Koch and Yang (1986) and Hallin and Saidi (2001) proposed model dependent tests and then applied this kind of test on the asymptotically uncorrelated residuals. For example, if $\hat{u}$ and $\hat{v}$ are residuals from ARMA models, a test may be defined by

$$Q(h) = \hat{\rho}_{uv}^2(-h)+...+\hat{\rho}_{uv}^2(h).$$

A slightly different situation arises in cases where there is some sort of a priori knowledge of which lag the largest cross-correlation might be (if any), the null hypothesis (2) can be tested by the asymptotic null distribution of (7); then one is left with the issue of how to estimate the variance $\lambda$ of equation (6). This approach is followed here because the other is fairly well investigated in the literature. In particular, two different approaches are investigated: (i) tests dependent upon a model,
and (ii) tests independent of model assumptions. Case (i) may be dealt with as follows: if \( X_t \) and \( Y_t \) are known to follow a finite-order ARMA process, then the autocorrelations \( \rho_{XX} \) and \( \rho_{YY} \) may be expressed as functions of the autoregressive parameters. For example, if \( X_t \) and \( Y_t \) are given by two ARMA(1,1) processes, that is, if

\[
X_t - \phi_X X_{t-1} = \varepsilon_{X,t} - \theta_X \varepsilon_{X,t-1}
\]

and

\[
Y_t - \phi_Y Y_{t-1} = \varepsilon_{Y,t} - \theta_Y \varepsilon_{Y,t-1}
\]

then the autocorrelations of \( X_t \) are known to be given by

\[
\rho_{XX}(1) = \frac{(1 - \phi_X \theta_X)(\phi_X - \theta_X)}{1 + \theta_X^2 - 2\phi_X \theta_X},
\]

\[
\rho_{XX}(\tau) = \phi_X \rho_{XX}(1), \quad \tau > 1.
\]

Hence, using obvious notation,

\[
\rho_{XX}(1) \rho_{YY}(1) = \frac{(1 - \phi_X \theta_X)(\phi_X - \theta_X)(1 - \phi_Y \theta_Y)(\phi_Y - \theta_Y)}{(1 + \theta_X^2 - 2\phi_X \theta_X)(1 + \theta_Y^2 - 2\phi_Y \theta_Y)}
\]

\[
\rho_{XX}(\tau) \rho_{YY}(\tau) = \phi_X^{\tau-1} \rho_{XX}(1) \phi_Y^{\tau-1} \rho_{YY}(1)
\]

results in

\[
\sum_{\tau=1}^{\infty} \rho_{XX}(\tau) \rho_{YY}(\tau) = \sum_{\tau=1}^{\infty} \phi_X^{\tau-1} \rho_{XX}(1) \phi_Y^{\tau-1} \rho_{YY}(1) = \rho_{XX}(1) \rho_{YY}(1) \left( \frac{1}{(1 - \phi_X \phi_Y)} \right).
\]

Thus, if \( X_t \) and \( Y_t \) are two ARMA(1,1) processes it follows that

\[
Var\left[ \hat{\rho}_{XY}(k) \right] = T^{-1} \frac{(1 - \phi_X \phi_Y)(\phi_X - \theta_X)(\phi_Y - \theta_Y)}{(1 + \theta_X^2 - 2\phi_X \theta_X)(1 + \theta_Y^2 - 2\phi_Y \theta_Y)} \left( \frac{1}{(1 - \phi_X \phi_Y)} \right)
\]

From (2.8) the variance for AR(1) or MA(1) processes are immediately obtained by setting the irrelevant parameter to zero. This estimator can easily be generalised to ARMA(\( p, q \)) processes of arbitrary orders by substituting \( \rho_{XX}(\tau) \) and \( \rho_{YY}(\tau) \) with the model-based autocorrelations. These are acquired by the autocorrelation generating function which can be found in the time series literature (see, for example, Hamilton, 1994). The unknown parameters of (8) should be replaced by any consistent estimates such as maximum likelihood estimates or non-linear least squares (see Brockwell & Davis, 1991; Box, et al., 1994 for further details on estimations of ARMA parameters).

An alternative way to use model based tests is to use the asymptotically independent residuals: If the parameters of the ARMA model were actually known, then the two marginal models \( X_t \) and \( Y_t \) could be reformulated according to

\[
\varepsilon_{X,t} = X_t - \phi_X X_{t-1} + \theta_X \varepsilon_{X,t-1}
\]

and

\[
\varepsilon_{Y,t} = Y_t - \phi_Y Y_{t-1} + \theta_Y \varepsilon_{Y,t-1}
\]

Thus, by replacing the true ARMA parameter by consistent estimates the resulting asymptotically white noise residuals, \( \hat{\varepsilon}_{X,t} \) and \( \hat{\varepsilon}_{Y,t} \), can be used to test for cross-correlations because the variance of the cross-correlation may be approximated by \( 1/T \), according to equation (3).

Residual based tests have been proposed earlier in the literature, including the citations above, but will still be considered for comparison. The advantage of ARMA based
tests is that they are parsimonious, although the disadvantage is that they are model-dependent and a rough approximation to the true unknown functional form may lead to an inconsistent variance estimate. Hence, it is of interest to also consider a variance estimate that does not rely on any model assumptions. In particular, the cross correlations \( \hat{\rho}_{XY}(\tau) \) of equation (6) could be substituted directly with the sample autocorrelations:

\[
\hat{\rho}_{XY}(\tau) = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(X_{t-\tau} - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}.
\]

However, as (6) is a sum of infinitely many parameters some care needs to be taken: If a stochastic process is absolutely summable with finite fourth-order moments, then \( \hat{\rho}_{XX}(\tau) \rightarrow N(\rho_{XX}(\tau), W_{\tau}) \) for some \( W_{\tau} < \infty \) (see Brockwell & Davis, 1991). Hence the variance and the bias of \( \hat{\rho}_{\tau\tau} \) are of the order \( O(T^{-\alpha}) \) for any \( \alpha \in (0,1/2) \), and \( \hat{\rho}_{\tau\tau}(\tau) \) converges in mean square to \( \rho_{\tau\tau}(\tau) \) at the rate \( O(T^{-\alpha}) \), that is, \( \hat{\rho}_{XX}(\tau) = \rho_{XX}(\tau) + O_p(T^{-\alpha}) \).

Accordingly it follows that \( \hat{\rho}_{XX}(\tau)\hat{\rho}_{YY}(\tau) = \rho_{XX}(\tau)\rho_{YY}(\tau) + O_p(T^{-\alpha}) \), and an estimate of (6) can be formulated. In particular, absolute summability of the original variables \( X \) and \( Y \) implies absolute summability of the sequence \( \{\rho_{XX}(\tau)\rho_{YY}(\tau)\}_{\tau=0}^{\infty} \). Thus, for some monotonically increasing function \( q = q(T) \),

\[
\delta(q) = \sum_{\tau=1}^{q} \hat{\rho}_{XX}(\tau)\hat{\rho}_{YY}(\tau) - \sum_{\tau=1}^{\infty} \rho_{XX}(\tau)\rho_{YY}(\tau) = \sum_{\tau=1}^{q} [\hat{\rho}_{XX}(\tau)\hat{\rho}_{YY}(\tau) - \rho_{XX}(\tau)\rho_{YY}(\tau)] + O(1)
\]

\( (10) \)

The literature concerning the convergence of sequences of the type \( \delta(q) \) is extensive, one of the most cited being Newey and West (1987). If \( \hat{\rho}_{XX}(\tau) = \rho_{XX}(\tau) + O_p(T^{-1/2+\epsilon}) \) for all \( \epsilon > 0 \) (which is the convergence rate met in most linear estimates) but the convergence of \( \hat{\rho}_{XX}(\tau) \) cannot be assumed to hold uniformly in \( \tau \), then \( q \) must be restricted to values below \( T^{1/4} \) in order to ensure that \( \delta(q) = o_p(1) \).

However, for linear processes with finite fourth moments, i.i.d. innovations and absolute summable coefficients, \( q \) may be relaxed to values below \( T^{1/2} \), \( \delta(T^{-1/2+\epsilon}) = o_p(1) \). Moreover, recent results (e.g., Lobato & Velasco, 2004; Robinson, 1998) have shown that, in many cases, sequences of the above type may converge for values up to \( T \), \( \delta(T^{-1/2+\epsilon}) = o_p(1) \). This is mainly a consequence of operating with sums containing stochastic down weighting such as \( \sum_{\tau=1}^{T} \hat{\rho}_{XX}(\tau)\hat{\rho}_{YY}(\tau) \) herein; \( \hat{\rho}_{XX}(\tau) \) down weights \( \hat{\rho}_{YY}(\tau) \) and vice versa and both decrease individually in \( \tau \). These properties indicate that restricting \( q \) to values below \( T^{1/4} \) might be unnecessarily stringent; therefore the compromise \( \delta(T^{-1/2+\epsilon}) \) is used in this article so that the proposed model-free estimate of (6) takes the form

\[
\hat{\lambda} = \sum_{\tau=1}^{q} \hat{\rho}_{XX}(\tau)\hat{\rho}_{YY}(\tau),
\]

where \( q \leq \text{int}(T^{-1/2+\epsilon}) \) . \( (11) \)

Hence, the variance estimate of (11) consistently estimates the variance component of (6). But, this estimate is not guaranteed to be positive in small samples, for this reason another variance estimate which is strictly non-negative is also considered:

\[
\widetilde{\lambda} = \sum_{\tau=0}^{q} (1 - \tau/(q+1)){\hat{\rho}_{XX}(\tau){\hat{\rho}_{YY}(\tau)}} . \]

\( (12) \)
The non-negativeness of (12) is easily established: When \(X_t\) and \(Y_t\) are two absolutely summable stochastic processes with finite fourth moments and \(\hat{\Gamma}_{X,T}\) and \(\hat{\Gamma}_{Y,T}\) are the matrices of the sample autocorrelations, it is well known that \(\hat{\Gamma}_{X,T}\) and \(\hat{\Gamma}_{Y,T}\) are both non-negative matrices (Brockwell and Davis, 1991). Moreover, because direct products (symbolized by \(\otimes\)) of non-negative matrices are also non-negative (Schott, 1997), it follows that \(\hat{\Gamma}_{X,T} \otimes \hat{\Gamma}_{Y,T}\) is non-negative as well. Hence there exists an \(L\) such that \(\hat{\Gamma}_{X,T} \otimes \hat{\Gamma}_{Y,T} = LL \geq 0\), then if \(1_q = (111...100..0)\) such that \(1_q L_1_q = q\), it follows that:

\[
\tilde{\lambda} = \sum_{\tau=0}^{q-1} (1-\tau/(q+1)) \hat{\rho}_{xx}(\tau) \hat{\rho}_{yy}(\tau) \\
= 1_q \hat{\Gamma}_{X,T} \otimes \hat{\Gamma}_{Y,T} 1_q \\
= 1_q L_1_q \\
= (L_1_q)' (L_1_q) \geq 0
\]

In other words, if \(X, Y\) are two linear processes with finite fourth order moments and absolute summable coefficients and \(q \leq \text{int} \left( T^{-1/2+e} \right)\), then \(\tilde{\lambda}\) is a non-negative and consistent estimate of (6). Truncating the sample autocorrelation function at a certain point, as in (11), is sometimes referred to as a rectangular lag window, and estimates of the kind in (12) are referred to as a triangular window. That terminology is adopted later, even though here work with products of correlations is employed as opposed to individual correlations (which is the usual case).

To sum up, four estimates of the variance of equation (6) have been proposed, two model-independent and two model-based estimates. The first two use the same information set, namely the ARMA model and its parameter estimates; the other two depend only upon the truncation point and the choice of lag window. Of particular interest is the potential difference between the model based and the model independent tests; how much gain is there in knowing the true model? It is also of interest to investigate the possible difference within each type of test, asking the questions: Does it matter how one makes use of the known model and does the choice of lag window make a difference?

**Results**

When investigating the properties of a test procedure, two aspects are of prime importance. First it is necessary to determine whether the actual size of the test - the probability of rejecting the null hypothesis when it is true - is close to the nominal size. Given that the actual size is a reasonable approximation to the nominal size, it is then necessary to investigate the actual power of the test - the probability of rejecting the null hypothesis when it is false - for a number of different parameter settings. The number of replicates in the computer simulations is 100,000 for each size and power simulation.

In this study the relevant factor is first and foremost the choice of test. Five different tests are considered based on the statistic (7) but with different estimates of the standard error \(\lambda\), namely (i) the ARMA based test using the asymptotically white noise residuals (so that \(\lambda = 1/T\)), (ii) standard error obtained from (8) using the true ARMA parameters, (iii) standard error obtained by (8) using maximum likelihood estimate of the ARMA parameters, (iv) standard error using the rectangular lag window (11) with truncation point \(q = \text{int} \left[ T^{-0.45} \right]\), and finally (v) the test based on the standard error using the triangular lag window (12), again with truncation point \(q = \text{int} \left[ T^{-0.45} \right]\).

It is critical to identify possible differences between these five tests, and in order to do so some different autocorrelation patterns must be considered. For that purpose the AR(1) process, MA(1) process and ARMA(1,1) processes are used with different values of autoregressive parameters, ranging from white noise (independent data) up to high autocorrelation. Moreover, two different sample sizes are used: 30 observations (which is usually considered as a small sample in time series
analysis) and 200 observations (medium-sized sample). Finally, in order to investigate the tests’ power to detect correlation, cross-correlations ranging from 0 (no correlation) up to 0.9 (very strong correlation) are considered. The significance level is set to the 0.05 level in all models so that the critical values are $-1.96$ and $1.96$ in all tests.

By counting the number of rejections the empirical significance level is identified for each test conducted. The results are presented in Tables 1-8. According to Table 1, which deals with the special case of two independent white noise processes, it is observed that all tests have an almost perfect size relative to their nominal sizes, except perhaps the residual test for the smallest samples. Although this is not an unexpected result (because the sample autocorrelations converge rapidly for white noise) it is still interesting because it reveals that the choice of test is almost irrelevant for white noise data. Unfortunately, the choice of test becomes less obvious when considering the size properties of autocorrelated data.

As shown in Table 2, there are some notable differences between the various tests. In particular, the rejection frequencies of the model-based tests (as functions of true respectively estimated parameters) reveal that there is no obvious gain in knowing the true ARMA parameters. Even though the under-rejection of both these tests seems to worsen for larger values of the autocorrelation parameter, the test of estimated ARMA parameters underestimates less when compared with the corresponding test of the true parameters. Moreover, there is also a somewhat drastic difference between the two model-independent tests. In fact, the test of the rectangular lag window seems to uniformly outperform that of the triangular lag window. Although the test of the rectangular lag window slightly over rejects for high autocorrelation, the effect is not that serious in contrast to that of the triangular lag window which shows a rejection frequency of 0.11 at high autocorrelation and small $T$. It is noteworthy that the residual-based test behaves satisfactorily at all sample sizes and autocorrelations.

Table 3 shows some interesting differences compared to Table 2. The residual-based test no longer maintains its good size properties, no difference exists between the two model-based tests and, additionally, the difference between the two model-independent tests is now very small (they both stay fairly close to the nominal size though the rectangular window is slightly closer).

Not unexpectedly, the rejection frequencies shown in Table 4 are a mixture of the results shown in Tables 2 and 3. Hence it is not easy to select a test that is generally better than another when it comes to size properties, though the residual-based test and the model-free test using the rectangular lag window may be said to have good overall properties.

The power simulations in Tables 5 and 6 present rejection frequencies for AR(1) properties at two sample sizes, 30 and 200 observations respectively. It is striking that the differences of the various tests are negligible for white noise, irrespective of whether the sample size is 30 or 200. Conversely, there appears to be a difference when the autoregressive parameter is 0.7.

The general pattern is that the model-based tests have surprisingly low power although the residual-based test has higher power than any other test. In fact, the difference is even more accentuated for the large sample size. The two model-independent tests have power properties between the model-based test and the residual-based one. The residual-based test maintains its superior power for the MA process (Tables 7 and 8) even if the difference to the other tests is now less drastic.

For most parameter values and sample sizes the model-free tests are not far behind those of the residual test. If one or two winners of the 5 tests are to be selected, one should start by considering tests that have fairly acceptable size properties - even for strong autocorrelation. This rules out the model-based tests (i) and (ii) as well as the model-independent test using a triangular lag window. The remaining two tests both have their own pros and cons; the residual-based test uniformly outperforms the model-independent test, but at the same time it should be noted that it is somewhat difficult to assume the model to be known. For this reason, and because the model-independent test is clearly consistent and not much weaker in power than
the residual-based test, one might want to recommend the test of the rectangular window for an applied situation unless the true model is known.

Table 1: Estimated Size for White Noise Process

<table>
<thead>
<tr>
<th>Sample Size $T$</th>
<th>Residual-Based Test</th>
<th>Model-Based, True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.060</td>
<td>0.052</td>
<td>0.052</td>
<td>0.050</td>
<td>0.052</td>
</tr>
<tr>
<td>30</td>
<td>0.058</td>
<td>0.049</td>
<td>0.049</td>
<td>0.048</td>
<td>0.048</td>
</tr>
<tr>
<td>40</td>
<td>0.056</td>
<td>0.050</td>
<td>0.050</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td>50</td>
<td>0.054</td>
<td>0.050</td>
<td>0.050</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td>70</td>
<td>0.054</td>
<td>0.051</td>
<td>0.051</td>
<td>0.049</td>
<td>0.050</td>
</tr>
<tr>
<td>100</td>
<td>0.052</td>
<td>0.052</td>
<td>0.052</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>200</td>
<td>0.052</td>
<td>0.052</td>
<td>0.052</td>
<td>0.053</td>
<td>0.052</td>
</tr>
<tr>
<td>500</td>
<td>0.050</td>
<td>0.051</td>
<td>0.051</td>
<td>0.052</td>
<td>0.051</td>
</tr>
</tbody>
</table>

Table 2: Estimated Size for AR(1) Process

<table>
<thead>
<tr>
<th>Sample Size $T$</th>
<th>Residual-Based Test</th>
<th>Model-Based, True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Phi = 0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.057</td>
<td>0.050</td>
<td>0.051</td>
<td>0.052</td>
<td>0.055</td>
</tr>
<tr>
<td>30</td>
<td>0.057</td>
<td>0.051</td>
<td>0.051</td>
<td>0.054</td>
<td>0.055</td>
</tr>
<tr>
<td>40</td>
<td>0.055</td>
<td>0.049</td>
<td>0.051</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td>50</td>
<td>0.057</td>
<td>0.051</td>
<td>0.051</td>
<td>0.054</td>
<td>0.055</td>
</tr>
<tr>
<td>70</td>
<td>0.054</td>
<td>0.050</td>
<td>0.049</td>
<td>0.050</td>
<td>0.053</td>
</tr>
<tr>
<td>100</td>
<td>0.056</td>
<td>0.050</td>
<td>0.049</td>
<td>0.051</td>
<td>0.052</td>
</tr>
<tr>
<td>200</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
<td>0.055</td>
<td>0.056</td>
</tr>
<tr>
<td>500</td>
<td>0.050</td>
<td>0.054</td>
<td>0.054</td>
<td>0.054</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phi = 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.056</td>
<td>0.028</td>
<td>0.034</td>
<td>0.057</td>
<td>0.069</td>
</tr>
<tr>
<td>30</td>
<td>0.056</td>
<td>0.035</td>
<td>0.039</td>
<td>0.057</td>
<td>0.066</td>
</tr>
<tr>
<td>40</td>
<td>0.056</td>
<td>0.038</td>
<td>0.039</td>
<td>0.056</td>
<td>0.065</td>
</tr>
<tr>
<td>50</td>
<td>0.057</td>
<td>0.041</td>
<td>0.042</td>
<td>0.055</td>
<td>0.062</td>
</tr>
<tr>
<td>70</td>
<td>0.054</td>
<td>0.043</td>
<td>0.045</td>
<td>0.055</td>
<td>0.063</td>
</tr>
<tr>
<td>100</td>
<td>0.056</td>
<td>0.046</td>
<td>0.047</td>
<td>0.055</td>
<td>0.061</td>
</tr>
<tr>
<td>200</td>
<td>0.052</td>
<td>0.050</td>
<td>0.050</td>
<td>0.054</td>
<td>0.060</td>
</tr>
<tr>
<td>500</td>
<td>0.050</td>
<td>0.052</td>
<td>0.052</td>
<td>0.053</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Phi = 0.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.057</td>
<td>0.001</td>
<td>0.006</td>
<td>0.071</td>
<td>0.114</td>
</tr>
<tr>
<td>30</td>
<td>0.056</td>
<td>0.003</td>
<td>0.011</td>
<td>0.070</td>
<td>0.111</td>
</tr>
<tr>
<td>40</td>
<td>0.055</td>
<td>0.009</td>
<td>0.015</td>
<td>0.068</td>
<td>0.110</td>
</tr>
<tr>
<td>50</td>
<td>0.059</td>
<td>0.016</td>
<td>0.020</td>
<td>0.065</td>
<td>0.104</td>
</tr>
<tr>
<td>70</td>
<td>0.053</td>
<td>0.027</td>
<td>0.031</td>
<td>0.064</td>
<td>0.097</td>
</tr>
<tr>
<td>100</td>
<td>0.056</td>
<td>0.033</td>
<td>0.037</td>
<td>0.065</td>
<td>0.093</td>
</tr>
<tr>
<td>200</td>
<td>0.053</td>
<td>0.041</td>
<td>0.043</td>
<td>0.059</td>
<td>0.081</td>
</tr>
<tr>
<td>500</td>
<td>0.050</td>
<td>0.049</td>
<td>0.049</td>
<td>0.056</td>
<td>0.072</td>
</tr>
</tbody>
</table>
Table 3: Estimated Size for MA(1) Process

<table>
<thead>
<tr>
<th>Sample Size $T$</th>
<th>Residual-Based Test</th>
<th>Model-Based True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>True Parameters</td>
<td>Estimated Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Theta = 0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.056</td>
<td>0.054</td>
<td>0.054</td>
<td>0.052</td>
<td>0.055</td>
</tr>
<tr>
<td>30</td>
<td>0.054</td>
<td>0.052</td>
<td>0.052</td>
<td>0.052</td>
<td>0.055</td>
</tr>
<tr>
<td>40</td>
<td>0.054</td>
<td>0.054</td>
<td>0.053</td>
<td>0.054</td>
<td>0.057</td>
</tr>
<tr>
<td>50</td>
<td>0.055</td>
<td>0.057</td>
<td>0.056</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td>70</td>
<td>0.057</td>
<td>0.058</td>
<td>0.058</td>
<td>0.059</td>
<td>0.061</td>
</tr>
<tr>
<td>100</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
<td>0.059</td>
<td>0.059</td>
</tr>
<tr>
<td>200</td>
<td>0.053</td>
<td>0.054</td>
<td>0.054</td>
<td>0.055</td>
<td>0.056</td>
</tr>
<tr>
<td>500</td>
<td>0.051</td>
<td>0.052</td>
<td>0.052</td>
<td>0.052</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>Theta = 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.058</td>
<td>0.052</td>
<td>0.051</td>
<td>0.053</td>
<td>0.059</td>
</tr>
<tr>
<td>30</td>
<td>0.054</td>
<td>0.051</td>
<td>0.051</td>
<td>0.054</td>
<td>0.061</td>
</tr>
<tr>
<td>40</td>
<td>0.054</td>
<td>0.054</td>
<td>0.054</td>
<td>0.057</td>
<td>0.063</td>
</tr>
<tr>
<td>50</td>
<td>0.054</td>
<td>0.058</td>
<td>0.058</td>
<td>0.059</td>
<td>0.064</td>
</tr>
<tr>
<td>70</td>
<td>0.058</td>
<td>0.057</td>
<td>0.058</td>
<td>0.061</td>
<td>0.064</td>
</tr>
<tr>
<td>100</td>
<td>0.058</td>
<td>0.058</td>
<td>0.057</td>
<td>0.058</td>
<td>0.061</td>
</tr>
<tr>
<td>200</td>
<td>0.053</td>
<td>0.055</td>
<td>0.055</td>
<td>0.056</td>
<td>0.060</td>
</tr>
<tr>
<td>500</td>
<td>0.051</td>
<td>0.053</td>
<td>0.053</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>Theta = 0.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.105</td>
<td>0.051</td>
<td>0.049</td>
<td>0.054</td>
<td>0.066</td>
</tr>
<tr>
<td>30</td>
<td>0.091</td>
<td>0.051</td>
<td>0.051</td>
<td>0.055</td>
<td>0.063</td>
</tr>
<tr>
<td>40</td>
<td>0.080</td>
<td>0.053</td>
<td>0.053</td>
<td>0.052</td>
<td>0.065</td>
</tr>
<tr>
<td>50</td>
<td>0.072</td>
<td>0.056</td>
<td>0.054</td>
<td>0.059</td>
<td>0.067</td>
</tr>
<tr>
<td>70</td>
<td>0.065</td>
<td>0.058</td>
<td>0.057</td>
<td>0.063</td>
<td>0.067</td>
</tr>
<tr>
<td>100</td>
<td>0.061</td>
<td>0.058</td>
<td>0.057</td>
<td>0.059</td>
<td>0.064</td>
</tr>
<tr>
<td>200</td>
<td>0.054</td>
<td>0.056</td>
<td>0.056</td>
<td>0.057</td>
<td>0.061</td>
</tr>
<tr>
<td>500</td>
<td>0.051</td>
<td>0.052</td>
<td>0.052</td>
<td>0.052</td>
<td>0.054</td>
</tr>
</tbody>
</table>
Table 4: Estimated Size for ARMA(1,1) Process

<table>
<thead>
<tr>
<th>Sample Size $T$</th>
<th>Residual-Based Test</th>
<th>Model-Based, True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.065</td>
<td>0.043</td>
<td>0.050</td>
<td>0.051</td>
<td>0.054</td>
</tr>
<tr>
<td>30</td>
<td>0.062</td>
<td>0.042</td>
<td>0.048</td>
<td>0.050</td>
<td>0.052</td>
</tr>
<tr>
<td>40</td>
<td>0.059</td>
<td>0.039</td>
<td>0.045</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td>50</td>
<td>0.055</td>
<td>0.040</td>
<td>0.048</td>
<td>0.050</td>
<td>0.051</td>
</tr>
<tr>
<td>70</td>
<td>0.052</td>
<td>0.041</td>
<td>0.051</td>
<td>0.051</td>
<td>0.054</td>
</tr>
<tr>
<td>100</td>
<td>0.052</td>
<td>0.043</td>
<td>0.051</td>
<td>0.054</td>
<td>0.053</td>
</tr>
<tr>
<td>200</td>
<td>0.051</td>
<td>0.047</td>
<td>0.051</td>
<td>0.051</td>
<td>0.052</td>
</tr>
<tr>
<td>500</td>
<td>0.049</td>
<td>0.046</td>
<td>0.047</td>
<td>0.048</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.068</td>
<td>0.035</td>
<td>0.041</td>
<td>0.053</td>
<td>0.059</td>
</tr>
<tr>
<td>30</td>
<td>0.057</td>
<td>0.037</td>
<td>0.044</td>
<td>0.054</td>
<td>0.061</td>
</tr>
<tr>
<td>40</td>
<td>0.057</td>
<td>0.041</td>
<td>0.049</td>
<td>0.056</td>
<td>0.063</td>
</tr>
<tr>
<td>50</td>
<td>0.051</td>
<td>0.037</td>
<td>0.045</td>
<td>0.052</td>
<td>0.056</td>
</tr>
<tr>
<td>70</td>
<td>0.055</td>
<td>0.044</td>
<td>0.050</td>
<td>0.053</td>
<td>0.058</td>
</tr>
<tr>
<td>100</td>
<td>0.051</td>
<td>0.044</td>
<td>0.048</td>
<td>0.050</td>
<td>0.054</td>
</tr>
<tr>
<td>200</td>
<td>0.052</td>
<td>0.049</td>
<td>0.050</td>
<td>0.052</td>
<td>0.056</td>
</tr>
<tr>
<td>500</td>
<td>0.054</td>
<td>0.052</td>
<td>0.053</td>
<td>0.053</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.075</td>
<td>0.029</td>
<td>0.026</td>
<td>0.061</td>
<td>0.084</td>
</tr>
<tr>
<td>30</td>
<td>0.061</td>
<td>0.037</td>
<td>0.037</td>
<td>0.060</td>
<td>0.076</td>
</tr>
<tr>
<td>40</td>
<td>0.055</td>
<td>0.042</td>
<td>0.042</td>
<td>0.061</td>
<td>0.081</td>
</tr>
<tr>
<td>50</td>
<td>0.055</td>
<td>0.042</td>
<td>0.043</td>
<td>0.060</td>
<td>0.072</td>
</tr>
<tr>
<td>70</td>
<td>0.053</td>
<td>0.043</td>
<td>0.045</td>
<td>0.055</td>
<td>0.067</td>
</tr>
<tr>
<td>100</td>
<td>0.049</td>
<td>0.045</td>
<td>0.047</td>
<td>0.054</td>
<td>0.064</td>
</tr>
<tr>
<td>200</td>
<td>0.046</td>
<td>0.042</td>
<td>0.042</td>
<td>0.047</td>
<td>0.055</td>
</tr>
<tr>
<td>500</td>
<td>0.052</td>
<td>0.049</td>
<td>0.049</td>
<td>0.050</td>
<td>0.056</td>
</tr>
</tbody>
</table>
Table 5: Estimated Power for AR(1) Process
Sample Size = 30

<table>
<thead>
<tr>
<th>Underlying Cross-Correlation</th>
<th>Residual-Based Test</th>
<th>Model-Based, True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phi = 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.058</td>
<td>0.054</td>
<td>0.054</td>
<td>0.052</td>
<td>0.054</td>
</tr>
<tr>
<td>0.1</td>
<td>0.093</td>
<td>0.090</td>
<td>0.088</td>
<td>0.086</td>
<td>0.087</td>
</tr>
<tr>
<td>0.2</td>
<td>0.191</td>
<td>0.194</td>
<td>0.193</td>
<td>0.188</td>
<td>0.192</td>
</tr>
<tr>
<td>0.3</td>
<td>0.365</td>
<td>0.382</td>
<td>0.380</td>
<td>0.369</td>
<td>0.376</td>
</tr>
<tr>
<td>0.4</td>
<td>0.589</td>
<td>0.619</td>
<td>0.616</td>
<td>0.606</td>
<td>0.613</td>
</tr>
<tr>
<td>0.5</td>
<td>0.802</td>
<td>0.833</td>
<td>0.829</td>
<td>0.822</td>
<td>0.828</td>
</tr>
<tr>
<td>0.6</td>
<td>0.945</td>
<td>0.959</td>
<td>0.957</td>
<td>0.954</td>
<td>0.957</td>
</tr>
<tr>
<td>0.7</td>
<td>0.992</td>
<td>0.994</td>
<td>0.995</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Phi = 0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.058</td>
<td>0.050</td>
<td>0.047</td>
<td>0.053</td>
<td>0.057</td>
</tr>
<tr>
<td>0.1</td>
<td>0.091</td>
<td>0.076</td>
<td>0.076</td>
<td>0.083</td>
<td>0.088</td>
</tr>
<tr>
<td>0.2</td>
<td>0.192</td>
<td>0.160</td>
<td>0.162</td>
<td>0.175</td>
<td>0.183</td>
</tr>
<tr>
<td>0.3</td>
<td>0.366</td>
<td>0.323</td>
<td>0.326</td>
<td>0.346</td>
<td>0.357</td>
</tr>
<tr>
<td>0.4</td>
<td>0.590</td>
<td>0.542</td>
<td>0.548</td>
<td>0.568</td>
<td>0.583</td>
</tr>
<tr>
<td>0.5</td>
<td>0.804</td>
<td>0.765</td>
<td>0.767</td>
<td>0.780</td>
<td>0.792</td>
</tr>
<tr>
<td>0.6</td>
<td>0.946</td>
<td>0.922</td>
<td>0.923</td>
<td>0.929</td>
<td>0.934</td>
</tr>
<tr>
<td>0.7</td>
<td>0.992</td>
<td>0.988</td>
<td>0.988</td>
<td>0.989</td>
<td>0.990</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Phi = 0.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.058</td>
<td>0.017</td>
<td>0.023</td>
<td>0.063</td>
<td>0.087</td>
</tr>
<tr>
<td>0.1</td>
<td>0.091</td>
<td>0.025</td>
<td>0.032</td>
<td>0.086</td>
<td>0.112</td>
</tr>
<tr>
<td>0.2</td>
<td>0.192</td>
<td>0.049</td>
<td>0.063</td>
<td>0.144</td>
<td>0.180</td>
</tr>
<tr>
<td>0.3</td>
<td>0.367</td>
<td>0.100</td>
<td>0.124</td>
<td>0.248</td>
<td>0.288</td>
</tr>
<tr>
<td>0.4</td>
<td>0.596</td>
<td>0.196</td>
<td>0.223</td>
<td>0.387</td>
<td>0.436</td>
</tr>
<tr>
<td>0.5</td>
<td>0.810</td>
<td>0.336</td>
<td>0.363</td>
<td>0.563</td>
<td>0.614</td>
</tr>
<tr>
<td>0.6</td>
<td>0.949</td>
<td>0.541</td>
<td>0.549</td>
<td>0.756</td>
<td>0.791</td>
</tr>
<tr>
<td>0.7</td>
<td>0.992</td>
<td>0.758</td>
<td>0.753</td>
<td>0.906</td>
<td>0.923</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>0.928</td>
<td>0.898</td>
<td>0.983</td>
<td>0.987</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>0.994</td>
<td>0.968</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 6: Estimated Power for AR(1) Processes
Sample Size = 200

<table>
<thead>
<tr>
<th>Underlying Cross-Correlation</th>
<th>Residual-Based Test</th>
<th>Model-Based, True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.052</td>
<td>0.051</td>
<td>0.051</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>0.1</td>
<td>0.292</td>
<td>0.290</td>
<td>0.291</td>
<td>0.289</td>
<td>0.290</td>
</tr>
<tr>
<td>0.2</td>
<td>0.811</td>
<td>0.815</td>
<td>0.816</td>
<td>0.815</td>
<td>0.816</td>
</tr>
<tr>
<td>0.3</td>
<td>0.990</td>
<td>0.991</td>
<td>0.990</td>
<td>0.991</td>
<td>0.991</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phi = 0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
</tr>
<tr>
<td>0.7</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>0.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phi = 0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
</tr>
<tr>
<td>0.7</td>
</tr>
<tr>
<td>0.8</td>
</tr>
<tr>
<td>0.9</td>
</tr>
</tbody>
</table>
### Table 7: Estimated Power for MA(1) Process

**Sample Size = 30**

<table>
<thead>
<tr>
<th>Underlying Cross-Correlation</th>
<th>Residual-Based Test</th>
<th>Model-Based, True Parameters</th>
<th>Model-Based, Estimated Parameters</th>
<th>Model-Free, Rectangle Window</th>
<th>Model-Free, Triangle Window</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theta = 0.0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.055</td>
<td>0.055</td>
<td>0.051</td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.081</td>
<td>0.083</td>
<td>0.080</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.179</td>
<td>0.187</td>
<td>0.179</td>
<td>0.182</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.346</td>
<td>0.374</td>
<td>0.371</td>
<td>0.361</td>
<td>0.366</td>
</tr>
<tr>
<td>0.4</td>
<td>0.566</td>
<td>0.607</td>
<td>0.604</td>
<td>0.590</td>
<td>0.598</td>
</tr>
<tr>
<td>0.5</td>
<td>0.788</td>
<td>0.827</td>
<td>0.824</td>
<td>0.812</td>
<td>0.821</td>
</tr>
<tr>
<td>0.6</td>
<td>0.932</td>
<td>0.954</td>
<td>0.953</td>
<td>0.947</td>
<td>0.952</td>
</tr>
<tr>
<td>0.7</td>
<td>0.987</td>
<td>0.995</td>
<td>0.994</td>
<td>0.994</td>
<td>0.995</td>
</tr>
<tr>
<td>0.8</td>
<td>0.998</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Theta = 0.3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.053</td>
<td>0.052</td>
<td>0.053</td>
<td>0.055</td>
<td>0.057</td>
</tr>
<tr>
<td>0.1</td>
<td>0.082</td>
<td>0.079</td>
<td>0.079</td>
<td>0.081</td>
<td>0.087</td>
</tr>
<tr>
<td>0.2</td>
<td>0.177</td>
<td>0.166</td>
<td>0.169</td>
<td>0.172</td>
<td>0.178</td>
</tr>
<tr>
<td>0.3</td>
<td>0.347</td>
<td>0.330</td>
<td>0.336</td>
<td>0.336</td>
<td>0.349</td>
</tr>
<tr>
<td>0.4</td>
<td>0.562</td>
<td>0.548</td>
<td>0.551</td>
<td>0.552</td>
<td>0.567</td>
</tr>
<tr>
<td>0.5</td>
<td>0.780</td>
<td>0.769</td>
<td>0.775</td>
<td>0.772</td>
<td>0.787</td>
</tr>
<tr>
<td>0.6</td>
<td>0.925</td>
<td>0.926</td>
<td>0.929</td>
<td>0.928</td>
<td>0.935</td>
</tr>
<tr>
<td>0.7</td>
<td>0.985</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.991</td>
</tr>
<tr>
<td>0.8</td>
<td>0.996</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>0.999</td>
<td>1.000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Theta = 0.7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.066</td>
<td>0.053</td>
<td>0.051</td>
<td>0.056</td>
<td>0.063</td>
</tr>
<tr>
<td>0.1</td>
<td>0.094</td>
<td>0.075</td>
<td>0.074</td>
<td>0.079</td>
<td>0.088</td>
</tr>
<tr>
<td>0.2</td>
<td>0.177</td>
<td>0.143</td>
<td>0.139</td>
<td>0.149</td>
<td>0.164</td>
</tr>
<tr>
<td>0.3</td>
<td>0.333</td>
<td>0.271</td>
<td>0.271</td>
<td>0.281</td>
<td>0.307</td>
</tr>
<tr>
<td>0.4</td>
<td>0.532</td>
<td>0.450</td>
<td>0.451</td>
<td>0.465</td>
<td>0.493</td>
</tr>
<tr>
<td>0.5</td>
<td>0.740</td>
<td>0.672</td>
<td>0.673</td>
<td>0.687</td>
<td>0.712</td>
</tr>
<tr>
<td>0.6</td>
<td>0.882</td>
<td>0.853</td>
<td>0.862</td>
<td>0.869</td>
<td>0.883</td>
</tr>
<tr>
<td>0.7</td>
<td>0.956</td>
<td>0.964</td>
<td>0.968</td>
<td>0.971</td>
<td>0.975</td>
</tr>
<tr>
<td>0.8</td>
<td>0.986</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.9</td>
<td>0.996</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Underlying Cross-Correlation</td>
<td>Residual-Based Test</td>
<td>Model-Based, True Parameters</td>
<td>Model-Based, Estimated Parameters</td>
<td>Model-Free, Rectangle Window</td>
<td>Model-Free, Triangle Window</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------------</td>
<td>-------------------------------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Theta = 0</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.049</td>
<td>0.049</td>
<td>0.049</td>
<td>0.050</td>
<td>0.049</td>
</tr>
<tr>
<td>0.1</td>
<td>0.282</td>
<td>0.284</td>
<td>0.282</td>
<td>0.283</td>
<td>0.283</td>
</tr>
<tr>
<td>0.2</td>
<td>0.809</td>
<td>0.813</td>
<td>0.814</td>
<td>0.813</td>
<td>0.813</td>
</tr>
<tr>
<td>0.3</td>
<td>0.990</td>
<td>0.991</td>
<td>0.991</td>
<td>0.990</td>
<td>0.990</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Theta = 0.3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.049</td>
<td>0.051</td>
<td>0.051</td>
<td>0.052</td>
<td>0.052</td>
</tr>
<tr>
<td>0.1</td>
<td>0.282</td>
<td>0.260</td>
<td>0.260</td>
<td>0.262</td>
<td>0.265</td>
</tr>
<tr>
<td>0.2</td>
<td>0.809</td>
<td>0.755</td>
<td>0.755</td>
<td>0.757</td>
<td>0.761</td>
</tr>
<tr>
<td>0.3</td>
<td>0.990</td>
<td>0.982</td>
<td>0.982</td>
<td>0.982</td>
<td>0.983</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Theta = 0.7</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.049</td>
<td>0.051</td>
<td>0.052</td>
<td>0.051</td>
<td>0.054</td>
</tr>
<tr>
<td>0.1</td>
<td>0.283</td>
<td>0.216</td>
<td>0.218</td>
<td>0.219</td>
<td>0.227</td>
</tr>
<tr>
<td>0.2</td>
<td>0.811</td>
<td>0.657</td>
<td>0.654</td>
<td>0.660</td>
<td>0.669</td>
</tr>
<tr>
<td>0.3</td>
<td>0.989</td>
<td>0.949</td>
<td>0.950</td>
<td>0.952</td>
<td>0.955</td>
</tr>
<tr>
<td>0.4</td>
<td>1</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Conclusion

This study used five tests for cross-correlation with the purpose of investi-gating the possible gain of knowing the true model, or the true parameters, relative to model independent tests. The size and power properties of five tests, each relying on different amounts of information, were investigated via the use of Monte Carlo simulations. It was observed that the size properties are essentially the same for all tests in case of white noise data. For autocorrelated data the size properties diverge; for slowly decaying autocorrelations the residual based test is markedly better than the others, although for rapidly decaying autocorrelations the residual based test is inferior to the others in that it over rejects, thus, none of the tests has uniformly best size properties.

The power properties of the tests are the same for white noise data, but in the case of autocorrelation there are some apparent differences. For slowly decaying autocorrelations the residual based test is markedly better than the others, but for rapidly decaying autocorrelations the power properties are about the same for all tests. It was also observed that the choice of lag window for the model independent estimates is of some importance. The size properties are uniformly better for the rectangular lag window but the power properties are about the same. In general, the residual based test dominates the model independent test in terms of power, but the potency of the residual based test should be weighed against the risk of using a misspecified model.

References


On Exact 100(1−α)% Confidence Interval of Autocorrelation Coefficient in Multivariate Data When the Errors are Autocorrelated

Madhusudan Bhandary
Columbus State University

An exact 100(1−α)% confidence interval for the autocorrelation coefficient $\rho$ is derived based on a single multinormal sample. The confidence interval is the interval between the two roots of a quadratic equation in $\rho$. A real life example is also presented.

Key words: Autocorrelation coefficient, confidence interval, quadratic equation.

Introduction

The autocorrelation coefficient $\rho$ is frequently used to measure the autocorrelation in a time series model. Weather patterns throughout the year change month by month and there is autocorrelation in the weather pattern from one month to the next. Similarly, the behavior of the stock-market pattern from day to day has an autocorrelation effect.

Statistical inference concerning $\rho$ for a single sample problem has been studied by Durbin and Watson (1950, 1951, 1971); some discussions are also given in Morrison (1983). Cochrane and Orcutt (1949) include a discussion about estimating the regression parameters when the errors are autocorrelated. Bhandary (2005) derived a likelihood ratio test for the equality of two autocorrelation coefficients based on two independent multinormal samples. Bhandary and Doetkott (in review) derived a likelihood ratio test for the equality of more than two autocorrelation coefficients based on more than two independent multinormal samples.

In this article, the problem of developing a method of obtaining an exact 100(1−α)% confidence interval for the autocorrelation coefficient $\rho$ based on a single multinormal sample is considered. A confidence interval for $\rho$ is found using the distributional property of a statistic. The confidence interval for $\rho$ is the interval between the two roots of a quadratic equation in $\rho$.

Methodology

Derivation of the Confidence Interval

The model for the multivariate data with autocorrelated error is as follows:

$$x = \mu + \varepsilon,$$

where

$$x = (x_1...x_p)'$$

is a $p \times 1$ vector of observations,

$$\mu = (\mu_1...\mu_p)'$$

is a $p \times 1$ vector of unknown means, and

$$\varepsilon = (\varepsilon_1...\varepsilon_p)'$$

is a $p \times 1$ vector of random errors.

It is assumed that $\varepsilon \sim N_p(0, \Sigma)$, where

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & \cdots & 1 \end{bmatrix},$$

(2.1)
and $N_p$ denotes $p$-variate normal distribution. The structure of the covariance matrix in (2.1) means that the errors are autocorrelated. The autocorrelatedness of the error is common in real practice and it can be tested from the data whether the error covariance structure is of (2.1) or not. In expression (2.1), $\sigma^2$ represents the variance of each error component and $\rho$ is called the autocorrelation coefficient.

Let $x_1, x_2, ..., x_n$ be $p \times 1$ vector of $n$ observations independently and identically distributed as $N_p(\mu, \Sigma)$, where $\Sigma$ is given by (2.1). The following transformation can be made: $u_i = T x_i$, $i = 1, 2, ..., n$, where

$$T = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \ldots & 0 \\ -\rho & 1 & 0 & \ldots & 0 \\ 0 & -\rho & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & -\rho & 1 \end{pmatrix} \tag{2.2}$$

Thus,

$$T \Sigma T' = \sigma^2 (1 - \rho^2) I_p, \tag{2.3}$$

where $I_p$ denotes identity matrix of order $p \times p$. Under the transformation (2.2),

$$u_i \sim N_p(\mu^*, \sigma^2 (1 - \rho^2) I_p), \quad i = 1, 2, ..., n \tag{2.4}$$

where $\mu^* = T \mu$.

Next, consider splitting the sample data into two parts— one with $n_1$ observations $x_1, x_2, ..., x_{n_1}$ (sample 1) and the other with $n_2$ observations $x_{n_1+1}, x_{n_1+2}, ..., x_{n_1+n_2}$ (sample 2) and $n_1 + n_2 = n$. Using the transformation (2.2), the data vector can be transformed from $x_i$ to $u_i$ as follows:

$$u_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{in} \end{pmatrix} \sim N_p(\mu^*, \Sigma^*), \quad i = 1, 2, ..., n \tag{2.5}$$

where $\mu^* = T \mu$, $\Sigma^* = \sigma^2 (1 - \rho^2) I_p$ and $T$ is given by (2.2).

Consider the following statistic:

$$\sum_{i=1}^{n} (u_i - \bar{u})' (u_i - \bar{u}) \tag{2.6}$$

where $u_i$'s are given by (2.5). Because

$$\frac{(n_2-1) \sum_{i=n_1+1}^{n} (u_i - \bar{u})' (u_i - \bar{u})}{(n_1-1) \sum_{i=1}^{n_1} (u_i - \bar{u})' (u_i - \bar{u})} \sim F_{n_1, n_2, p}$$

(using (2.5)), it can be stated that the exact value of the constant $C^*$ can be obtained from the equation

$$\sum_{i=1}^{n_1} (u_i - \bar{u})' (u_i - \bar{u}) > C^* \alpha \quad P(\sum_{i=n_1+1}^{n} (u_i - \bar{u})' (u_i - \bar{u}) > C^* \alpha) = \alpha$$

as

$$C^* = \left( \frac{n_1 - 1}{n_2 - 1} \right) F_{\alpha; n_1, n_2} \tag{2.7}$$

where, $F_{\alpha; n_1, n_2}$ is the upper $100\alpha\%$ point of the F-distribution with d.f. $n_1, n_2$ respectively and $n_2 = n - n_1$.

The inequality inside the probability in (2.7) can be written as
EXACT CONFIDENCE INTERVAL OF AUTOCORRELATION COEFFICIENT

\[
\sum_{i=1}^{n} (x_i - \bar{x})^T T T (x_i - \bar{x}) > C^*
\]  
(2.8)

where \( T \) is given by (2.2).

From (2.2) it may be observed that

\[
T^T = \begin{bmatrix}
1 & -\rho & 0 & 0 & \cdots & 0 \\
-\rho & 1 + \rho^2 & -\rho & 0 & \cdots & 0 \\
0 & -\rho & 1 + \rho^2 & -\rho & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & -\rho & 1 + \rho^2 & -\rho \\
0 & 0 & \cdots & \cdots & -\rho & 1 \\
\end{bmatrix}
\]  
(2.9)

Using (2.9) in (2.8), the following inequality results:

\[
\sum_{i=1}^{n} (x_i - \bar{x})^2 + (1 + \rho^2) \sum_{r=2}^{p-1} (x_{ir} - \bar{x}_r)^2 \\
+ (x_{p} - \bar{x}_p)^2 \\
-2\rho \sum_{r=1}^{p-1} (x_{ir} - \bar{x}_r)(x_{ir+1} - \bar{x}_{r+1}) \\
> C^*
\]

\[
\frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2 + (1 + \rho^2) \sum_{r=2}^{p-1} (x_{ir} - \bar{x}_r)^2}{\sum_{i=n+1}^{n} (x_i - \bar{x}_i)^2 + (1 + \rho^2) \sum_{r=2}^{p-1} (x_{ir} - \bar{x}_r)^2} > C^*
\]

i.e., \( a \rho^2 + b \rho + c > 0 \)  
(2.10)

where

\[
a = \sum_{i=1}^{n} \sum_{r=2}^{p-1} (x_{ir} - \bar{x}_r)^2 - C^* \sum_{i=n+1}^{n} \sum_{r=2}^{p-1} (x_{ir} - \bar{x}_r)^2
\]

\[
b = 2C^* \sum_{i=1}^{n} \sum_{r=1}^{n} (x_{ir} - \bar{x}_r)(x_{ir+1} - \bar{x}_{r+1})
\]

and

\[
c = \sum_{i=1}^{n} \sum_{r=1}^{p} (x_{ir} - \bar{x}_r)^2 - C^* \sum_{i=n+1}^{n} \sum_{r=1}^{n} (x_{ir} - \bar{x}_r)^2
\]

and \( C^* \) is given by (2.7). Note that the data is split and called sample 1 and sample 2 in such a way that \( a > 0 \) i.e., if \( a < 0 \) then sample 2 is called as sample 1 and sample 1 as sample 2, and in that case \( a > 0 \), where \( a \) is given by (2.11).

Let the roots of \( a \rho^2 + b \rho + c = 0 \) be

\[
\hat{\rho}_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}
\]  
(2.12)

and

\[
\hat{\rho}_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}
\]

where \( a, b \) and \( c \) are given by (2.11). Note that in (2.12), if \( b^2 - 4ac \) happens to be negative, then sample 2 is called as sample 1 and sample 1 as sample 2 and, in that case, \( b^2 - 4ac \) becomes positive.

Because \( a > 0 \),

\[
\alpha = P( \sum_{i=1}^{n} (u_i - \bar{u})(u_i - \bar{u}) > C^* ) = P( a \rho^2 + b \rho + c > 0 )
\]

\[
= P(\rho < \hat{\rho}_1, \rho > \hat{\rho}_2) \text{ (using (*))}
\]

92
Therefore, $P(\hat{\rho}_1 < \rho < \hat{\rho}_2) = 1 - \alpha$, where $\hat{\rho}_1$ and $\hat{\rho}_2$ are given by (2.12). Thus, $100(1 - \alpha)\%$ confidence interval for $\rho$ is $(\hat{\rho}_1, \hat{\rho}_2)$ where $\hat{\rho}_1$ and $\hat{\rho}_2$ are given by (2.12).

A Real Life Example
A real data set from Anderson (1976) and Hand, Daly, Lunn, McConway, and Ostrowski (1994) containing a sample of the monthly average air temperature (°F) from January to April at Nottingham Castle for 10 years - hence the data is autocorrelated over months - is used to provide an example. (See Table 1).

Table 1: Average Monthly Temperatures from January to April for Nottingham

<table>
<thead>
<tr>
<th>Obsvn.</th>
<th>Year</th>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1921</td>
<td>44.2</td>
<td>39.8</td>
<td>45.1</td>
<td>47.0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1937</td>
<td>40.8</td>
<td>41.0</td>
<td>38.4</td>
<td>47.4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1922</td>
<td>37.5</td>
<td>38.7</td>
<td>39.5</td>
<td>42.1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1934</td>
<td>39.4</td>
<td>38.2</td>
<td>40.4</td>
<td>46.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1927</td>
<td>40.6</td>
<td>38.5</td>
<td>45.3</td>
<td>47.1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1920</td>
<td>36.2</td>
<td>40.8</td>
<td>44.4</td>
<td>46.7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1933</td>
<td>39.2</td>
<td>39.3</td>
<td>44.5</td>
<td>48.7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1926</td>
<td>40.0</td>
<td>43.4</td>
<td>43.4</td>
<td>48.9</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1935</td>
<td>40.0</td>
<td>42.6</td>
<td>43.5</td>
<td>47.1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1925</td>
<td>40.0</td>
<td>40.5</td>
<td>40.8</td>
<td>45.1</td>
<td></td>
</tr>
</tbody>
</table>

Results
Data in Table 1 is a p-variate data set for $p = 4$ and $n = 10$; it is split into two parts as $n_1 = 5$ and $n_2 = 5$, and sample 1 is the first 5 observations and sample 2 is the remaining 5 observations.

The formulas for computing the sum of square and sum of products are as follows:

Using formula (2.11) results in $a = 4.9997$, $b = 25.7179$ and $c = 3.4117$ and $\alpha = 0.05$ is used, hence $F_{0.05, 20, 20} = 2.12$. Using (2.12), results in a 95% confidence interval for $\rho$ as (-5.0076, 0.1363) which is approximated as (-1.0, 0.1363) because $|\rho| < 1$.

References

Bhandary, M., & Doetkott, C. Test for the equality of autocorrelation coefficients for several populations in multivariate data when the errors are autocorrelated. Submitted to *American Journal of Mathematical and Management Science*.


Beyond Alpha: Lower Bounds for the Reliability of Tests

Nol Bendermacher
Radboud University,
Nijmegen, The Netherlands

The most common lower bound to the reliability of a test is Cronbach’s alpha. However, several lower bounds exist that are definitely better, that is, higher than alpha. An overview is given as well as an algorithm to find the best: the greatest lower bound.

Key words: test reliability; Cronbach’s alpha.

Introduction
The concept of reliability is based on the notion of accuracy or precision of a measurement. This article is confined to the reliability of tests - psychological or other - consisting of a number of items and to the situation where a test is administered only once. A person’s score on such a test is the sum of his/her scores on the individual items.

According to classical test theory, the score $x_{ij}$ of person $i$ on item $j$ consists of two parts: the true score $\tau_{ij}$ and an error component $\varepsilon_{ij}$: $x_{ij} = \tau_{ij} + \varepsilon_{ij}$. Moreover, classical test theory assumes that the error components are uncorrelated with the true parts as well as with each other. As a consequence the covariance matrix $\Gamma$ of the items is the sum of two components: the covariance matrix ($\Gamma_\tau$) of the true parts and the covariance matrix ($\Gamma_\varepsilon$) of the error components:

$$\Gamma = \Gamma_\tau + \Gamma_\varepsilon$$

The assumption of uncorrelated errors implies that $\Gamma_\varepsilon$ is a diagonal matrix; thus the off-diagonals of $\Gamma$ and $\Gamma_\tau$ are identical. The assumption of independent errors is essential for all measures discussed herein. Many conditions exist which lead to the violation of the assumption of independent errors, for example: in a test with a time limit where an unanswered item results in a minimum score, the errors of the last items may correlate, or in a long or difficult test errors may become correlated due to the effect of fatigue or declining motivation during the test administration.

The reliability of a test consisting of $v$ items is defined as:

$$\rho_{tt} = 1 - \frac{\sigma_e^2}{\sigma_t^2}$$ (1)

where $\sigma_e^2$ is the error variance and $\sigma_t^2$ is the total variance of the test scores:

$$\sigma_t^2 = \sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}$$ (2)

$$\sigma_e^2 = \text{TR}(\Gamma_\varepsilon) = \sum_{i=1}^{v} \Gamma_{eii}$$ (3)

Based on these formulae the definition of reliability can be rewritten as:

$$\rho_{tt} = 1 - \frac{\sigma_e^2}{\sigma_t^2} = 1 - \frac{\sum \Gamma_{eii}}{\sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}}$$ (4)
It should be noted that this definition leaves undecided whether the unique variances (item variance components not correlated with any other item) are treated as error or as true variance. The lower bounds discussed herein are lower bounds according to both definitions.

Lower Bounds

If no other assumptions are added to those of the classical model it is impossible to assess the reliability of a test from a single administration; only lower bounds can be derived. From (4) it is clear that - given covariances $\Gamma$ - the reliability is maximal if the trace of the error covariance matrix $\Gamma_e$ is minimal. As Jackson and Agunwamba (1977) remarked, the only restrictions that the classical model imposes on the elements of $\Gamma_e$ are

1. $0 \leq \Gamma_{e ii} \leq \Gamma_{ii}$, and
2. $\Gamma_e = \Gamma - \Gamma_e$ is non-negative definite.

Thus, if the set of values $\Gamma_e$ that maximizes its trace $\sum_i \Gamma_{e ii}$ under these restrictions can be located, the result would give the smallest possible value for the reliability given the covariance matrix $\Gamma$; this value is the greatest possible lower bound to the reliability. Jackson and Agunwamba (1977) and ten Berge, Snijders and Zegers (1981) described algorithms to find this largest lower bound; however, several well-known lower bounds are first put forth.

Guttman (1945) introduced a series of lower bounds called $\lambda_1$ through $\lambda_6$.

$$\lambda_1: \frac{1}{\sqrt{v-1}} \sum_{i \neq j} \sqrt{\sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}^2}$$

This $\lambda_1$ is the sum of the off-diagonal cells in $\Gamma$ divided by the sum of all cells. The larger the item covariances, as compared to the variances, the larger $\lambda_1$.

$$\lambda_2 = \lambda_1 + \sqrt{\frac{\sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}^2}{\sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}^2}}$$

Because $\lambda_2 \geq \lambda_1$, $\lambda_2$ should always be preferred over $\lambda_1$.

$$\lambda_3: \frac{v}{v-1} \lambda_1 = \frac{v}{v-1} \left( 1 - \frac{\sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}}{\sum_{i=1}^{v} \sum_{j=1}^{v} \Gamma_{ij}} \right)$$

This $\lambda_3$ is better known as Cronbach’s alpha. Guttman (1945) remarked “$\lambda_3$ is easier to compute than $\lambda_2$, since only the total variance and the item covariances are required. If the covariances are all positive and homogeneous, then $\lambda_3$ will not be much less than $\lambda_2$ and may be an adequate lower bound. If the covariances are heterogeneous, and in particular, if some are negative, then $\lambda_2$ will be definitely superior to $\lambda_3$. $\lambda_2$ can be positive and useful when $\lambda_3$ is negative and useless” (pp. 274-275). In brief, $\lambda_1 \leq \lambda_3 \leq \lambda_2$. Therefore, with modern computational facilities, $\lambda_2$ should always be preferred over $\lambda_3$. In actual practice, however, researchers tend to use $\lambda_3$, which is better known as Cronbach’s alpha or, with dichotomous items, the Kuder-Richardson 20 (KR20).

Ten Berge and Zegers (1978) showed that $\lambda_3$ and $\lambda_2$ are members of a series of bounds $\mu_0, \mu_1, \mu_2, ...$, defined by the following general formula:

$$\mu_r = \frac{1}{\sigma_r^2} \left[ p_0 + \left( p_1 + \left( p_2 + ... \left( p_{r-1} + \left( p_r \frac{1}{2} \right) \frac{1}{3} \right) \frac{1}{2} \right) \frac{1}{3} \right) \frac{1}{2} \right]$$

where $r = 0, 1, 2, ...$

$$p_h = \sum_{i \neq j} \sigma_{ij}^2, h = 1, 2, ..., r-1$$
\[ p_h = \frac{v}{v-1} \sum_{i \neq j} \sigma_{ij}^{(2^h)}, h = r \]

From this formula it is observed that \( \mu_0 = \lambda_3 = \text{Cronbach's alpha} = \text{KR20} \) and \( \mu_1 = \lambda_2 \). The differences between \( \mu_{r+1} \) and \( \mu_r \) rapidly converge to zero, thus, there is not much use in going further than \( \mu_3 \).

\[ \lambda_4: 2 \left( 1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_3^2} \right) \]  

(9)

where \( \sigma_1^2 \) and \( \sigma_2^2 \) are the variances of two test halves:

\[ \sigma_1^2 = \sum_{i \neq j} \sum_{j} \Gamma_{ij} \]

where \( i \) and \( j \) run over the items in the first test half and similarly,

\[ \sigma_2^2 = \sum_{i \neq j} \sum_{j} \Gamma_{ij} \]

with \( i \) and \( j \) running over the items in the second test half.

A problem with \( \lambda_4 \) is that many ways exist by which to split a test into two parts, meaning that there are many different values for \( \lambda_4 \): the most interesting of them is the largest. In the statistical package SPSS (release 15.0.0) the value of \( \lambda_4 \) depends on the order of the items in the scale: it assigns the first \( v/2 \) items (with odd \( v \) the first \( (v+1)/2 \)) to the first test half and the remaining items to the second half.

A simple algorithm to find a good split is based on the following: Imagine that the rows and columns of the covariance matrix are rearranged such that the items of the first test half come first, \( \sigma_1^2 \) and \( \sigma_2^2 \) are the sums of the upper left and the lower right quarter of the covariance matrix \( \Gamma \) respectively. Because the sum (\( \sigma_3^2 \)) of the entire matrix is fixed, \( \lambda_4 \) is maximal if the sum of the lower left (and the upper right) quarter is maximal. This leads to the following algorithm:

1) Locate the pair of items with the highest covariance and assign one of them to test 1 and the other to test 2.
2) Try each ordered pair \((i, j)\) of items not yet assigned. Compute the covariance between the two test parts if item \( i \) is assigned to test 1 and item \( j \) to test 2. After all pairs are tried, make the assignment that resulted in the highest covariance between the tests.
3) Repeat step 2 until all items have been assigned to one of the test halves. In the case of an odd number of items, the last item is added to the group for which the mean covariance with the item is the smallest.

Given a specific split, Jackson and Agunwamba (1977) described a method to determine whether the resulting value of \( \lambda_4 \) is the greatest possible lower bound. Define: \( b \text{ a vector with } v \text{-elements, with } b_i = 1 \text{ if item } i \text{ belongs to test half 1 and } b_i = -1 \text{ if it belongs to test half 2}; A = \Gamma b; \theta_i = b_i A_i, i = 1, v \) (this \( \theta \) is the vector with error variances); and \( \Gamma_i = \Gamma - \text{diag}(\theta_i) \). If \( \Gamma_i \) is non-negative definite and all \( \theta_i \geq 0 \), \( \lambda_4 \) is the greatest possible lower bound.

\[ \lambda_5: \lambda_1 + \frac{2 \sqrt{\max_i (\Gamma_{ii})}}{\sum \sum (\Gamma_{ij})} \text{ with } \Gamma_{ii} = \sum \sum (\Gamma_{ij})^2 - \Gamma_{ii}^2 \]  

(10)

As Guttman (1945) noted, this measure will be larger than \( \lambda_2 \) if one item has large covariances with the other items compared with the covariances among those items. Otherwise \( \lambda_5 \) is less than or equal to \( \lambda_2 \).

\[ \lambda_6: 1 - \frac{\sum_i \Gamma_{ii} (1 - P_i^2)}{\sum \sum \Gamma_{ij}} = 1 - \frac{\sum_i \frac{1}{\Gamma_{ii}^{-1}}}{\sum \sum \Gamma_{ij}} \]  

(11)

where \( \Gamma_{ii}^{-1} \) denotes the \( i \text{th} \) diagonal of the inverse of \( \Gamma \). In these formulæ \( P_i^2 \) is the squared multiple correlation in the multiple regression of
item \(i\) on the remaining \(v-1\) items: 
\[ P_i^2 = 1 - \frac{1}{P_{ii}^{-1}}. \]
\((P_{ii}^{-1})\) denotes the \(i\)th diagonal of the inverse of the correlation matrix from \(\Gamma\).

Guttman (1945) explained that \(\lambda_6\) will be larger than \(\lambda_2\) if the multiple correlations are relatively large as compared to the zero-order correlations. Otherwise \(\lambda_6\) will tend to be less than or equal to \(\lambda_2\). Jackson and Agunwamba (1977) reported that \(\lambda_6\) should be particularly advantageous in the fairly typical situation where the inter-item correlations are positive, moderate in size and somewhat similar. Jackson and Agunwamba (1977) added a seventh bound, called \(\lambda_7\):

\[ \lambda_7: 1 - \sum_{i=1}^{v} \frac{\sigma_i^2}{\sigma_i^2} \left[ \sqrt{\frac{v}{v-1}} \sum_{j \neq i} d_{ij}^2 \right] \]  

(12)

where \(d_{ij}^2\) is defined as follows:

\[ g = \text{the value of } j \text{ for which } \frac{\Gamma_{ij}^2}{\Gamma_{jj}} \text{ is largest} \]

\[ k = \text{the value of } i \neq j \text{ for which } \frac{\Gamma_{ij}^2}{\Gamma_{ii}} \text{ is largest} \]

\[ r_{ij} = \text{the correlation between items } i \text{ and } j \]

\[ d_{ij}^2 = \sigma_i^2 \sigma_j^2 \max\left( r_{ig}^2 r_{kj}^2, r_{ij}^2 \right) \]

Jackson and Agunwamba remarked that this bound will be substantially better than \(\lambda_2\) when there is considerable variation among the squared correlations.

Woodhouse and Jackson (1977) showed some partial orders in the bounds \(\lambda_1\) through \(\lambda_7\): \(\lambda_1 \leq \lambda_3 \leq \lambda_2 \leq \lambda_7\), \(\lambda_1 \leq \lambda_4\), \(\lambda_1 \leq \lambda_5\), \(\lambda_1 \leq \lambda_6\). Table 1 shows a covariance matrix of four items and the lower bounds discussed for the reliability of their sum.

| Table 1: Variances and Covariances of Four Items and Lower Bounds for the Reliability of their Sum |
|-----------------------------------------------|-----------------------------------------------|
| \(\lambda_1\) | \(0.3992\) |
| \(\lambda_2\) | \(\mu_1 = 0.5867\) |
| \(\lambda_3\) | \(\alpha = \mu_0 = 0.5323\) |
| \(\lambda_4\) | \(0.5574\) |
| \(\lambda_5\) | \(0.6125\) |
| \(\lambda_6\) | \(0.5817\) |
| \(\lambda_7\) | \(0.5904\) |
| \(\mu_2\) | \(0.5936\) |
| \(\mu_3\) | \(0.5957\) |

When Experimental Independence Does Not Hold

Guttman (1953) provided some lower bounds for the situation where the assumption of independent errors does not hold by introducing an additional quantity \(\delta\), for which in some specific situations upper bounds can be defined. Such situations are tests with a time limit and more general tests where the completion of an item depends on the completion of its predecessor. The adjusted measures are:

\[ \lambda_1^* = \lambda_1 - \frac{\delta}{\sigma_i^2} \]  

(13)

\[ \lambda_2^* = \lambda_2 - \frac{\delta}{\sigma_i^2} \]  

(14)

\[ \lambda_3^* = \frac{v}{v-1} \lambda_1^* \]  

(15)

\[ \lambda_4^* = \lambda_4 - \frac{2\delta}{\sigma_i^2} \]  

(16)

For the situation where the assumption of uncorrelated errors is violated only by the fact that the completion of an item depends on the completion of its predecessor, Guttman (1955) gives three upper bounds for \(\delta\), assuming that an item that is omitted results in the lowest possible score. Thus, the assumption of uncorrelated errors is weakened to the following: “The basic assumption from now on is that, if person \(i\) attempts item \(j\), then his score on any later item...”
g (g > j) will be experimentally independent of his score on this attempted item j. That is, we are considering here the case where dependence is due solely to *omissions*, so that if a part is attempted, no further experimental dependence holds" (Guttman, 1955, p. 119). Defining

\[ \nu' = \text{the number of items with a non-zero variance} \]

\[ l_i = \text{the minimum score on item i, also the score for an unattempted item} \]

\[ h_i = \text{the maximum score on item i} \]

\[ m_i = h_i - l_i \]

\[ x_i = \text{mean score on item i with } l_i \text{ subtracted} \]

\[ p_i = \text{proportion of persons that attempt item i} \]

the (estimates of the) upper bounds for \( \delta \) are:

\[ d_1 = \sum_{i=1}^{\nu'} \frac{m_i x_i (1 - p_i)}{\sum_{j=i+1}^{\nu'} m_j} \]  

(17)

If \( m_i = 1 \) for all i, this formula reduces to:

\[ d_1 = \frac{\sum_{i=1}^{\nu'} \nu' - i}{\nu'} \cdot x_i (1 - p_i) \]  

(18)

\[ d_2 = 2 \sum_{i=1}^{\nu'} \left( m_i \sqrt{1 - p_i} \sum_{j=i+1}^{\nu'} \sqrt{m_j x_j} \right) \]  

(19)

If \( m_i = 1 \) for all i, this formula reduces to:

\[ d_2 = 2 \sum_{i=1}^{\nu'} \sqrt{1 - p_i} \sum_{j=i+1}^{\nu'} \sqrt{x_j} \]  

(20)

\[ d_3 = 2 \sum_{i=1}^{\nu'} \sum_{j=i+1}^{\nu'} e_{ij} \]  

(21)

where

\[ e_{ij} = \max( m_i \sqrt{m_j x_j (1 - p_i)}, m_j \sqrt{m_i x_i (1 - p_j)} ) \]  

(22)

From these formulas it is clear that \( d_3 \) is at least as high as \( d_1 \) and \( d_2 \).

Finding the Greatest Lower Bound

Woodhouse and Jackson (1977) described an algorithm that finds the greatest lower bound (GLB) for the reliability of a test if only the assumptions of classical test theory hold. However, ten Berge, Snijders and Zegers (1981) showed that this algorithm will not always produce the correct lower bound. They described another algorithm that avoids these shortcomings and also is less time consuming. The algorithm, as implemented in this study, proceeds as follows: Define:

\[ C = \text{the given covariance matrix} \]

\[ C_0 = C - \text{Diag}(C) \]

and

\[ R_i = \text{The } i\text{-th row of } C \]

1) Construct a \( v \) by \( r \) matrix \( T \) with \( r \leq v \) and not too small. Ten Berge, Snijders and Zegers (1981) advised that \( r = \text{the number of non-negative eigenvalues in } \Gamma_0 \). In order to be safe, choose \( r = v \).

Similar to Bentler & Woodward (1980) the cells \( T_{ij} \) of \( T \) are defined as follows:

\[ \text{If } i > j \text{ } T_{ij} \text{ is set to } - \frac{1}{\sqrt{i}} \]

\[ \text{If } i = j \text{ } T_{ij} \text{ is set to } \frac{1}{\sqrt{i}} \]

\[ \text{If } i < j \text{ } T_{ij} \text{ is set to } 0 \]

(By this choice all rows have length 1.)

2) Perform the following steps for each row \( i \) of \( T \):

2.1) Compute \( a = \text{MIN}(0, \sqrt{R_i^T T T^T R_i}) \); \( a \) is the provisional estimate of the true variance of item \( i \).
2.2) If $0 < C_{ii} < a$, row $i$ is replaced by
\[ -\frac{1}{C_{ii}} T^T R_i \]
If $0 < a \leq C_{ii}$, row $i$ is replaced by
\[ -\frac{1}{a} T^T R_i. \]

2.3) If $a = 0$, rescale row $i$ to length 1.

3) Compute the (estimated) sum of error variances: $E = TR(T^TCT)$ and check for convergence. The process has converged if the following conditions hold:
   a) $E$ has not (sufficiently) decreased since the last check
   b) All rows of $T$ have length $\geq 1$

If the process has not converged go back to step 2.

4) Compute the resulting estimate of $\Gamma_t$ by copying $C$ and replacing $C_{ii}$ by
\[ \text{MIN}\left(C_{ii}, \sqrt{R_i^T TT^T R_i}\right), \quad i = 1, v \]
and check whether its smallest eigenvalue is zero. If not, the whole procedure should be repeated with another starting value of $T$, but we wonder if such a situation will ever occur.

5) Define $E = C_{ii} - C_{iii}$ and estimate
\[ \text{GLB} = 1 - \frac{E}{\sum_{i=1}^{v} \sum_{j=1}^{v} C_{ij}} \quad (23) \]

If this algorithm is applied to the example of Table 1 the result is GLB = 0.7324. Ten Berge and Sočan (2004) provide several sources from which other programs can be obtained that compute the greatest lower bound.

The Effect of Sampling Error
A problem exists with several of the lower bounds described in the preceding text. When estimated from a small sample, $\lambda_4$, $\lambda_5$, $\lambda_6$, $\lambda_7$ and GLB will capitalize on chance; meaning that their estimates from the sample tend to overestimate the true population values. As Shapiro and ten Berge (2000) remarked: “It is well known that the g.l.b., based on small samples (even a sample of one thousand subjects is not generally enough) may severely overestimate the population value, and statistical treatment of the bias has been badly missing” (p. 413). They show that bias tends to increase with decreasing sample size and with lower values of GLB. Moreover, the bias is expected to be larger with more parameters to be estimated, that is, with greater $v$.

In absence of an analytical solution the use of brute (computing) force is suggested. The following bootstrapping approach could be used:

1) Compute from the sample covariance matrix $C$ the selected lower bound, $G_0$. If the sample from which $C$ is computed is available, steps 2 through 5 may be skipped and the sample plays the role of $X$ in step 6.

2) Generate a $n$ by $v$ matrix $F$, filled with drawings from a standard normal distribution; $n$ must be not too small and always larger than $v$: 1,000 or 2,000 is adequate.

3) Rotate the columns of $F$ to orthogonality and scale them to mean 0 and length $\sqrt{n}$; $F$ will act as the set of components from a principal components analysis.

4) Perform a principal components analysis on $C$, resulting in a diagonal matrix $\Lambda$ with eigenvalues and the matrix $V$ with the corresponding eigenvectors. Compute the factor matrix $A = V\Lambda^{\frac{1}{2}}$ and make sure that $A$ is square; add zero columns if needed.

5) Construct the matrix $X = FA^T$. The resulting $X$ has a multivariate normal distribution with covariance matrix $\frac{1}{n}X^T X = C$.

6) Draw $k$ random samples from $X$. For each estimate the covariance matrix and the
chosen lower bound. The sampling consists of random selections (with replacement) of rows from $X$. Compute the mean $G_m$ and the standard deviation $s_m$ of these lower bound estimates. The standard error $\frac{s_m}{\sqrt{k}}$ may be used as a stop criterion during the simulations.

7) The difference $G_0 - G_m$ is an estimator of the bias by capitalization on chance and $G_0$ is corrected by taking $2G_0 - G_m$ instead. The correction may not be perfect, but it will be close if the sample is not too small and $G_0$ not too great.

In the example of Table 1 and assuming a sample size $n = 200$ the bias is estimated as 0.002839; taking $n = 100$ the bias estimate becomes 0.002942. These bias estimates are very small, possibly due to the small number of items.

A computer program, called Reliab, that computes some of the lower bounds to the reliability, including the GLB, is available at http://www.ru.nl/socialewetenschappen/rtog/software/statistische/kunst/

The Factor Analytic Approach

Factor analysis explains the correlations between a set of items by a limited set of underlying latent variables, called factors. The model allows the estimate for scores of individuals on the factors as weighted sums of their item scores. In this model it is possible not only to find lower bounds, but also to find real estimates of the reliability of the estimated factor scores from a single test administration.

In factor analytic models, the variance of an item is viewed as composed of two parts:

1) Common variance, i.e. variance that is shared with other items, and

2) Unique variance (or unicity), i.e. variance that is unique for the item: it consists of specificity and genuine error.

Defining:

$Z = n \times v$ matrix of standardized scores ($z$-scores) of $n$ individuals on $v$ items.

$F = n \times v$ matrix of true scores of the individuals on $f$ factors; $F$ is unknown.

$B_z = v \times f$ matrix of weights to estimate the factor scores $F$ from the item scores $Z$: $\hat{F} = ZB_z$

Assume that the weights are scaled such that the variances of $\hat{F}$ (i.e., diagonal values of $\frac{1}{n} \hat{F}^T \hat{F}$) are unity. Thus if

$A = v \times f$ factor pattern, i.e., the matrix containing the weights of the factors in the reconstruction of $Z$: $Z = FA^T + \text{error} + \text{unicities};$

$S = v \times f$ factor structure; it contains the correlations between $Z$ and $F$;

$U = v \times v$ diagonal matrix with unicities; and

$R_{ff} = f \times f$ matrix with correlations between the factors;

then the correlations between the factors and the factor score estimates are:

$$ R_{ff} = \frac{1}{n} F^T \hat{F} $$

$$ = \frac{1}{n} F^T ZB_z $$

$$ = \frac{1}{n} F^T (FA^T + U) B_z $$

$$ = \frac{1}{n} F^T FA^T B_z $$

$$ = R_{ff} A^T B_z $$

$$ = S^T B_z $$

If this model is adhered to, the latent factors play the role of true scores, and although they are latent, $R_{ff}$ contains estimates of the correlations between them and the factor score.
estimates. The squares of these correlations can be interpreted as the reliabilities of the factor score estimates. This measure is also called the “factor determinacy” (McDonald, 1974, p. 213).

In this context two remarks must be made:

1) Factors, as they result from a factor analysis, are not completely defined: they function as axes in an f-dimensional space and any other set of f axes in that space will explain the correlations between the items equally well. Therefore, the orientation of the factors must be selected on the basis of additional criteria, for example their interpretability from a given theory, and

2) As with all regression models, the squared correlations between factors and factor score estimates tend to be inflated, especially when the analysis is based on a small sample.

Conclusion
A number of lower bounds to the reliability of a test have been discussed; all are based on the covariance matrix of the items in the test. It is clear that the most commonly used measure, known as Cronbach’s alpha, KR20 or \( \lambda_3 \), is a poor choice; its only advantage over Guttman’s \( \lambda_2 \) is its ease of computation by hand.

It is clear that - under the assumptions of the classical test theory and without additional assumptions - the measure known as the Greatest Lower Bound is the highest possible lower bound. Its only weakness, one shared with several of the other measures, is its sensibility to capitalization on chance if it is estimated from a relatively small sample. In the absence of analytical methods to correct this bias a bootstrapping approach using brute computing force is suggested in order to minimize the bias.

References


Impact of Measurement Model Modification on Structural Parameter Integrity
When Measurement Model is Misspecified

Weihua Fan
University of Houston

In the process of model modification, parameters of residual covariances are often treated as free parameters to improve model fit. However, the effect of such measurement model modifications on the important structural parameter estimates under various measurement model misspecifications has not been systematically studied. Monte Carlo simulation was conducted to compare structural estimates before and after measurement model modifications of adding residual covariances under varying sample sizes and model misspecifications. Results showed that researchers should pay attention when such measurement model modifications are made to initially misspecified model with missing path(s).

Key words: Structural equation modeling; modification indices; Lagrange multiplier (LM) tests; residual covariance; misspecification.

Introduction
Model modification, also known as specification search, has been widely used in the application of structural equation modeling, in hope of improving models. After obtaining a model that fails to meet accepted goodness of fit standards, many researchers frequently turn toward model modification information in an attempt to find a parsimonious model to fit the sample data. Two approaches commonly used for model modification are the Lagrange Multiplier test (LM test; also referred to as modification index) and the Wald (W) test. The LM test reduces constraints by freely estimating parameters such as residual covariances that are currently fixed (usually to zero). The W test increases constraints by fixing parameters that are currently free (Bentler, 1995; Chou & Bentler, 1990). The more applied approach is usually the LM test, because freely estimating parameters rather than fixing parameters improves model fit.

Anderson and Gerbing (1988) proposed a two-step SEM process as some guidance for latent variable path models which is commonly recommended in practice (see, Kline, 2004; Schumacker & Lomax, 2004). The SEM process consists of two steps: the measurement step and the structural step. The measurement step in the process considers the measurement model, which specifies the relations between the underlying factors and the measured variables. It allows researchers opportunities to improve the data-model fit through model modification within the measurement model while temporarily inserting a saturated latent structure. After obtaining satisfactory data-model fit in the first step, the second step involves the structural model which hypothesizes relations between the latent variables. Assessing the structural relations in an SEM application is usually the focal point of an investigation. Examples of this two-step process include studies by Mattanah, Hancock, and Brand (2004), Joiner, Leveson, and Langfield-Smith (2002) and Chong and Chong (2002). Also see Mulaik and Millsap (2000) for a four-step process and Green, Thompson and Poirier (1999) for a 2-stage specification search procedure, and Green, Thompson and Poirier (2001) for an adjusted Bonferroni method for eliminating parameters in specification searches.

Although the application of the model modification procedures is commonly observed...
across disciplines (Brekler, 1990; Hutchinson, 1998; MacCallum, Roznowski & Necowitz, 1992), it is data-driven in nature and is characterized by capitalization on chance (e.g., MacCallum, Roznowski & Necowitz, 1992). Researchers have thus become prudent with model modifications regarding important theory related parameters such as structural parameters or loading parameters. In the meantime, parameters of residual covariances, which are much less theoretically concerned, are often treated as free parameters to help improve the model-data fit. This is exemplified in work by Newcomb and Bentler (1988) in which 77 residual covariances were added prior to structural analyses.

Given that the primary purpose of latent variable path models is to assess theoretical relations between latent variables (e.g., Kerlinger, 1986), it is not uncommon that parameters of residual covariances are used in abundance to help improve model fit. The liberal manner in which residual covariances are made to measurement models does not appear to be due to a prevailing belief that the true measurement model will be found as a result; research (see, e.g., MacCallum, Roznowski & Necowitz, 1992) has documented that sample-based respecifications seldom arrive at the true population model with consistency.

Two types of errors can result in model modifications such as incorrectly reducing fixed constraints: (1) errors due to sampling fluctuation, and (2) errors due to misspecification (Green, Thompson & Poirier, 1999). Errors related to sampling fluctuation occur if model modifications fit the specific characteristics of sample data but not the population. Errors due to misspecification can occur in two situations. When the initially hypothesized model is correctly specified, model modification of freely estimated correctly fixed parameters to maximally increase model fit is unnecessary and results in Type I error capitalizations on random sample covariation. Fan and Hancock (2006) have shown that the overspecification of measurement model modifications impacts the structural parameter estimates under certain conditions, however, this effect is usually negligible. When the initially hypothesized model is incorrectly specified, which is almost inevitable in practice, model modifications to revise the model by freely estimating additional parameters can lead to two possible results: a revised model with less misspecification errors than the initially hypothesized model or a revised model with a greater number of errors due to misspecification. Two common types of measurement model misspecification include: (1) relevant parameters are incorrectly fixed to zero, and (2) irrelevant parameters that should be fixed to zero are freely estimated.

If a researcher starts with an initial model with incorrectly fixed relevant parameters set to zero, it is expected that freely estimating incorrectly fixed parameters such as cross-loadings or residual covariances would be beneficial to the structural parameters. However, incorrectly fixing relevant parameters to zero can manifest in significant residual covariances, which exist as a function of fixing the truly non-zero parameters (e.g., cross-loadings) to zero. With this scenario it is likely that the LM test also suggests to freely estimate residual covariances that are correctly fixed to zero, this results in a model with more misspecification errors. In addition, if a researcher starts with an initial model incorrectly freely estimating irrelevant parameters that should be fixed to zero, it is expected that the LM test suggesting more freely estimating incorrectly fixed parameters will lead to a model with more misspecification errors. This begs the question of whether the prevailing modification of freely estimating residual covariances is indeed price-free to our structural parameter estimates under measurement model misspecifications. No study of specification searches has considered this issue, thus the current focus of this investigation.

Recently with the development of the statistical software, a handful of Monte Carlo studies have been conducted to assess the performances of model modifications. Three such studies on model modifications are Hutchinson (1998), Chou and Bentler (2002) and Fan and Hancock (2006). Hutchinson (1998) extended the work of MacCallum, Roznowski and Necowitz (1992) to examine the sampling stability of post hoc model modifications. The study found that modifications tended to be inconsistent unless the sample size is very large.
or the model is large. Chou and Bentler (2002) focused on the W test procedure and found a satisfactory success rate of the model modification of structural relationships among factors. However, the results of the test are based on the assumptions of a correct measurement model and a known sequence of latent factors, which pose great challenges to applied researchers. Other relevant studies include Green and Babyak (1997), Green, Thompson and Babyak (1998), Hutchinson (1993), Kaplan (1988, 1989), and Silvia and MacCallum (1988).

Fan and Hancock (2006) investigated the impact of measurement model respecification on structural parameter integrity. Their study compared interfactor correlations before and after measurement model respecifications of crossloadings, intrafactor residual covariances and interfactor residual covariances for a five-factor confirmatory model. The research suggested that some effect on structural parameter estimates arises under conditions of modification in the measurement model; however, in general, the impact is negligible. Although some comfort has been provided for researchers regarding the two-step process of measurement model modification, the study is limited in two ways. First, the study focused on the impact of overspecification of measurement model modifications on structural parameters where paths were added to a correctly specified model. Second, it studied models with interfactor correlations, which appear to be less representing in practice than models involving direct structural relations.

This investigation extends the study of Fan and Hancock (2006) by assessing the effects of freely estimating fixed residual covariances under conditions of measurement model misspecification for a theoretical model involving direct structural relations. This study aims to determine whether model modifications of freely estimating fixed residual covariances are structurally benign under the situations of measurement model misspecification with the goal of gaining insights into the extent to which caution must be exercised in measurement model modification prior to structural model evaluation. For a three-factor confirmatory factor model, Monte Carlo simulation is used to compare structural estimates before and after measurement model modifications under varying sample sizes and model misspecifications.

Methodology

Model Specification

The true model derived from an example illustrated by Paxton, et al. (2001) is a three-factor model measured by nine observed variables. Seven out of nine observed variables load on a single factor and the remaining two load on two factors (See Figure 1). In addition, factor two is regressed on factor one with the coefficient 0.6, and factor three is regressed on factor two with the same coefficient. The model has factor loadings (λ) with an unstandardized value of 1.0 and standardized value of 0.70, while the two cross-loadings have an unstandardized coefficient of 0.30 and a standardized coefficient of 0.21.

Five model specifications are considered in the study. The first model correctly specifies the structure that exists in the population, which is the true model (specification 1). The second model is misspecified by omitting one relevant cross-loading path linking factor 1 with item 4 (specification 2); that is, one relevant cross-loading path is incorrectly fixed to zero. The third model is misspecified by additionally omitting one more relevant cross-loading path linking factor 2 with item 7 (specification 3); that is, one more relevant cross-loading path is incorrectly fixed to zero. The fourth model is misspecified by containing one irrelevant loading from factor 2 to item 3 (specification 4), thus one irrelevant loading path that should be fixed zero is freely estimated. The fifth model is misspecified by containing one additional irrelevant loading from factor 3 to item 6 (specification 5), so one more irrelevant loading path which should be fixed zero is freely estimated.

Note that all modifications of residue covariance suggested to the above models do not exist in the population measurement model. For specification 1, the overspecification of the measurement model simply constitutes Type I error capitalizations on random sample covariation. For the remainder of the specified
true models, the modifications of residue covariance are incorrectly freed to maximally improve data-model fit, leading to a revised model with more misspecification errors.

Sample Size, Replications, Data Generation and Modeling

Four different sample sizes are manipulated in the study, ranging from small to large: 100, 200, 400 and 800. A total of 5 (model specifications) × 4 (sample sizes) = 20 experimental conditions. For each condition, seven separate runs were conducted. Data generation and estimation were carried out using EQS 6.1 (Bentler, 1998) and SAS IML (1990). Within each of the 20 conditions, multivariate normal data were generated and modeled with ML estimation as described by Paxton, et al. (2001).

For each replication within each cell, each of the specified models was first imposed upon the sample data yielding structural estimates of $\phi_1$ and $\phi_2$ (as well as of loadings and error variances) before any model modification. Second, each of the specified models with saturated structural correlations was imposed upon the sample data yielding residual covariances suggested by the multivariate Lagrange Multiplier (LM) test. This step is to mimic the measurement step in the two-step modeling process (Anderson & Gerbing, 1988). Next, the residual covariances suggested by the multivariate LM test with cumulative statistical significance $p<.05$ were added to the measurement model and the new structural estimates of $\phi_1$ and $\phi_2$ were obtained after the model modifications.

In this investigation, the multivariate LM test is restricted to suggesting residual covariances only, the purpose being to diagnose the impact of the most commonly applied measurement model modification on structural relation assessment. Three different measurement model respecifications were examined by the study separately: residual covariances within factors (intrafactor residual covariances) and residual covariances between factors (interfactor residual covariances) and the combination of both. That is, the multivariate LM test is restricted to suggesting only
intrafactor residual covariances in one respecification of each initial model, to suggesting only interfactor residual covariances in another respecification of each initial model, and to suggesting both intrafactor and interfactor residual covariances in yet another modification of each initial model. Thus, each initial solution was subjected to three separate modifications. As the purpose of this study is to diagnose the potential impacts of different types of the residual covariances on structural parameters, different modifications were made in isolation.

To summarize, in the 20 cells of the design, each replication’s data were analyzed seven times: (1) initially specified model without any modification, (2) each of the specified models with saturated structural correlations yielding intrafactor residual covariances, (3) specified model with saturated structural correlations yielding interfactor residual covariances, (4) specified models with saturated structural correlations yielding both intrafactor and interfactor residual covariances, (5) respecified model with suggested intrafactor residual covariances added, (6) respecified model with suggested interfactor residual covariances added, and (7) respecified model with both intrafactor and interfactor residual covariances added.

Convergent replications were generated for each condition in which a convergent replication reached convergence within 50 iterations in all analyses based on the same data set and did so without yielding any offending estimates (e.g., Heywood cases). This strategy was adopted because improper solutions in SEM might affect the estimation of structural parameters, thus threatening the results of the study. To maintain 1,400 replications per condition, an initial set of 1,500 replications was generated for each of the conditions.

Analyses

All analyses are based on the 1,400 convergent replications. For each replication on the same sample, the two structural parameters of interest were estimated (a) for the initial model before any modification, (b) after adding to the initial model any intrafactor residual covariances suggested by the multivariate LM test \( p < .05 \), (c) after adding to the initial model any interfactor residual covariances suggested by the multivariate LM test \( p < .05 \), and (d) after adding to the initial model any intrafactor and interfactor residual covariances suggested by the multivariate LM test \( p < .05 \). As a result, each given cell’s 1,400 replications had four estimates for each of the two structural parameters.

After initial examinations of general convergence, three primary analyses were conducted on the 1,400 sets of four structural parameter estimates. First, relative deviation values between parameter estimates before and after model modifications were computed for each cell, each of which indicates the bias of the parameter estimates from the true parameter values. Specifically, for each replication, the deviation was computed for each structural parameter estimate comparing the estimates after each of the three modifications back to the initial estimate. These values are treated as relative biases due to measurement model modifications and are equivalent to the difference in bias with respect to the true parameter value:

\[
\text{Relative bias (RB)} = \hat{\phi}_{\text{after}} - \hat{\phi}_{\text{before}}. \quad (1)
\]

Thus, for each parameter estimate within a replication, three such deviations were computed, comparing the parameter estimate after each of the three modifications back to the initial estimate. These values are treated as relative biases due to measurement model modifications and are equivalent to the difference in bias with respect to the true parameter value:

\[
(\hat{\phi}_{\text{after}} - \phi) - (\hat{\phi}_{\text{before}} - \phi).
\]

The difference between the estimated structural parameter \( \hat{\phi} \) and the true parameter value \( \phi \) is due to sampling fluctuation and any misspecification errors. Thus, the relative biases are the differences in structural estimates after counterbalancing the effects of sampling fluctuation and the existing misspecification errors for the initial model; in other words, they show the effects of adding residual covariances to structural estimates. A negative relative bias value indicates that adding residual covariances tend to decrease the structural estimates comparing to the initial structural estimates. A positive relative bias value, on the other hand, indicates that adding residual covariances increases the structural estimates comparing to the initial structural estimates.
For each structural parameter, three one-sample \( t \) tests were conducted comparing the average relative bias (across the 1,400 replications) to zero; note that these are equivalent to dependent sample \( t \) tests comparing biases before \((\hat{\phi}_{\text{before}} - \phi)\) and after \((\hat{\phi}_{\text{after}} - \phi)\) modification. Identical analyses were performed for each of the two structural parameters in all 20 cells. The first analysis focused on the relative bias of parameter estimates to detect potential differential bias as a function of parameter value.

Within each condition and for each specific parameter value the variance of the above bias values before and after modification was tested using a test for dependent sample variances (Hinkle, Wiersma & Jurs, 1988). Whereas the first analysis addressed the relative bias of parameter estimates, the second analysis attempted to examine parameter estimate variability. That is, because relative bias estimates can be positive or negative, their average could be near zero, and yet the experimental condition under scrutiny might actually (or additionally) be inducing instability in those estimates. Thus, to evaluate variability in the cells, the following procedures were conducted.

Recall that each cell has 1,400 sets of bias estimates, including those before modification, those after intrafactor residual covariances, those after interfactor residual covariances, and those after both intrafactor and interfactor residual covariances. Within each cell, the 1,400 bias estimates prior to modification were mean centered using their cell mean; this was repeated for the bias estimates intrafactor residual covariances, interfactor residual covariances, and both using their respective cell means. This within-cell centering removed bias and allowed specific focus on variability. Centered values were squared to eliminate sign and these squared values were used in the following analysis:

\[
\text{Squared mean centered bias} = (\text{Bias} - \overline{\text{Bias}})^2,
\]

where \( \text{Bias} = (\hat{\phi} - \phi) \), and \( \overline{\text{Bias}} \) is the average of bias for each cell before or after each model modification. Each cell now has 1,400 squared mean centered biases for the initial, intrafactor error covariance, interfactor residual covariance and the combined residual covariances conditions. Three dependent-sample \( t \) tests were conducted comparing squared mean centered biases before and after modification. Given the centering and squaring, this is tantamount to dependent-sample tests of variance of the bias estimates. Identical analyses were performed for both structural parameters in all 20 cells.

Eight series of dependent-sample \( t \) tests were conducted on the above relative bias estimates to detect the differences of the structural parameter estimate values among the five specified models. First, four dependent-sample \( t \) tests were conducted to compare the relative bias estimates in each cell for each of the misspecified model specifications (2, 3, 4 or 5) to those in the corresponding cell of the correctly specified model (specification 1). That is, for every cell for the misspecified models, four dependent-sample \( t \) tests were conducted for comparisons between specification 2 versus 1, specification 3 versus 1, specification 4 versus 1, and specification 5 versus 1.

These analyses are to detect potential different effects of adding different residual covariances on the structural parameter estimates for different model misspecification conditions. Similar procedures were also conducted to detect differences on the relative biases of the structural parameter estimate values between specifications 2 and 3, specifications 4 and 5, specifications 2 and 4, and specifications 3 and 5. The purpose is to assess if different measurement model modifications have different impacts with models under different types of misspecification conditions, and also under the same type of misspecification but different degrees of misspecification.

Results

Convergence

Only specification 5 had failures to converge. In 17 of the 20 cells, precisely 1,400 replication attempts were required to obtain
1,400 convergent replications. In the remaining three cells, additional replications were required. For the condition \( n = 100 \) after modification by adding interfactor residual covariances and the combination of both intrafactor as well as interfactor residual covariances, 1,402 replication attempts were required to yield 1,400. For the condition of \( n = 800 \) (after modification by adding intrafactor residual covariances), 1,401 replication attempts were required to yield 1,400. Thus, even under the most challenging conditions, the rate of converge was still excellent.

However, in order to ensure that the structural parameter estimates were generated based the same series of data sets across different respecifications and model specification conditions, the three data sets giving nonconvergent replications were eliminated for all the conditions. That is, 1,403 replications were conducted for all conditions yielding 1,400 convergent replications. All analyses were all conducted based on the 1,400 convergent replications.

Description of Modifications

Table 1 shows the average number of modifications of each type made under the different study conditions. Overall, the combination of intrafactor and interfactor residual covariances were added as expected, with interfactor residual covariances only slightly less frequent, followed by intrafactor residual covariances. Note that specification 1 is a correctly specified model and any modifications added are Type I errors due to random sampling fluctuation.

Sample size appeared to have a small but systematic influence for specifications 2 and 3. When the model was misspecified by omitting relevant cross-loading path(s), increases in sample size led to a slight increase in number of the residual covariance respecifications made as the power to identify the misspecification increased. In addition, specifications 2 and 3 with missing relevant paths provided a slightly larger number of residual covariance respecifications than other specifications, increasing by the severity of misspecification. Because paths are omitted to variables that serve as indicators to other factors in specifications 2 and 3, this explains the larger number of modifications for interfactor residual covariances, as well as other residual covariances. Conversely, specifications 4 and 5 with additional irrelevant paths had generally slightly smaller numbers of number of residual covariance respecifications than other specifications. Considering possible other paths explaining the covariances within the model, it is expected that less modifications are to be made.

Average Bias

Table 1 shows the results of average structural parameter estimates before and after any modifications for each condition. For specification 1, the true model is fit to the data. Thus the bias between the estimated structural parameters and true structural parameter is due to sampling error. All modifications suggested by the ML test under this situation are also due to sampling fluctuation, which makes the structural parameter estimates less accurate by slightly increasing the estimates. As sample size increased, the effects of modifications of adding residual covariances tended to decrease as the power increased. It was also noticed that all effects were minimal.

The greatest biases were observed both before and after any modifications for specifications 2 and 3 with missing relevant path(s) and especially for the under-estimated structural parameters. For specification 2, when the path from F1 to V4 is incorrectly fixed to zero, the first structural parameter estimates (\( \hat{\phi}_1 \)) were much more biased than the appropriately estimated structural parameter (\( \hat{\phi}_1 \)). The structural parameter estimates of \( \hat{\phi}_1 \) were inflated up to 0.07, while the structural parameter estimates of \( \hat{\phi}_2 \) were inflated to no more than 0.015.

The modifications of residual covariances seemed to help control the inflation for the structural parameter estimates of \( \hat{\phi}_1 \) to some degree, but with limited effects. For specification 3, when both paths from F1 to V4 and from F2 to V7 are incorrectly fixed to zero, both structural parameter estimates were biased.
and inflated to the similar degree and pattern. For specifications 4 and 5 with irrelevant path(s) - when truly zero path(s) are freely estimated - the structural parameter estimates were moderately inflated at a sample size of 100; as sample size increased, the inflation of the structural parameter estimates became very small.

**Relative Bias**

Further concerning the impact of different residual covariance modifications on relative bias of estimates for each of the two parameters, two one-sample  \( t \) tests were conducted for each of the 20 cells respectively comparing the average relative bias to zero. Given that two structural parameters appear in each cell, it may be considered that there are two \( t \) test pairs in each cell, with each of the two parameters being within a model for each modification type. In Table 2, results are presented for the average relative bias (magnified by 1,000) for each parameter across sample sizes, model specifications and different modification types.

Table 2 shows all relative biases to be no larger than hundredths. Most cells with statistically significant values occurred at specifications 2 and 3, when the model is misspecified with missing relevant paths. Very few significant cells occurred at specifications 1, 4 and 5. For the first structural parameter \( \phi_1 \) at specification 2 when the path from F1 to V4 is missing, all of the relative biases were negative, indicating a propensity for the modification to cause the parameter estimate to become slightly smaller than it is prior to modification.

For the second structural parameter \( \phi_2 \) at specification 2, the significant relative biases were positive when sample sizes are moderate or large, indicating a propensity for the modification to cause the parameter estimate to become slightly larger than it was prior to
modification. The results showed that the structural parameter estimates became smaller and closer to the true parameter value when modifications of residual covariances were made when the initial first structural path is under-represented in the unmodified misspecified model. Conversely, for the initial appropriately estimated second structural path, the effects of modifications were no longer helpful and tended to increase the estimated structural parameter and cause greater biases. However, the practical effects for the second structural parameter were much smaller than for the first under-estimated structural parameter. In addition, it is notable that most of the significant cases occurred when sample sizes were moderate or large.

More significant cells were observed as the severity of misspecification increased when one more relevant path from F2 to V7 is missing at specification 3. Relative biases associated with intrafactor residual covariance modification or the combination of intrafactor and interfactor were almost all negative across sample sizes, whereas in the interfactor case the propensity was for mostly positive relative bias. That is, intrafactor residual covariance or the combination of intrafactor and interfactor modification had a propensity to cause a parameter estimate to become smaller than it was prior to modification.

The interfactor covariance modification, by contrast, had the tendency to cause a

<table>
<thead>
<tr>
<th>Specification</th>
<th>Covariance Condition*</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100 200 400 800</td>
</tr>
<tr>
<td></td>
<td>( \hat{\phi}_1 )</td>
<td>( \hat{\phi}_2 )</td>
</tr>
<tr>
<td>1</td>
<td>BF</td>
<td>0.610 0.612 0.601 0.605 0.603 0.603 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Intra</td>
<td>0.612 0.613 0.603 0.605 0.604 0.603 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>0.610 0.613 0.601 0.607 0.604 0.602 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>0.612 0.614 0.602 0.607 0.604 0.602 0.603 0.600</td>
</tr>
<tr>
<td>2</td>
<td>BF</td>
<td>0.669 0.613 0.660 0.607 0.662 0.604 0.661 0.602</td>
</tr>
<tr>
<td></td>
<td>Intra</td>
<td>0.661 0.607 0.655 0.608 0.653 0.607 0.649 0.608</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>0.662 0.607 0.657 0.607 0.658 0.607 0.658 0.604</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>0.667 0.614 0.659 0.609 0.656 0.608 0.655 0.606</td>
</tr>
<tr>
<td>3</td>
<td>BF</td>
<td>0.668 0.665 0.660 0.657 0.662 0.655 0.661 0.653</td>
</tr>
<tr>
<td></td>
<td>Intra</td>
<td>0.668 0.647 0.656 0.627 0.655 0.615 0.652 0.608</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>0.669 0.667 0.662 0.661 0.662 0.659 0.660 0.659</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>0.668 0.650 0.658 0.631 0.656 0.619 0.653 0.611</td>
</tr>
<tr>
<td>4</td>
<td>BF</td>
<td>0.614 0.612 0.604 0.605 0.604 0.603 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Intra</td>
<td>0.615 0.613 0.606 0.605 0.602 0.603 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>0.615 0.612 0.605 0.607 0.604 0.602 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>0.617 0.613 0.607 0.607 0.603 0.603 0.602 0.600</td>
</tr>
<tr>
<td>5</td>
<td>BF</td>
<td>0.614 0.616 0.604 0.606 0.604 0.603 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Intra</td>
<td>0.616 0.615 0.605 0.605 0.603 0.603 0.601 0.600</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>0.618 0.612 0.604 0.607 0.603 0.602 0.602 0.600</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>0.618 0.616 0.605 0.608 0.602 0.603 0.601 0.600</td>
</tr>
</tbody>
</table>

*Note. BF refers to the initial model without any modification, Intra refers to the intrafactor residual covariance condition, Inter refers to the interfactor residual covariance condition, and Comb refers to the combination of intrafactor and interfactor residual covariance condition.
parameter estimate to become slightly greater than it was prior to modification. The results indicate that modification to intrafactor residual covariance or the combination of intrafactor and interfactor modification can help make more accurate structural parameter estimates when they are under-represented in the initial model. However, modifications of interfactor residual covariance tend to bias the structural parameter estimates even more under the same situation, although the effects are smaller.

Most of the cells for specifications 2 and 3 with moderate and large sample sizes were significant compared to specification 1, while there were fewer significant cells for specifications 4 and 5 compared to 1. In addition, while comparing the relative biases among specifications 2 to 4 and 3 to 5, many significant results were observed with moderate and large sample sizes. Results showed that misspecifications with missing path(s) had greater impact on the structural parameter estimates after model modifications than misspecifications with additional irrelevant path(s), and comparisons between specifications 2 and 3, and 4 and 5 showed the effects of the severity of the misspecification on the structural parameter estimates after model modifications.

The results have two implications. First, for misspecifications with additional irrelevant paths (specifications 4 and 5), adding one more irrelevant path does not impact the structural parameter estimates to a great degree; second, for misspecifications missing relevant path(s) (specifications 2 and 3), the biases on the structural parameter estimates after the intrafactor or the combined factor model modifications increase as the severity of the misspecification increases; the effects are somewhat ambiguous for interfactor model modification.

Relative Variability

Within each cell, the 1,400 bias estimates prior to modification were mean centered using their cell mean; this was repeated for the bias estimates from the intrafactor residual covariances, the interfactor residual covariances, and the combination of the two. All centered values were squared to eliminate sign; thus, the average value for a cell was the empirical parameter estimate variance.

In Table 3, the statistical significance of the difference between the original empirical parameter estimate variance and that after each modification is reported for each parameter estimate pair for intrafactor residual covariances, interfactor residual covariances, and the combination of both respectively. Also shown is the average percentage change for each pair of parameter estimates in empirical standard errors after modification, computed as the square root of the empirical parameter estimate variance before modification minus the square root of the empirical parameter estimate variance before modification, divided by the square root of the empirical parameter estimate variance before modification and then multiplied by 100 to create a percentage. That is,

$$
\text{% change} = \frac{\sqrt{s_{\text{after}}^2} - \sqrt{s_{\text{before}}^2}}{\sqrt{s_{\text{before}}^2}} = \frac{s_{\text{after}} - s_{\text{before}}}{s_{\text{before}}},
$$

where $s_{\text{after}}^2$ refers to the empirical parameter estimate variance after modification, $s_{\text{before}}^2$ refers to the empirical parameter estimate variance before modification, $s_{\text{after}}$ refers to the empirical parameter estimate standard deviation after modification, and $s_{\text{before}}$ refers to the empirical parameter estimate standard deviation before modification.

In general, model modification by adding residual covariances appeared to lead to a significant increase in variability of the parameter estimates relative to those prior to modification because more specification errors are introduced into the model. A couple of exceptions occurred in specification 2 with small sample sizes after modification of intrafactor residual covariance or interfactor error covariance, where modification significantly decreased variability of the parameter estimates relative to those prior to modification. Second, empirical standard errors were typically
inflated by less than 8% for specification 1 and varied for other specification conditions when model was misspecified. The most extreme values appeared under specification 2, with inflation reaching between -32.048% and 45.446%. For the other misspecified conditions, the empirical standard error values inflated to no larger than 18%. The modification of interfactor residual covariances generally yielded the smallest inflation in empirical standard errors for specifications 3, 4 and 5.

Conclusion
This study investigated the effects of measurement model modifications of adding residual covariances on structural parameter estimates with different sample sizes under different model specifications. The Monte Carlo study performed a systematic examination of the impact on the structural parameter estimates for three different common measurement model modifications of residual covariances under five different model specifications, which include adding adding intrafactor residual covariances, adding interfactor residual covariances and adding both intrafactor and interfactor residual covariances respectively. Overall, the model specifications with missing relevant path(s) have the most impact on the structural parameter estimates, while the impact increases as the severity of the misspecification increases.

The propensity is noted that the modifications of either adding intrafactor residual covariance or adding both intrafactor and interfactor residual covariances tended to decrease the structural parameter estimates.

Table 4: Percentage Change in Empirical Standard Error of Parameter Estimates across Modifications

<table>
<thead>
<tr>
<th>Specification</th>
<th>Covariance Condition*</th>
<th>Sample Size</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(\hat{\phi}_1)</td>
<td>(\hat{\phi}_2)</td>
<td>(\hat{\phi}_1)</td>
<td>(\hat{\phi}_2)</td>
</tr>
<tr>
<td>1</td>
<td>Intra</td>
<td>3.271(^a)</td>
<td>0.527</td>
<td>4.917(^a)</td>
<td>0.899(^a)</td>
<td>5.467(^a)</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>2.191(^a)</td>
<td>2.569(^a)</td>
<td>1.846(^a)</td>
<td>7.222(^a)</td>
<td>2.542(^a)</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>5.345(^a)</td>
<td>2.606(^a)</td>
<td>6.763(^a)</td>
<td>7.608(^a)</td>
<td>7.204(^a)</td>
</tr>
<tr>
<td>2</td>
<td>Intra</td>
<td>-29.844(^a)</td>
<td>-31.260(^a)</td>
<td>7.735(^a)</td>
<td>3.942(a)</td>
<td>10.685(^a)</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>-31.114(^a)</td>
<td>-32.048(^a)</td>
<td>-15.347(a)</td>
<td>-15.207(a)</td>
<td>7.250(a)</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>5.441(^a)</td>
<td>3.342(^a)</td>
<td>9.161(^a)</td>
<td>6.896(^a)</td>
<td>9.802(^a)</td>
</tr>
<tr>
<td>3</td>
<td>Intra</td>
<td>4.213(^a)</td>
<td>5.781(^a)</td>
<td>8.109(^a)</td>
<td>11.294(^a)</td>
<td>9.719(^a)</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>2.228(^a)</td>
<td>0.941</td>
<td>2.715(^a)</td>
<td>2.709(^a)</td>
<td>2.777(^a)</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>5.423(^a)</td>
<td>6.326(^a)</td>
<td>9.037(^a)</td>
<td>12.760(^a)</td>
<td>9.422(^a)</td>
</tr>
<tr>
<td>4</td>
<td>Intra</td>
<td>4.596(^a)</td>
<td>0.341</td>
<td>10.827(^a)</td>
<td>0.914(^a)</td>
<td>15.636(^a)</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>2.174(^a)</td>
<td>2.703(^a)</td>
<td>2.359(^a)</td>
<td>7.280(^a)</td>
<td>1.995(^a)</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>6.239(^a)</td>
<td>2.610(^a)</td>
<td>10.134(^a)</td>
<td>7.721(^a)</td>
<td>13.483(^a)</td>
</tr>
<tr>
<td>5</td>
<td>Intra</td>
<td>3.969(^a)</td>
<td>-1.289</td>
<td>11.371(^a)</td>
<td>0.357</td>
<td>15.096(^a)</td>
</tr>
<tr>
<td></td>
<td>Inter</td>
<td>4.430(^a)</td>
<td>-5.187(^a)</td>
<td>4.168(^a)</td>
<td>3.043(^a)</td>
<td>5.925(^a)</td>
</tr>
<tr>
<td></td>
<td>Combine</td>
<td>7.939(^a)</td>
<td>-2.998(^a)</td>
<td>12.647(^a)</td>
<td>4.016(^a)</td>
<td>17.311(^a)</td>
</tr>
</tbody>
</table>

*Note. Intra refers to the intrafactor residual covariance condition, inter refers to the interfactor residual covariance condition, and comb refers to the combination of intrafactor and interfactor residual covariance condition.\n
The empirical parameter estimate variance was statistically significantly different (p<.05) from the corresponding variance prior to model modification.
compared to those prior to modification, while the interfactor modification tended to increase the structural parameter estimates. By contrast, the model specifications with additional irrelevant path(s) did not have much impact on the structural parameter estimates because very few random significant cells were observed at specifications 4 and 5. In sum, only the under-estimated structural parameter estimates from the misspecified models with missing relevant path(s) approached the true values when the intrafactor residual covariances or the combination of intrafactor and interfactor residual covariances were added. In addition, all three model modifications tended to significantly increase the variability of the parameter estimates relative to those prior to modification across model specification conditions. Interestingly, sample size did not appear to influence the impact of the measurement model modifications of residual covariances on the structural parameter estimates. Thus, more attention should be paid to a misspecified model with missing path(s) when conducting measurement model modification.

Although many significant cells were observed, the impacts of the different types of model modification were found to be small, usually no larger than hundredths. However, it is notable that this study focused on conditions with normally distributed data. It is expected that the impact of measurement model modification would be enlarged under the inevitable conditions with nonnormally distributed data in practical research. Based on results from this investigation it may be stated that liberally saturating measurement models with additional residual covariance parameters does not necessarily help with the structural parameter estimates if a researcher begin with a misspecified model, especially with missing parameters. Thus, the role of theory cannot be more salient for all model types (Hancock, 1999).

However, researchers should exercise caution with the results. As with any simulation study, there are innumerable conditions to manipulate and choices must be made to keep the design manageable. It is possible that another research study could produce different results under different models and experimental conditions. For example, the extent to which measurement model modification alters structural parameter estimates under conditions of nonnormality remains to be studied. Similarly, the amount of modification that affects parameters within a mean structure, such as latent means and intercepts, is also a subject for further investigation.

References


On A Comparison between Two Measures of Spatial Association

Faisal G. Khamis  
Al-Zaytoonah University of Jordan, 
Amman, Jordan

Abdul Aziz Jemain  
University Kebangsaan Malaysia, 
Selangor, Malaysia

Kamarulzaman Ibrahim

Two measures of spatial association between two variables were used by many researchers. These are the Wartenberg (1985) and Lee (2001) measures. Based on simulation for lattice data, the sensitivity of both measures was studied and compared with different choices of spatial structures, spatial weights and sample sizes using bias and mean square error. Different scenarios are used in terms of assumed numbers and sample sizes. Moran’s $I$ is used to examine the spatial autocorrelation of such a variable with itself. Both the Wartenberg and Lee measures are found to be sensitive, however, Wartenberg’s measure is found to be somewhat better than Lee’s measure because it is slightly more sensitive when sample size is small.

Key words: Wartenberg and Lee measures, simulation study, spatial association, sensitivity, spatial structures, weights.

Introduction

It is argued that lattice data are spatially correlated. The Wartenberg (1985) and Lee (2001) are two measures used for investigating the spatial association between two or more variables taking into account neighboring information. Lee criticized Wartenberg’s measure and suggested two criteria for developing a measure for bivariate spatial association. First, the measure should conform to Pearson’s $r$ between two variables in terms of direction and magnitude. Second, a bivariate spatial association measure should reflect the degrees of spatial autocorrelation for both variables under investigation. Lee developed an index, $L$, that combines Pearson’s bivariate correlation with Moran’s spatial autocorrelation measures, to measure spatial association. Lee stated that Wartenberg’s measure is vulnerable to a reverse of the direction of spatial association. Also, Lee’s measure has the spatial lags of two variables while Wartenberg’s measure has the spatial lag for one variable. Thus, this study makes a comparison between these measures in terms of their sensitivity. The observations for each particular sub-area can be either univariate or multivariate data. When the data are univariate, Moran’s $I$ statistic can be used to describe the spatial autocorrelation of such a variable. If the observations are multivariate then the Wartenberg and Lee measures can be used.

Methodology

Real data are important for the development of statistical methods and ideally their analysis also stimulates research in statistical theory. Simulated data is also important and has a different role. This role is particularly valuable when several competing methods are available but little or no theory exists to indicate which is superior. Simulating spatial data is important because statistical inference for spatial data often relies on randomization tests. The ability to simulate realization of a hypothesized process quickly and efficiently is important to allow a sufficient number of realizations to be produced (Schabenberger & Gotway, 2005). The
The performance of the Wartenberg and Lee measures is evaluated based on simulated data. The spatial association measures of Wartenberg and Lee can be given respectively as

\[ M_{X,Y} = \frac{N}{S_0} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - \bar{x})(y_j - \bar{y}) \]

\[ L_{X,Y} = \frac{N}{S_0} \sum_{i=1}^{N} \left[ \sum_{j=1}^{N} w_{ij} (x_j - \bar{x}) \times \left( \sum_{j=1}^{N} w_{ij} (y_j - \bar{y}) \right) \right] \]

where \( S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}, i \neq j \), \( N \) is the sample size and \( w_{ij} \) is the binary spatial weight (1, 0).

The univariate statistic for spatial association or autocorrelation of Moran’s \( I \) is defined as (Cliff & Ord, 1981)

\[ I = \frac{N}{S_0} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - \bar{x})(x_j - \bar{x}) \]

Because the neighbor structure is the basic structure for the covariance model of lattice data, a careful definition of spatial neighbors is a crucial analysis step (Kaluzny, et al., 1998). Neighbors may be defined as locations which border each other or as locations within a certain distance of each other. If neighbors are defined as locations bordering each other, then there are several types of spatial neighbors. For example, the first order method (the rook pattern) identifies neighbors as those to left, to the right, or above or below each location, that is, the rook makes links in four cardinal directions. The diagonal method (the bishop pattern) makes only diagonal links. The second order method (the queen pattern) includes the first-order neighbors and those diagonally linked, that is, the queen makes links in all eight directions. Figure 1 shows these three types of spatial connectivity.

The sensitivity to the choice or the definition of spatial structures of neighbors was studied for both Wartenberg and Lee measures. The simulation study was based on six spatial structures: sharing boundary (rook), sharing boundary (bishop), sharing boundary (queen), distance apart (1.5), distance apart (2.25) and distance apart (3).

If the spatial structure was made based on distance apart, the distances between location \( i \) and all its surrounding neighbors will be calculated. These distances were calculated in the SPLUS program based on such distance measures, for example, Euclidian. If the calculated distances were found within for example, distance apart (1.5), the surrounded locations will be considered as neighbors to the location \( i \).

Kaluzny, et al., (1998) stated the choice of spatial weights between such \( ith \) location and its neighbors is a crucial step. They recommended that several choices of spatial weights be tried so that the sensitivity of the results can be determined. However, three different spatial weights \( w_{ij} \) were used (\( w_{ij} = 1, w_{ij} = 1/d_{ij}, w_{ij} = 1/d_{ij}^2 \)), where \( d_{ij} \) is the distance between location \( i \) and location \( j \) and when \( d_{ij} \) is large, the \( w_{ij} \) will be less. This means that \( w_{ij} \) for the nearest neighbors will be higher than that for the farthest neighbors.

The bias and mean square error (MSE) were used to decide which statistic is better. Let \( \theta \) be the parameter of interest, then the MSE of \( \hat{\theta} \) is defined as follows (Garthwaitw, et al., 1995)

\[ MSE(\hat{\theta}) = E \left[ (\hat{\theta} - \theta)^2 \right] \]

and the estimated value is calculated using

\[ \hat{MSE}(\hat{\theta}) = \hat{Var}(\hat{\theta}) + [\hat{bias}(\hat{\theta})]^2 \]

where
and \( \hat{\theta} \) is the estimated bivariate spatial association measure based on simulated data, \( \theta \) is the actual value of bivariate spatial association measure based on Wartenberg’s or Lee’s measure and \( s \) is the total number of runs (in this study, \( s = 10,000 \)).

The estimate \( \hat{\theta} \) is an unbiased estimator for \( \theta \) if \( E(\hat{\theta}) = \theta \); otherwise it is biased. The bias of \( \hat{\theta} \) is defined to be (Garthwaitw, et al., 1995)

\[
\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta
\]

and the estimated value of \( E(\hat{\theta}) \) is given by

\[
\bar{\theta} = \frac{\sum_{i=1}^{s} \hat{\theta}_i}{s},
\]

then the estimated value of \( \text{bias}(\hat{\theta}) \) is given by

\[
\text{bias}(\hat{\theta}) = \bar{\theta} - \theta.
\]

Simulation Process

The process of the simulation study included several steps which were considered somewhat complicated. The complication arose from allowing three kinds of spatial correlations before starting the spatial analysis and because the simulation must be made under a randomness assumption. The spatial correlations were: the bivariate spatial correlation between two variables and spatial autocorrelation for each variable.

The simulation study was carried out using SPLUS programming and accomplished in four steps. First, the original samples for two variables were designated to act as the population for sampling purposes. In the second step, the original samples were re-sampled a specified number of times (up to several thousands) to generate a large number of new samples, where each sample was a random subset of the original sample. In the third step, the bivariate spatial measures of Wartenberg and Lee were estimated for each new sample. In the last step, the estimated values of bias and MSE were calculated using the computed spatial measures found in step 3.

During the process of generating new samples, the simulation program may change certain characteristics of the sample to meet the researcher’s objectives. For example, the degree of correlation between variables may be varied across the generated samples in some systematic manner. The simulation process was run using 10000 runs, where Wartenberg’s measure, \( W \) and Lee’s measure \( L \), were each estimated 5,000

\[
\text{Var}(\hat{\theta}) = \frac{\sum_{i=1}^{s} (\hat{\theta}_i - \theta)^2}{s},
\]
times. The bias and MSE were then calculated for each measure. The mechanism proposed herein contains certain assumed form of univariate spatial correlation (autocorrelation) for each variable as shown from the distribution of assumed observed values, and hence there is also a bivariate spatial correlation between these two autocorrelated variables based on the actual value of Wartenberg and Lee measures.

Results
To assess the performance of both the Wartenberg and Lee measures, a series of simulations were conducted. Several scenarios were studied to investigate the sensitivity of both the Wartenberg and Lee measures based on different choices of spatial structures and spatial weights using bias and MSE. The values of two variables, X and Y, were generated based on their assumed true means and standard deviation one for each observation. Autocorrelation values of each X and Y variables based on Moran’s I were found positive, negative, high or low because different choices of spatial structures and spatial weights were used; the nine resulting scenarios follow.

Scenario 1: Sample Size (4×4 = 16)
Figures 3a and 3b show two study areas with their assumed true means. Table 1, shows the autocorrelation values for both variables X and Y based on global Moran’s I statistic. Table 2 shows the actual values of the Wartenberg and Lee measures and their bias and MSE using different choices of spatial structures.

Figure 3 Proposed Area Divided into 16 Quadrates in a 4×4 Lattice

The numbers in the quadrates are the assumed true means for two different variables X and Y, where is a gradient patch from low values (in the center) to high values (in the edges) for X variable, and (b) is the opposite direction of (a) for Y variable.

Table 1: Autocorrelation for Variables X and Y Based on Moran’s I Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.31</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>-0.24</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.07</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.07</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>-0.18</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>-0.19</td>
</tr>
</tbody>
</table>
Table 2: The *bias* and *MSE* of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>W Actual</th>
<th>W bias</th>
<th>W MSE</th>
<th>L Actual</th>
<th>L bias</th>
<th>L MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>-0.31</td>
<td>0.14</td>
<td>0.03</td>
<td>-0.17</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>0.24</td>
<td>-0.11</td>
<td>0.02</td>
<td>-0.86</td>
<td>0.38</td>
<td>0.16</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>-0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>-0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>0.18</td>
<td>-0.08</td>
<td>0.01</td>
<td>-0.10</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>0.19</td>
<td>-0.08</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.05</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Scenario 2: Sample Size (4×4 = 16)

Figures 4a and 4b show two study areas with their assumed true means. Table 3 shows the autocorrelation values for both variables X and Y based on Moran’s I statistic. Table 4 shows the actual values of Wartenberg and Lee measures and their *bias* and *MSE* using different choices of spatial structures.

Figure 4: Proposed Area Divided into 16 Quadrates in a 4×4 Lattice

The numbers in the quadrates are the assumed true means for two different variables X and Y, where both (a) and (b) are a gradient patches from low values (in the center) to high values (in the edges) for X and Y variables respectively.

Table 3: Autocorrelation for Variables X and Y Based on Moran’s I Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>X Autocorrelation</th>
<th>Y Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>-0.24</td>
<td>-0.24</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>-0.18</td>
<td>-0.18</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>-0.19</td>
<td>-0.19</td>
</tr>
</tbody>
</table>
Table 4: The bias and MSE of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>W</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual</td>
<td>bias</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.31</td>
<td>-0.24</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>-0.24</td>
<td>0.18</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.07</td>
<td>-0.06</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.07</td>
<td>-0.06</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>-0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>-0.19</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Scenario 3: Sample Size (4×4 = 16)

Figures 5a and 5b show two study areas with their assumed true means. Table 5 shows the autocorrelation values for both variables X and Y based on Moran’s I statistic. Table 6 shows the actual values of Wartenberg and Lee measures and their bias and MSE using different choices of spatial structures.

Table 5: Autocorrelation for Variables X and Y Based on Moran’s I Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.57</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>0.22</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.42</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.42</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>0.14</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

The numbers in the quadrates are the assumed true means for two different variables X and Y, where (a) is a gradient patch from low values (the bottom left side) to high values (the upper right side) for X variable, and (b) is the opposite direction of (a) for Y variable.
Table 6: The *bias* and *MSE* of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>W</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual</td>
<td>bias</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>-0.58</td>
<td>0.30</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>-0.28</td>
<td>0.14</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>-0.45</td>
<td>0.23</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>-0.45</td>
<td>0.23</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>-0.14</td>
<td>0.07</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>-0.09</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Scenario 4: Sample Size (4x4 = 16)

Figures 6a and 6b show two study areas with their assumed true means. Table 7 shows the autocorrelation values for both variables X and Y based on Moran’s I statistic. Table 8 shows the actual values of Wartenberg and Lee measures and their *bias* and *MSE* using different choices of spatial structures.

Table 7: Autocorrelation for Variables X and Y Based on Moran’s I Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.59</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>0.28</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.46</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.46</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>0.14</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

The numbers in the quadrates are the assumed true means for two different variables X and Y, where both (a) and (b) are a gradient patches from low values (the bottom left side) to higher values (the upper right side) for X and Y variables respectively.
Table 8: The \textit{bias} and \textit{MSE} of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>W</th>
<th></th>
<th>L</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual bias</td>
<td>MSE</td>
<td>Actual bias</td>
<td>MSE</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.58</td>
<td>-0.31</td>
<td>0.11</td>
<td>0.48</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>0.25</td>
<td>-0.14</td>
<td>0.03</td>
<td>0.27</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.44</td>
<td>-0.24</td>
<td>0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.44</td>
<td>-0.24</td>
<td>0.06</td>
<td>0.34</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>0.14</td>
<td>-0.08</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>0.09</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Scenario 5: Sample Size (6×6 = 36)

Figures 7a and 7b show two study areas with their assumed true means. Table 9 shows the autocorrelation values for both variables \(X\) and \(Y\) based on Moran’s \(I\) statistic. Table 10 shows the actual values of Wartenberg and Lee measures and their \textit{bias} and \textit{MSE} using different choices of spatial structures.

Figure 7: Proposed Area Divided into 36 Quadrates in a 6×6 Lattice

The numbers in the quadrates are the assumed true means for two different variables \(X\) and \(Y\), where (a) is a gradient patch from low values (in the center) to high values (in the edges) for \(X\) variable, and (b) is the opposite direction of (a) for \(Y\) variable.

Table 9: Autocorrelation for Variables \(X\) and \(Y\) Based on Moran’s \(I\) Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>(X) Autocorrelation</th>
<th>(Y) Autocorrelation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>0.33</td>
<td>0.35</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>0.10</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Table 10: The *bias* and *MSE* of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>W</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual</td>
<td>bias</td>
</tr>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>-0.63</td>
<td>0.23</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>-0.35</td>
<td>0.12</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>-0.50</td>
<td>0.18</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>-0.50</td>
<td>0.18</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>-0.18</td>
<td>0.06</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>-0.10</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Scenario 6: Sample Size *(8×8 = 64)*

Figures 8a and 8b show two study areas with their assumed true means. Table 11 shows the autocorrelation values for both variables *X* and *Y* based on Moran’s *I* statistic. Table 12 shows the actual values of Wartenberg and Lee measures and their *bias* and *MSE* using different choices of spatial structures.

Figure 8: Proposed Area Divided into 64 Quadrates in a 8×8 Lattice

The numbers in the quadrates are the assumed true means for two different variables *X* and *Y*, where (a) is a gradient patch from low values (in the center) to high values (in the edges) for *X* variable, and (b) is the opposite direction of (a) for *Y* variable.
Table 11: Autocorrelation for Variables $X$ and $Y$ Based on Moran’s $I$ Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>Autocorrelation $X$</th>
<th>Autocorrelation $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>0.48</td>
<td>0.54</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>0.61</td>
<td>0.66</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>0.42</td>
<td>0.44</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 12: The bias and $MSE$ of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Structures

<table>
<thead>
<tr>
<th>Type of Spatial Structure</th>
<th>W Actual</th>
<th>W bias</th>
<th>W MSE</th>
<th>L Actual</th>
<th>L bias</th>
<th>L MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharing Boundary (Rook)</td>
<td>-0.75</td>
<td>0.15</td>
<td>0.02</td>
<td>-0.59</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>Sharing Boundary (Bishop)</td>
<td>-0.52</td>
<td>0.10</td>
<td>0.01</td>
<td>-0.31</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Sharing Boundary (Queen)</td>
<td>-0.64</td>
<td>0.13</td>
<td>0.02</td>
<td>-0.41</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>Distance Apart (1.5)</td>
<td>-0.64</td>
<td>0.13</td>
<td>0.02</td>
<td>-0.41</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>Distance Apart (2.25)</td>
<td>-0.43</td>
<td>0.09</td>
<td>0.01</td>
<td>-0.19</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Distance Apart (3)</td>
<td>-0.36</td>
<td>0.07</td>
<td>0.01</td>
<td>-0.14</td>
<td>0.03</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Scenario 7: Sample Size (4 × 4 = 16)

Figures 9a and 9b show two study areas with their assumed true means. The spatial structure is defined using the distance apart (1.5). Table 13 shows the autocorrelation values for both variables $X$ and $Y$ based on Moran’s $I$ statistic. Table 14 shows the actual values of Wartenberg and Lee measures and their bias and $MSE$ using different choices of spatial weights.

Figure 9: Proposed Area Divided into 16 Quadrates in a 4 × 4 Lattice

The numbers in the quadrates are the assumed true means for two different variables $X$ and $Y$, where (a) is a gradient patch from low values (in the center) to high values (in the edges) for $X$ variable, and (b) is the opposite direction of (a) for $Y$ variable.
Table 13: Autocorrelation for Variables $X$ and $Y$ Based on Moran’s $I$ Using Different Types of Spatial Weights

<table>
<thead>
<tr>
<th>Type of Spatial Weight</th>
<th>Autocorrelation $X$</th>
<th>Autocorrelation $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{ij} = 1$</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}$</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}^2$</td>
<td>0.16</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 14: The bias and MSE of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Weights

<table>
<thead>
<tr>
<th>Type of Spatial Weight</th>
<th>W Actual</th>
<th>W bias</th>
<th>W MSE</th>
<th>L Actual</th>
<th>L bias</th>
<th>L MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{ij} = 1$</td>
<td>-0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}$</td>
<td>-0.12</td>
<td>0.05</td>
<td>0.01</td>
<td>-0.06</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}^2$</td>
<td>-0.16</td>
<td>0.07</td>
<td>0.01</td>
<td>-0.05</td>
<td>0.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Scenario 8: Sample Size ($6 \times 6 = 36$)

Figures 10a and 10b show two study areas with their assumed true means. The spatial structure is defined using the distance apart (1.5). Table 15 shows the autocorrelation values for both variables $X$ and $Y$ based on Moran’s $I$ statistic. Table 16 shows the actual values of Wartenberg and Lee measures and their bias and MSE using different choices of spatial weights.

Figure 10: Proposed Area Divided into 36 Quadrates in a $6 \times 6$ Lattice

The numbers in the quadrates are the assumed true means for two different variables $X$ and $Y$, where (a) is a gradient patch from low values (in the center) to high values (in the edges) for $X$ variable, and (b) is the opposite direction of (a) for $Y$ variable.
Table 15: Autocorrelation for Variables $X$ and $Y$ Based on Moran’s $I$ Using Different Spatial Weights

<table>
<thead>
<tr>
<th>Type of Spatial Weight</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_{ij} = 1$</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}$</td>
<td>0.52</td>
<td>0.53</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}^2$</td>
<td>0.54</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Table 16: The bias and MSE of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Weights

<table>
<thead>
<tr>
<th>Type of Spatial Weight</th>
<th>$W$</th>
<th>$L$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual</td>
<td>bias</td>
</tr>
<tr>
<td>$w_{ij} = 1$</td>
<td>-0.50</td>
<td>0.18</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}$</td>
<td>-0.53</td>
<td>0.19</td>
</tr>
<tr>
<td>$w_{ij} = 1/d_{ij}^2$</td>
<td>-0.55</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Scenario 9: Sample Size (8×8 = 64)

Figures 11a and 11b show two study areas with their assumed true means. The spatial structure is defined using the distance apart (1.5). Table 17 shows the autocorrelation values for both variables $X$ and $Y$ based on Moran’s $I$ statistic. Table 18 shows the actual values of Wartenberg and Lee measures and their bias and MSE using different choices of spatial weights.

Figure 11: Proposed Area Divided into 64 Quadrates in a 8×8 Lattice

The numbers in the quadrates are the assumed true means for two different variables $X$ and $Y$, where (a) is a gradient patch from low values (in the center) to high values (in the edges) for $X$ variable, and (b) is the opposite direction of (a) for $Y$ variable.
Conclusion
The results from these scenarios show that the Wartenberg and Lee measures differ slightly in terms of their sensitivity to different choices of spatial structures and spatial weights. Results show that Wartenberg’s measure is somewhat more sensitive than Lee’s measure to the different choices of spatial structures and spatial weights when the sample size is small; for the large sample sizes the results of both measures are approximately the same. Several techniques in statistics are sensitive - meaning they sometimes provide inaccurate results when a small sample size is used - because the information in a small sample is less than that of a large sample size.

Wartenberg’s equation is vulnerable to a reverse in direction of association as stated by Lee. This reverse in direction was found in scenarios 1 and 2 as shown in the column of actual value of Wartenberg’s measure in Tables 2 and 4 respectively.

Table 17: Autocorrelation for Variables X and Y Based on Moran’s I Using Different Spatial Weights

<table>
<thead>
<tr>
<th>Type of Spatial Weight</th>
<th>Autocorrelation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>( w_{ij} = 1 )</td>
<td>0.61</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>( w_{ij} = 1/d_{ij} )</td>
<td>0.63</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>( w_{ij} = 1/d_{ij}^2 )</td>
<td>0.65</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>

Table 18: The bias and MSE of Wartenberg’s Measure (W) and Lee’s Measure (L) with Actual Values Using Different Types of Spatial Weights

<table>
<thead>
<tr>
<th>Type of Spatial Weight</th>
<th>W Actual</th>
<th>bias</th>
<th>MSE</th>
<th>L Actual</th>
<th>bias</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( w_{ij} = 1 )</td>
<td>-0.64</td>
<td>0.13</td>
<td>0.02</td>
<td>-0.41</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>( w_{ij} = 1/d_{ij} )</td>
<td>-0.66</td>
<td>0.13</td>
<td>0.02</td>
<td>-0.33</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>( w_{ij} = 1/d_{ij}^2 )</td>
<td>-0.68</td>
<td>0.13</td>
<td>0.02</td>
<td>-0.27</td>
<td>0.05</td>
<td>0.00</td>
</tr>
</tbody>
</table>

References
An Evaluation of Multiple Imputation for Meta-Analytic Structural Equation Modeling

Carolyn F. Furlow
Georgia State University

S. Natasha Beretvas
University of Texas at Austin

A simulation study was used to evaluate multiple imputation (MI) to handle MCAR correlations in the first step of meta-analytic structural equation modeling: the synthesis of the correlation matrix and the test of homogeneity. No substantial parameter bias resulted from using MI. Although some SE bias was found for meta-analyses involving smaller numbers of studies, the homogeneity test was never rejected when using MI.

Key words: Meta-analysis, structural equation modeling, multiple imputation, missing data.

Introduction

Meta-analytic structural equation modeling (MASEM) has been recommended as a useful approach for supporting theoretical models and combines the benefits of both meta-analysis and structural equation modeling (SEM). The meta-analytic benefits include the use of quantitative synthesis methods which allow a researcher to cull correlations from multiple studies that can then be combined across those studies to provide individual, more precise estimates of each relevant correlation. This can be conducted for each element of a correlation matrix that describes the full set of relationships between the variables of interest to the MASEM researcher. The resulting meta-analytically pooled correlation matrix can then be analyzed using SEM procedures.

Several methodological dilemmas and impediments are frequently encountered by MASEM researchers. Most commonly, applied researchers synthesize elements of a correlation matrix which is then analyzed using SEM software. The MASEM researcher is likely to encounter problems with missing data. In the MASEM context, this can be in the form of missing studies or missing correlations (Pigott, 1994). The focus of this study concerns performance of multiple imputation for handling missing correlations for the first step of MASEM, the synthesis of the correlation matrices across studies.

Missing Data in MASEM

If a researcher were interested in summarizing elements of a correlation matrix describing relationships among five variables, ideally data from each contributing study would include estimates of each of the correlations in the matrix. This is rarely the case. At the primary study level, several possible reasons exist to explain why a correlation might not be reported. The authors of the study might not have been interested in measuring one of the five variables of interest to the meta-analyst, or at the time when one of the primary studies was conducted, a variable of interest to the meta-analyst might not yet have been conceptualized as a construct that exhibits an interesting relationship with other variables in the matrix (Furlow & Beretvas, 2005). Thus, in either scenario, the study would not include correlations of that variable with each of the remaining four.

Carolyn F. Furlow conducted this study while in the Department of Educational Policy Studies. Current email: cfurlow@cdc.gov. Natasha Beretvas is an Associate Professor and chair of Quantitative Methods in the Department of Educational Psychology. Her interests are in multilevel and meta-analytic modeling techniques. Email her at: tasha.beretvas@mail.utexas.edu.
Another plausible reason why a correlation might be missing from a study may be the file drawer problem (Rosenthal, 1979), most commonly referred to as publication bias. Publication bias describes the tendency of authors (and editors) to provide statistical results (either descriptive or inferential) only for statistically significant results. Authors either fail to mention uninteresting (commonly meaning statistically non-significant) results, or journal space limitations restrict the presentation of the relevant values offering only the phrase “not statistically significant”.

Authors using MASEM reported using a variety of methods for handling missing correlation estimates. Hom, et al. (1992) utilized listwise deletion (LD) by only incorporating results from studies that provided the full set of correlations of interest. The vast majority of MASEM researchers (e.g., Brown & Stayman, 1992; Conway, 1999; Manfredo, et al., 1996; Parker, et al., 2003; Premack & Hunter, 1988; Tett & Meyer, 1993) used pairwise deletion (PD). A few used single value imputation to handle missing correlations in their MASEM studies (Bailey, 2001; Colquitt, LePine & Noe, 2000). The single imputation method involves either mean imputation (using the mean of the correlation estimates provided in other studies in the meta-analysis) or substituting a value based on related results from other meta-analytic research conducted outside the domain of the focal MASEM study (Colquitt, et al., 2000). It is unclear in some MASEM articles how the missing correlations were handled (e.g., Verhaeghen & Salthouse, 1997). To date, no applied study has used multiple imputation (MI) to handle missing correlations.

As with any statistical analysis, the source of the missing data impacts how well the method used to handle the missing data will function. Little and Rubin (1987) categorized missing data mechanisms into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). What distinguishes these missing data mechanisms is the relationship between the missing (unreported or unobserved) data point and the complete set of data. In traditional statistical analyses the cases (rows) in a dataset are a single study’s participants and the data points are participants’ scores on each variable (columns). In meta-analysis, the columns (variables) consist of correlation estimates for each row, which represents each study.

With MCAR data, the correlation’s missingness is unrelated to any of the observed correlations in the dataset. As an example, when data are not gathered in a primary study due to lack of funding and that funding is not related to any of the variables and thus to correlations in the dataset, then the missingness can be considered MCAR (Pigott, 1994). Another example of MCAR data occurs when a primary researcher does not measure a variable of interest for the MASEM because it is not theoretically relevant to his/her study and thus that variable’s correlations with the other variables would be missing.

With MAR data, a correlation’s missingness ($M_{ij} = 1$ if correlation $r_{ij}$ is missing, and $M_{ij} = 0$, otherwise) would be related to the correlation’s value but only indirectly - specifically only through another observed correlation (or correlations) in the dataset (Collins, Schafer & Kam, 2001). For example, the likelihood of a correlation, $r_{WX}$, being missing in a study might be greater for higher values of another observed correlation, $r_{YZ}$. There would then appear to be a simple relationship between $r_{WX}$ and $M_{WX}$. However, if within levels of $r_{YZ}$, values of $r_{WX}$ are unrelated to $M_{WX}$, then the data are MAR. In other words, if the relationship between $r_{WX}$ and its likelihood of being missing (i.e., $M_{WX}$) is fully explained by the relationship between $r_{YZ}$ and $M_{WX}$, then the missingness can be considered MAR.

As a meta-analytic example, studies being synthesized might involve an assessment of the relationship between constructs $W, X, Y$ and $Z$. There might be a variety of scales that are designed to assess each of $Y$ and $Z$. Researchers who use certain (more reliable psychometrically) measures of $Y$ and $Z$ might espouse a theoretical framework that also means they are more likely to be interested in the relationship between measures $X$ and $W$. Researchers who use
different measures of $Y$ and $Z$ may be less likely to assess $X$ and $W$. When using more reliable measures of $Y$ and $Z$, the resulting $r_{YZ}$ will tend to be stronger than the $r_{YZ}$ based on less reliable scores. Thus, for higher values of $r_{YZ}$, the likelihood that $r_{XW}$ is reported is higher than for lower values of $r_{YZ}$. And for lower values of $r_{YZ}$, it is more likely that $r_{XW}$ will be missing. However, controlling for $r_{YZ}$, there is no relationship between $r_{XW}$ and the likelihood that $r_{XW}$’s value is missing from a study. Thus missing $r_{XW}$ values could be considered MAR.

MNAR data result when the likelihood of a missing correlation is related to the value of the (missing) correlation itself. Publication bias provides a likely cause of MNAR data. As mentioned earlier, if a correlation estimate is not statistically significant, an author might not report the relevant statistical information and/or an editor might censor the presentation of such results. If publication bias explains the missingness, then the likelihood of missingness is negatively related to the correlation estimate’s value, all other factors being equal. The opposite pattern of MNAR (in which there is a positive relationship between the missing correlation’s value and the likelihood of its being missing) is also possible. It can occur when a researcher purposely neglects to report a correlation that is stronger than would be expected theoretically.

Use of listwise deletion (LD) to handle missing data can be advocated in situations in which only a few data points are missing. LD has been found to result in unbiased parameter estimates for models estimated with MCAR data (Allison, 2003). However, LD can also result in a drastic reduction in statistical power under conditions with high proportions of missing data. Graham and Hofer (2000) recommend that if only five percent or less of the dataset is MCAR, then LD can be used. Unfortunately, LD is usually not a feasible alternative in MASEM research. A high proportion of study authors do not report all correlations of interest to MASEM researchers (Furlow & Beretvas, 2005), for example, in Premack and Hunter’s (1988) MASEM study, if LD had been used it would have resulted in a completely empty dataset.

Many MASEM researchers do not use LD but instead employ available case analysis (PD) as the preferred method for handling missing correlations (Furlow & Beretvas, 2005). When using PD, no information is deleted; each element of the correlation matrix is instead obtained by synthesizing all available, observed correlation estimates. Use of PD with conventional (i.e., not meta-analytic) data has been found to result in approximately unbiased parameter estimates for MCAR data, however, PD can lead to biased estimates if data are MAR or MNAR (Graham & Hofer, 2000). Use of PD has also been found to lead to non-positive definite correlation matrices for typical, non-meta-analytic datasets (Arbuckle, 1996; Graham & Hofer, 2000). To date, this problem has been reported in only one applied MASEM study (Kubeck, 2002). Even the few MASEM simulation studies that have been conducted to evaluate the performance of PD with missing data have not encountered non-positive definite matrices (S. F. Cheung, 2000; M. Cheung & Chan, 2005; Furlow & Beretvas, 2005).

Another method to handle missing data in MASEM research could be through the use of mean imputation to impute a missing data point’s value (Graham & Hofer, 2000). The problem with mean imputation is that it deflates the associated variability of the relevant estimate (the correlation in MASEM); this holds even when the missing data mechanism is MCAR, thus mean imputation is not recommended. To
compensate for the reduction in variability, it is possible to use Bayesian multiple imputation (MI) procedures (Rubin, 1978; Rubin, 1987; Schafer, 1997). MI has not been used in meta-analysis in general and the goal of the current study is to investigate its use with MASEM.

Multiple Imputation

No applied MASEM study to date has examined the performance of multiple imputation (MI); however, MI is a promising technique for handling missing data found in MASEM research. MI expands upon single imputation and its resultant attenuation of variability. MI takes into account the uncertainty involved in missing data and imputes \( m \) plausible values (where \( m > 1 \)) to replace each single missing data point (each correlation estimate in MASEM research) resulting in \( m \) datasets. Each imputed dataset will have the same values for the non-missing correlation estimates. The values imputed for the missing data points will distinguish the \( m \) datasets. Each of the \( m \) datasets is analyzed using the statistical procedure of interest (i.e., the meta-analysis) and the results can be summarized across the imputed datasets. To obtain unbiased estimates using MI, the missing data are assumed to be at most MAR (thus, MI will also work well with MCAR data).

MI uses the Bayesian Markov chain Monte Carlo (MCMC) algorithm to impute values for missing data points. The reader is referred to several excellent chapters, texts and articles that provide more information on the technical process underlying MI (Allison, 2003; Peugh & Enders, 2004; Graham & Hofer, 2000; Schafer & Graham, 2002; Schafer & Olsen, 1998).

In traditional use of MI, the researcher calculates the statistic of interest (whether it is a sample mean, a correlation, a regression coefficient, etc.) represented generally as \( \hat{q}_i \) for imputed dataset \( i \). The simple average of the \( m \) estimates can be combined across imputed datasets to provide the multiply imputed estimate of the statistic using:

\[
\bar{q} = \frac{1}{m} \sum_{i=1}^{m} \hat{q}_i
\]  

(Rubin, 1987). The variance estimate associated with \( \bar{q} \) is a function of the variance within each imputed dataset

\[
\bar{u} = \frac{1}{m} \sum_{i=1}^{m} u_i
\]  

as well as the variability between the imputed datasets (Rubin, 1987)

\[
b = \frac{1}{(m-1)} \sum_{i=1}^{m} [\hat{q}_i - \bar{q}]^2
\]  

The total variance can be calculated using

\[
t = \bar{u} + \frac{1}{(1 + m)} b
\]  

In MASEM, the meta-analysis involves synthesizing correlations across studies. If MI were used, values for correlations would be imputed leading to the construction of \( m \) complete (imputed) datasets of correlations. A synthesized correlation is calculated for each correlation (e.g., \( r_{WX} \), \( r_{WZ} \), \( r_{XZ} \), etc.) in each dataset and each resulting synthesized correlation corresponds to the relevant \( \hat{q} \) (previously mentioned), thus, equations 1 - 4 can be used to calculate the MI estimate of each synthesized correlation and its associated variance.

Although parameter and standard error estimates can be easily combined using Equations 1 - 4, multivariate inferences, such as the test of homogeneity in meta-analysis, require different formulas. For example, Schafer’s (1997) formula for combining \( \chi^2 \) values (such as the one from the test of homogeneity) across studies is a relatively simple function of each imputation’s \( \chi^2 \) statistic value and its \( df \). The formula provides an \( F \)-ratio statistic for which an associated \( p \)-value can be estimated that can be interpreted as the significance test associated with the \( \chi^2 \). The formula is:
where $df_{\chi^2}$ is the $df$ associated with the $\chi^2$, $\bar{\chi}^2$ is the mean of the $m$ imputations’ $\chi^2$ values, $df_{\text{Error}}$ is the error degrees of freedom of the $F$-ratio statistic calculated as follows:

$$df_{\text{Error}} = \frac{(m-1)(1+\bar{r}_2^{-1})^2}{df_{\chi^2}^m}$$

and (Equation 5), $r_2$ is the sample variance of $\sqrt{\chi^2}$ across imputations where $r_2$ is calculated as follows:

$$r_2 = \frac{(1+1/m)}{(m-1)} \left[ \sum_{i=1}^{m} \chi_i^2 - \frac{\left( \sum_{i=1}^{m} \chi_i^2 \right)^2}{m} \right]^{1/2}.$$ 

Rubin also derived the formula for the efficiency of estimates based on $m$ imputed datasets

$$(1 + \gamma/m)^{-1}$$

where $\gamma$ is the fraction of missing information. In most cases between five and ten imputations are sufficient to achieve efficient results, however, with a large degree of missingness, more imputations may be necessary (Allison, 2003; Hershberger & Fisher, 2003). With a higher number of imputations, estimates of parameters become more stable (Allison, 2003). Allison notes that one diagnostic test of whether more imputations are necessary requires a check of the degrees of freedom for each parameter estimate. If the degrees of freedom are appreciably below 100 then more imputations should improve the efficiency of the estimates.

Typical MI procedures assume that data are multivariately normal. In MASEM, the typical unit of analysis is the correlation, $r$. The sampling distribution of $rs$ sampled from non-zero $\rho$, however, tends to be increasingly skewed for larger $|\rho|$ (Hedges & Olkin, 1985). The use of Fisher’s $Z_r$ transformation results in a more normal sampling distribution even for larger $\rho$ (Steiger, 1980). While the resulting $Z_s$ are approximately normally distributed, research has suggested that MI is reasonably robust to violations of the assumption of normality (Enders, 2001; Graham & Schafer, 1999). Graham and Schafer’s (1999) simulation study found that - even for extremely non-normal variables and small sample sizes - MI worked very well.

A benefit of using MI to handle missing data involves the less restrictive MAR missingness mechanism that can be assumed (unlike with PD and LD where only MCAR is assumed). Maximum likelihood (ML) methods and the expectation maximization (EM) algorithm, among others, also offer alternatives for handling missing data (Collins, Schafer & Kam, 2001). Use of MI, however, is less computationally intensive than ML (Sinharay , et al., 2001) and most MI programs use the EM algorithm to estimate starting values for the ensuing data augmentation iterations. Use of MI is further facilitated by its availability in several software packages including NORM (Schafer, 1999), SAS PROC MI (SAS Institute, 2005) S-PLUS (version 6.0, Insightful Corporation, 2001), and SPSS (version 14.0, SPSS, 2006).

To date, no meta-analytic researchers have used MI when handling missing data. The focus of the current study is to evaluate use of MI for synthesizing correlation matrix elements and their corresponding standard errors for use in MASEM. After missing correlations have been handled in MASEM, the researcher can synthesize the correlation matrix elements across studies. Before this synthesizing can occur, however, the researcher must decide whether to synthesize the correlations univariately or multivariately.
Synthesizing Correlations: Univariate Synthesis

Several methods exist that are used to synthesize effect sizes (here, correlations) across the k studies included in a meta-analysis (Hedges & Olkin, 1985). Synthesis methods typically involve weighting each effect, e, by a function of its associated sample size. The weight, w, most commonly used to obtain the pooled estimate of the effect size, \( \hat{\epsilon} \), is the inverse of the effect’s conditional variance (Cooper & Hedges, 1994):

\[
\hat{\epsilon} = \frac{\sum_{i=1}^{k} w_i e_i}{\sum_{i=1}^{k} w_i} .
\] (9)

This weighting assigns more weight to the more precise correlation estimates that are associated with larger sample sizes.

As noted, the sampling distribution of \( r \) is increasingly skewed for larger values of |\( \rho \)| (Hedges & Olkin, 1985). In addition, the large-sample variance of this distribution depends on the value of the parameter itself (Becker, 2000). For this reason, several meta-analytic researchers and in particular MASEM researchers (for example, Becker & Fahrbach, 1994; Hafdahl, 2001; 2007) recommend using Fisher’s (1928) \( r \)-to-\( Z_r \) normalizing and variance-stabilizing transformation:

\[
Z_r = .5\{\ln[(1 + r)/(1 - r)]\}
\] (10)

when synthesizing correlation estimates. The variance estimate associated with \( Z_r \) is:

\[
\hat{\sigma}^2_{Z_r} = \frac{1}{n_i - 3} .
\] (11)

The weight associated with the correlation estimate for study \( i \) is thus \( w_i = (n_i - 3) \), thus, to obtain the pooled estimate of the transformed correlation, \( \hat{\rho}_r \), between variables \( X \) and \( Y \), the following equation is used:

\[
\hat{\rho}_r = \frac{\sum_{i=1}^{k} (n_i - 3) Z_{r_i}}{\sum_{i=1}^{k} (n_i - 3)}
\] (12)

The resulting pooled, transformed correlation estimate, \( \hat{\rho}_r \), is then back-transformed to the correlation metric using the back-transformation formula:

\[
\hat{\rho} = \frac{\exp(2\hat{\rho}_r) - 1}{\exp(2\hat{\rho}_r) + 1}
\] (13)

to obtain the pooled estimate of the correlation, \( \hat{\rho} \). This univariate synthesis method can be used for each correlation in the matrix of interest. The standard error of \( \hat{\rho}_r \) is calculated using

\[
s_{\hat{\rho}_r} = \frac{1}{\sqrt{\sum_{i=1}^{k} n_i - 3}} .
\] (14)

Synthesizing Correlations: Multivariate Synthesis

Becker (1992b) introduced a multivariate synthesis method using generalized least squares (GLS) estimation that recognizes the possible dependencies among the \( p^* \) effect sizes (where, \( p^* = \frac{p(p-1)}{2} \) correlations among \( p \) variables in a correlation matrix):

\[
\hat{\theta} = \left(X\Sigma^{-1}X\right)^{-1}X\Sigma^{-1}T
\] (15)

where \( \hat{\theta} \) is a \( p^* \times 1 \) column vector containing the multivariately synthesized estimates of the \( p^* \) effect sizes, \( X \) is the design matrix consisting of \( k \) stacked \( p^* \times p^* \) that identifies matrices for \( p^* \) effect sizes per study, \( \Sigma \) is a block-diagonal matrix containing the covariance matrix for each study’s set of effect sizes as blocks along its diagonal and \( T \) is a \( kp^* \times 1 \) column vector containing each study’s effect size estimates. The omnibus \( Q \)-statistic is used to test the null hypothesis of the homogeneity of effect sizes.
(correlation matrices) across studies. It can be calculated using:

\[ Q = T \Sigma^{-1} T - \hat{\theta}'(X \Sigma^{-1} X)\hat{\theta} \quad (16) \]

and is assumed to follow a \( \chi^2 \) distribution with \((k-1)p^*\) degrees of freedom.

Olkin and Siotani’s (1976) large-sample approximation to the covariance between two correlations should be used to calculate elements of \( \Sigma \) in equations 15 and 16:

\[
\sigma_{z_{iu},z_{iv}} = \left[ 0.5 \rho_{iu} \rho_{iv} (\rho_{iu}^2 + \rho_{iv}^2 + \rho_{iu} \rho_{iv}) \\
+ \rho_{iu}^2 \rho_{iv} + \rho_{iv}^2 \rho_{iu} - (\rho_{iu} \rho_{iv})^2 \\
+ \rho_{iu} \rho_{iv} \rho_{iu} \rho_{iv} + \rho_{iu} \rho_{iv} \rho_{iu} \rho_{iv} \\
+ \rho_{iu} \rho_{iv} \rho_{iu} \rho_{iv} \right] / n_i
\]

(17)

Alternatively, multivariate synthesis with GLS estimation (see equation 15) could be used to synthesize \( Z_r \)-transformed correlations (equation 10). Elements of the covariance matrix, \( \Sigma \) for the \( Z_r \)s are a function of the covariances between the correlations (equation 17) and can be calculated using:

\[
\sigma_{z_{iu},z_{iv}} = \frac{n_i \sigma_{z_{iu},z_{iv}}}{(n_i - 3)(1 - \rho_{iu}^2)(1 - \rho_{iv}^2)}
\]

(18)

(Steiger, 1980). Initially, when demonstrating use of GLS synthesis in a simulation study (Becker, 1992b), individual study estimates of \( \rho \) were used in equations 17 or 18 when calculating \( \Sigma \). Use of these less efficient single study estimates of \( \rho \) was later found to be one cause of GLS’ poor performance for synthesizing correlation matrix elements (Becker & Fahrbach, 1994).

Researchers have found that multivariate GLS tends to outperform univariate synthesis methods when a pooled estimate of \( \rho \) is instead substituted for each \( \rho \) in equations 17 or 18 (Becker & Fahrbach, 1994; S. Cheung, 2000; Furlow & Beretvas, 2005). Lastly, it should be mentioned that when there is no missingness, or the missing values have been replaced using some type of imputation, the results with \( Z_r \)-transformed correlations from GLS and univariate synthesis are mathematically equivalent (Gagné, Furlow, & Beretvas, 2004).

Hafdahl (2007) conducted a study evaluating the performance of univariate and multivariate synthesis methods paired with \( r \) and Fisher’s transformation, \( Z_r \), as well as using the more efficient estimates of \( \rho \) in the relevant weight (univariate or multivariate) matrix. Hafdahl found support for using the \( Z_r \) transformation over \( r \), for multivariate over univariate synthesis methods and for substituting the pooled estimates of \( \rho \) instead of using individual study estimates. Combining these options led to better parameter estimation accuracy, efficiency and precision and for improved Type I error control for the test of homogeneity.

Hafdahl (2007) only investigated the performance of synthesis methods when no data (i.e., correlation estimates) were missing in any of the studies being meta-analyzed. In cases where not all correlation estimates are provided in every study, the relevant rows and columns are deleted from the matrices (specifically in the \( T \), \( \Sigma \), and \( X \) matrices) used in GLS (equation 15) and the \( Q \)-statistic (equation 16) estimation (Becker & Schram, 1994). Other researchers have assessed the impact of missing data on MASEM. Similar to Hafdahl’s results, Furlow and Beretvas (2005) found that the \( Z_r \) transformation along with use of pooled average estimates of \( \rho \) substituted for \( \rho \) in the elements of the \( \Sigma \) matrix worked best as a synthesis method. Furlow and Beretvas (2005) also compared the results from their study when correlations were MCAR and MNAR and when LD versus PD was used to handle the missing correlations. They found that MNAR data produced high levels of relative bias in the correlation estimates while percent relative bias among the correlations for MCAR data was never above 3%. Use of PD resulted in enhanced estimation of synthesized correlations when compared with LD when it was used along with the more efficient method for GLS.

Cheung and Chan (2005) demonstrated the use of multi-group SEM (where each study comprises a group) and model parameter constraints across groups as a way to conduct
MASEM analyses. They compared the performance of their procedure with the more typically used MASEM procedure (in which the elements in the correlation matrix are first meta-analytically synthesized and then analyzed using canned SEM software without the covariance matrix associated with the synthesized correlation matrix being analyzed).

The authors evaluated both procedures when data were missing and found support for their procedure, however, the authors used the earlier version of GLS with individual study estimates of $\rho$ when calculating the covariance matrix, $\Sigma$, used in GLS’ multivariate synthesis (equation 15) and for the $Q$-statistic (equation 16). Therefore, it was not surprising that GLS did not perform well. In addition, Cheung and Chan only considered $k = 5, 10$ and $15$ in their study (well below the median $k$ of 26 that they reported in their review of the applied literature). Although they acknowledged that their largest $k$ was smaller than the average reported, they indicated that because their method involved the cross-group constraints (where each study is considered as a group) their method was too computationally intensive to involve larger $k$s, thus providing a weakness to their method for MASEM.

Both methodological MASEM studies (Furlow & Beretvas, 2005; Cheung & Chan, 2005) had also generated data such that the sample size associated with each study was the same. While use of a single sample size for every study in the simulated meta-analysis might simplify interpretation of results, it does not provide an authentic simulation of reality. Instead, in a real-world meta-analysis the sample size for each study is typically different.

Methodology

This simulation study was designed to investigate the use of MI for pooling estimates of correlation matrices when some correlation estimates were missing in the primary studies being synthesized. For this exploration of the use of MI, the synthesis of elements of a four-variable correlation matrix was investigated with MCAR data. Manipulated conditions in the study included the degree of missingness (25% and 50% of all correlations), the number of studies ($k = 25, 50$ and $100$) involved in the meta-analysis, and the average sample size per study ($\bar{n} = 50$ and $100$).

Because MI assumes data are multivariate normal, it was of interest in this study to transform the correlations to Fisher’s $Z$ metric since its sampling distribution is more normal than that of $\rho$. As noted, when $Z_r$ is used, results from the more efficient version of GLS are equivalent to using univariate weighting when no data are missing or missing data values have been imputed (Gagné, Furlow & Beretvas, 2004). The results from the omnibus test of the homogeneity of variance, however, will not be the same for univariate and multivariate synthesis. Thus, the performance of univariate versus multivariate synthesis methods was compared when assessing the Type I error performance for the homogeneity test. Use of MI was also assessed in terms of resulting parameter and standard error estimation for only the univariate synthesis of correlations.

In applied meta-analysis, study results tend to be based on uneven sample sizes. To mimic this, each study’s sample size, $n_i$, was generated using the same distribution as that used in Hafdahl’s (2007) simulation study:

$$n_i = \left\lceil \frac{\bar{n}}{2} \left( \frac{X_i - 3}{\sqrt{6}} \right) + \bar{n} \right\rceil$$

(19)

where $X_i$ (for $i = 1$ to $k$) was sampled for each study from a $\chi^2$ distribution with 3 degrees of freedom. The value of $\bar{n}$ depended on the sample size condition. In the current study, for small and moderate $\bar{n}$ conditions, the values of $\bar{n}$ were 50 and 100, respectively. Last, $\{y\}$ represents the closest integer to the value of $y$.

Data Generation

Multivariate normal fixed-effects data were generated in SAS (SAS Institute, 2005) using the Cholesky root of the generating covariance matrix. For each combination of conditions, 1,000 replications were conducted. The relevant degree of missingness was introduced into the dataset, the missingness was then handled using MI, and correlation estimates were synthesized.
Model Generation

To simplify this exploration of the use of MI with MASEM analyses, a scale-invariant model was selected to generate the data. Scale-free parameter and standard error estimation results for a scale-invariant SEM model are equivalent whether a correlation or a covariance matrix is analyzed (Cudeck, 1989). A four-variable, one-factor (scaled to have a variance of one) model was used. Values for the elements of the correlation matrix used to generate the data are those implied by the relevant generating values for the true factor loadings (with loading values of 0.5, 0.6, 0.7 and 0.8 for variables V1, V2, V3 and V4, respectively). Table 1 shows the model-implied values of the correlations used to generate the data. To simplify the study, the variables’ and factors’ variances were each standardized to have a value of one.

Table 1: Generating Values of Model-Implied Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V2</td>
<td>0.30</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V3</td>
<td>0.35</td>
<td>0.42</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>V4</td>
<td>0.40</td>
<td>0.48</td>
<td>0.56</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Corresponding generating loading values were 0.5, 0.6, 0.7 and 0.8 for V1, V2, V3 and V4, respectively.

After data were generated, MCAR missingness was introduced. The premise underlying the general design of missingness in this simulation study was that once a study was (randomly) selected to have missingness introduced, then a variable was randomly selected as one that was not measured in a study. Once a variable was selected to be missing, all correlations involving that variable were designated as missing. Thus, if variable V1 was selected then \( r_{12}, r_{13}, r_{14} \), would each be missing for a study: this mimics a realistic meta-analytic scenario in which a variable is not measured in a study and thus associated correlations could not be reported.

To determine reasonable values for the degree of missing data and the number of studies synthesized in the meta-analysis for this study, a review of applied MASEM studies in the literature was conducted through a search of the PsycInfo database using the search criteria “meta-analysis” and each of “structural equation modeling”, “path analysis” and “confirmatory factor analysis”. In addition, other applied MASEM articles cited in the resulting sources or known to the authors were also examined. This led to the identification of 24 applied MASEM studies. The amount of missing correlations could only be determined for 13 of these 24 studies because authors did not report the information needed to calculate these percentages. Two studies reported no missing correlations while at the other extreme, three studies reported over 80% of all correlations missing. The mean rate of missing correlations was 67.8% while the median rate was 70%. The mean number of studies synthesized across all 24 MASEM studies was 49.6 with correlations being pooled from a minimum of four to a maximum of 155 studies’ results.

The number of studies in the simulated meta-analysis used in this study were chosen to reflect small (25), moderate (50) and large (100) numbers of studies. Per-study sample size was varied as described in equation 19 with two levels for the average per-study sample size (\( \bar{n} = 50 \) and 100). Two percentage levels of missing correlations were chosen (25% and 50%) to reflect the amounts of missingness found in applied MASEM studies. In conditions where 25% of the correlations were missing in a meta-analysis, 30% of the studies were first selected and then 50% (2) of the four variables within those studies were chosen to be missing (resulting in one correlation out of six remaining in those studies). In conditions with 50% of correlations missing, 60% of the studies were selected to have missingness and 50% of the variables within those studies were designated as missing along with their correlations. Baseline conditions where no correlations were missing were also examined for each combination of \( k \) and \( \bar{n} \).
Synthesis Method

Following dataset generation for each condition and introducing the missingness, MI was used. To multiply impute the data, SAS PROC MI was utilized, employing a noninformative prior (the default for PROC MI) and assuming a multivariate normal posterior distribution. Because a relatively large degree of missingness was simulated in this study, forty imputations were used rather than the typical five to ten (Allison, 2003). The forty imputed datasets each consisted of a full set of correlations for each study in each simulated meta-analysis.

After the forty multiply imputed datasets had been calculated for each replication the contents of each dataset were used to obtain forty synthesized matrices. To synthesize each correlation estimate, each study’s $r_{ij}$ value was transformed to $Z_{rij}$ using equation 10. These were pooled together using equation 12 to obtain the univariately pooled $\hat{\rho}_{ij}$ for each pair of variables $i$ and $j$. The standard error estimates were also calculated using equation 14. The resulting estimates of the Fisher-transformed correlation matrix elements and associated standard error estimates were then combined across the 40 imputed datasets per replication using Rubin’s combination rules (see equations 1-4) through PROC MIANALYZE.

Performance of the $Q$-statistic for correct identification of the homogeneity of the correlation matrices across studies was also evaluated. The $Q$-statistic was calculated with the covariance matrix, $\Sigma$ in equation 16 containing only variances of the $Z_{s}$s along the diagonal for the test of homogeneity for the univariate synthesis. The $Q$-statistic was also calculated using the full covariance matrix (see equations 17 and 18) for GLS. Rather than using single-study estimates of $\rho$, the more efficient pooled estimates were used because they have been found to enhance the performance of GLS (Hafdahl, 2007). Per-imputation $\chi^2$ estimates and associated p-values were combined across imputations using Allison’s SAS macro COMBCHI (2000) (which utilizes equations 5, 6, and 7). A correction to COMBCHI corrected a small error in the code (Enders, personal communication, December 8, 2005).

Data Analysis

The relative percent bias, $B(\hat{\theta})$, was used to evaluate estimation of correlations (Hoogland & Boomsma, 1998). Hoogland and Boomsma recommended identification of bias when the magnitude of $B(\hat{\theta})$ exceeds five percent of the corresponding population value. The accuracy of the standard error estimates associated with each correlation was assessed using the standard error’s relative percent bias. Hoogland and Boomsma suggested that standard error relative percent bias of magnitude 10 percent or less indicates an acceptable degree of bias. Finally, the proportion of correct fixed-effects model identifications were tallied using the univariately and multivariately weighted $Q$-statistic (see equation 16).

Results

Parameter Estimation Bias

No substantial bias was found under any of the conditions examined for estimation of the correlations. Relative percent bias for each element across conditions and matrices never exceeded a magnitude of 1%.

Standard Error Estimation Bias

Table 2 lists the results from all study conditions and all correlations for the standard error bias. In cells with no missing correlations, percent relative bias was always well below Hoogland and Boomsma’s (1998) 10% cutoff with a highest magnitude of 5.5%. In cells with missing data, the bias was always positive and a distinction was apparent in the bias for the small (25%) and large (50%) degree of missingness conditions.

For conditions with 25% of correlations missing, bias magnitude was always below 10% for cells with $k$s of 50 and 100, except unexpectedly for $\rho_{24}$ with $k = 100$ and $n = 100$. With a $k$ of 25 and $n = 50$, bias was consistently above 10%. With a $k$ of 25 and $n = 100$, the magnitude of the bias decreased below 10% for all $\rho$s except $\rho_{14}$. In cells with 25% missing data and $k = 25$, bias ranged from 4% to
In cells with 50% of correlations missing the bias increased, with larger bias for smaller \( k \). In these cells with a \( k \) of 25, bias ranged from 12.1% to 22.3% with a mean of 17.2% (\( SD = 3.1 \)).

Q-Statistic’s Correct Model Rejection Rates

All data were generated with homogeneous correlation matrices. Table 3 shows the proportion of instances that the \( Q \)-statistic led to an incorrect inference that there was heterogeneity in the correlation matrices for cells with no missing correlations. Univariate weighting of the \( Q \)-statistic led to lower incorrect model rejection rates than did multivariate weighting. However, the average rejection rate (\( M = 8.3\% \), \( SD = 0.8 \)) for univariate weighting exceeded the nominal level of 5%. The average for multivariate weighting was 17.6% (\( SD = 5.4 \)). In general, for GLS, the rejection rates increased as \( k \) increased. This was not the case for univariate synthesis. For conditions with missing correlations, MI never led to an incorrect rejection of the null hypothesis (i.e., the rejection rate was always 0%), regardless of synthesis method, thus the results are not presented in a table.

Conclusion

Use of MI for meta-analysis resulted in synthesized correlations without substantial parameter estimation bias when data were...
MASEM WITH MULTIPLE IMputation

MCAR. Substantial positive standard error bias was found, but typically only for smaller meta-analyses ($k$ of 25) and this bias was higher with larger degrees of missing data. Cheung and Chan (2005) also identified problems with MASEM when values of $k$ were low. Based on the review of applied MASEM studies, however, it appears that most MASEM studies involve larger $k$ values. From these results, use of MI with meta-analysis might be advocated. The results for the test of homogeneity, however, indicate that MI should not be used for testing the homogeneity of correlation matrices across studies. Although MI never resulted in an incorrect inference that there was heterogeneity in the correlation matrices, future research is likely to indicate that when this test should be rejected (i.e., when there is heterogeneity present), MI will have insufficient power.

While parameter and SE estimate results from GLS and univariate weighting utilizing the $z$-transformation are equivalent when the data being analyzed has no missingness, their results differ for the test of homogeneity. GLS resulted in substantially higher incorrect rejection rates than did the univariate weighting method and the rates for GLS increased as $k$ increased. These findings are consistent with those from Hafdahl (2007) where GLS synthesis resulted in higher Type I error rates when compared with univariate synthesis when no data were missing. Thus in the context of testing for heterogeneity, the univariate weighting method is recommended.

Limitations and Future Directions

One limitation of this study is the use of only a fixed-effects model both for data generation and model estimation. While this is a limitation, applied MASEM studies most frequently assume a fixed-effects model and therefore the performance of the conditions in this study under this assumption provide an important starting point for this research. A random-effects model, however, might provide a more appropriate fit in many MASEM studies, particularly when important between-study characteristics impact the variability found among studies’ correlations. Hafdahl (2008) recently compared GLS with univariate synthesis under a random-effects model with no missing correlations. Hafdahl found that while both methods had high power to reject homogeneity when at least 50 studies were used in the meta-analysis, when fewer than 50 studies were used GLS had far superior power performance over univariate weighting. This difference was particularly noticeable when the average per study sample size was at least 100. Future research should continue to explore the differences between the univariate and multivariate synthesis methods for their power and Type I error control particularly when missing data occur. Given MI’s Type I error performance for the test of homogeneity, its performance with between-study heterogeneity should also be evaluated to see if it exhibits the weak power that would be expected.

As noted, MI worked well in most conditions for estimation of correlations and their standard errors (typically substantial standard error bias was only found with 50% of correlations missing). The results for the test of homogeneity with MI, however, seem to indicate a problem with its use. Because Schaefer’s (1997) equations for combining the relevant $p$-values from the $\chi^2$ test of homogeneity seem not to have worked well in the conditions examined here, it would seem that future research should explore whether these

<table>
<thead>
<tr>
<th>Study Conditions</th>
<th>Synthesis Method</th>
<th>GLS</th>
<th>Univariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$\tilde{n}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>13.1</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>10.9</td>
<td>9.1</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>17.2</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>16.8</td>
<td>7.6</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>25.7</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>21.7</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Note: See equation 19 for $\tilde{n}$. Results for cells with missing correlations are not reported here because MI always resulted in never rejecting the test of homogeneity.

Table 3: Homogeneity Assumption Rejection Rates for Omnibus Q-Statistic for Conditions with No Missing Correlations
equations can be refined or new equations developed. Additionally, future research should also evaluate the performance of MI with missing data mechanisms other than MCAR.

Use of MASEM techniques will continue to increase as educational researchers use meta-analysis to summarize past research and SEM to investigate relationships between observed and latent variables. It is hoped that the results from this study will help inform the use of, and lead to continued refinement of, MASEM techniques for educational and psychological research.

References


Median-Unbiased Optimal Smoothing and Trend Extraction

Dimitrios D. Thomakos
University of Peloponnese,
Tripolis, Greece

The problem of smoothing a time series for extracting its low frequency characteristics, collectively called its trend, is considered. A competitive approach is proposed and compared with existing methods in choosing the optimal degree of smoothing based on the distribution of the residuals from the smooth trend.

Key words: Local linear, moving average, singular spectrum analysis, smoothing, splines, time series, trend extraction.

Introduction
A fundamental problem in time series analysis is smoothing a realization and extracting its low-frequency characteristics, collectively called its trend. In the process of solving this problem a practitioner is faced with three underlying sub-problems: (a) to define the nature of the trend (e.g., deterministic or stochastic) and its perceived degree of smoothness, (b) to decide on a particular class of models to use (e.g., polynomial or non-parametric approximations), and (c) to select, usually with a data-based approach (e.g., cross-validation) the degree of approximation (or smoothness) that will enable accurate extraction of the required trend features. A large amount of literature exists which deals with these problems and includes various proposed methods for addressing them. Although it is not possible to review this literature here; many related references can be found in books and monographs, such as, Härdle (1990), Fan and Gijbels (1996), Hart (1997), Golyandina, et al. (2001) and Fan and Yao (2003).

The methods depend on various assumptions about the data generating process (DGP) itself and its stability over time. However, in many applications one does not know or is not willing to make assumptions about the structure of the DGP and, consequently, is led to use an approach unrelated to such specific assumptions. Examples include moving average (MA) smoothing, singular spectrum analysis (SSA) smoothing and all the known forms of non-parametric smoothing, like smoothing splines (SS) and local linear (LL) smoothers. This choice of a non-parametric approximation usually takes care of problem (b), and partially (c) if methods such as cross-validation or plug-in bandwidths are used.

As for problem (a), it is usually the case that the nature of the trend that one wants to extract is application-specific, as is its perceived degree of smoothness. However, some characteristics exist that are commonly accepted about the notion of a trend, such as: (i) it has most of its power concentrated in (a band of) the lower frequencies of the spectrum, (ii) it is more smooth (less volatile) than the actual observations, (iii) it reflects the central tendency of the process, and (iv) the observations are usually located in clusters above or below the trend component.

Problem (c) is thus left, i.e. that of selecting the appropriate optimal degree of smoothing the observations for extracting the trend component. In the context of non-
parametric methods, such as SS and LL, the choice of the degree of smoothing is guided by
the bias-variance trade-off and a proximity criterion - such as the mean squared error (MSE) or
the integrated mean squared error - is minimized directly or by variants of cross-validation/plug-in methods. However, such criteria are invariably linked to the notion of fit
(of various degrees) to the observations themselves not to the notion of an underlying trend. This runs contrary to the notion of the trend that passes through the center of the clusters of observations without tracking all their swings. In addition, for methods such as MA or SSA there are no formal procedures for selecting the degree of smoothing; the results of the proposed methodology can be applied in making such selections to these two methods as will be illustrated.

Methodology
Consider a stochastic process \( \{ X_t \}_{t \in \mathbb{Z}} \) and assume that a realization of size \( n \) from this process is available, for example, \( \{ x_t \}_{t=1}^n \). The problem is how much to smooth the realization so as to successfully extract the low frequency characteristics, or the trend, of the process. No assumptions are made as to whether the trend is deterministic or stochastic. Such smoothing will lead to an additive decomposition of the form:

\[
X_t = s^k_t + u^k_t
\]

where \( s^k_t \) is the estimated smoothed component (the trend) of the series, that depends on a smoothing parameter \( k \), and \( u^k_t \) is the estimated residual that also depends on \( k \). Note that the above decomposition is not taken as the data generating process; rather it is the result of the smoothing operation. In particular, \( u^k_t \) is not assumed to be a realization from a true error process acting on \( X_t \). As such, the representation of equation (1) has applicability both in cases where a deterministic slowly varying function of time exists and where it does not, for example, in the context of a financial time series that possibly follows a random walk.

The way the residuals \( u^k_t \) are distributed is important in understanding whether a component that roughly corresponds to the characteristics (i) to (iv) attributed to the trend of a realization has been successfully extracted. First, recall that any smoothing operation that successfully extracts a measure of central tendency leads to residuals with an approximately zero mean. It does not, however, necessarily lead to residuals that have zero median, so as to have a residual distribution where equal probability is placed in observing positive (above the trend) and negative (below the trend) residuals.

This probabilistic symmetry of the residuals should be important because an extracted trend cannot possibly be accurate if it leads, on average, to more positive than negative residuals (or vice versa). In such a case the trend would be biased, either over- or under-estimating the low frequency movement of the process. If the problem of trend extraction is considered in the above context of symmetrizing the probability assigned to positive and negative residuals, it is necessary to look for a measure different than the MSE. A plausible way of proceeding is as follows.

Let \( \text{sgn}(x) = I(x > 0) - I(x \leq 0) \) denote the sign function and note that for any continuous random variable \( X \), with \( F_X(\cdot) \) as its distribution function,

\[
\text{Esgn}(X) = \text{EI}(X > 0) - \text{EI}(X \leq 0) = 1 - 2F_X(0),
\]

(2)

the absolute value of the expected sign of \( X \), \( |\text{Esgn}(X)| \), is symmetric around \( F_X(0) = 0.5 \) where it attains its unique minimum. It therefore follows that if the distribution of \( X \) is symmetric around zero (i.e., has a zero median) then \( |\text{Esgn}(X)| \) is minimized.

This can be adapted into a smoothing context and the absolute value of the expected sign of the residuals \( u^k_t \) can serve as the
MEDIAN-UNBIASED OPTIMAL SMOOTHING AND TREND EXTRACTION

objective function that should be minimized in choosing the degree of smoothing. Essentially this amounts to choosing the degree of smoothing so as to assign roughly equal probability to positive and negative residuals in accordance with characteristics (i) to (iv). This leads to consideration of the following:

\[
\mathbb{E}\text{sgn}(u_i^k) = |1 - 2F_u(0)| \quad (3)
\]

where \( F_u(\cdot) \) denotes the distribution function of the residuals. As noted, this function is minimized when \( F_u(0) = 0.5 \iff u_{(0.5)}^k = 0 \), that is, when the residuals are made to have a zero median. The trend component which will correspond to such residuals can now be called a median-unbiased trend.

To practically implement this idea consider the empirical version of equation (3) which can be estimated in two equivalent ways as follows:

\[
|\text{MRS}(k)| \equiv \left| n^{-1} \sum_{i=1}^n \text{sgn}(u_i^k) \right| \\
= \left| 1 - (2n)^{-1} \sum_{i=1}^n I(u_i^k \leq 0) \right| \quad (4)
\]

where MRS denotes the mean residual sign based on the sample of observations. As can be observed from equation (4), the MRS can be obtained either using the average sign or using the empirical distribution function evaluated at zero. The most practical way of optimizing the \( |\text{MRS}(k)| \) is by direct search over a grid of plausible values for the smoothing parameter. If \( K \equiv \{k_{\min}, k_{\max}\} \) denotes such a grid then the optimal value \( k^* \) is given by:

\[
k^* = \arg\min_{k \in K} |\text{MRS}(k)| \quad (5)
\]

The range of grid values to consider is both problem-specific and method-specific and no general guidelines can be given. For example, if a moving average is to be used for smoothing, then \( k \) takes only integer values; if a kernel smoother is to be used then \( k \) takes real values – possibly in a pilot interval. To overcome this potential shortcoming one can alternatively consider using data-dependent, sub-sampling approaches. One variant of such a sub-sampling approach could be as follows:

1. Split the observations into \( M \) non-overlapping sections each of equal length \( m \leq n / 2 \), for \( j = 1, \ldots, M \), with \( m \to \infty \) as \( n \to \infty \) and \( m / n \to c \) for some constant \( c \).
2. Select a range of plausible values for each section, for example \( K_j \).
3. Compute the optimal value of the smoothing parameter for each section, for example \( k_j^* \).
4. Select the full sample optimal value of the smoothing parameter as the average of the parameters from each section, i.e., \( k^* = M^{-1} \sum_{j=1}^M k_j^* \).

The above is just one sub-sampling method. Alternatively, the series can be split using a sliding window of length \( m \), thus having \( M \) overlapping sections each of length \( m \). This alternative is not further pursued herein but is easily implementable.

Results

The above methodology was applied to simulated time series and a real time series using different smoothers: symmetric MA, SSA, LL, SS and the Kalman fixed point (KF) smoother. All methods are appropriate under different conditions for the data generating process. For the LL smoothing and SS methods the degree of smoothing selected by the present methodology was compared with the degree of smoothing selected using generalized cross-validation (GCV) and plug-in (plug) methods respectively. The SSA smoother was used as in Thomakos (2008) with an asymptotically optimal decomposition of the covariance of the process, when the process has stochastic trends. All computations reported below were performed in R.
Simulated Series

Two types of data generating process (DGP) were considered. The first is given as the sum of a deterministic, slowly varying function \( g(t/n) \) and stationary errors and the second is given as the sum of a stochastic trend (a random walk) and stationary errors. Specifically, for the first DGP:

\[
\text{DGP I: } x_t = g\left(\frac{t}{n}\right) + u_t \tag{6}
\]

with \( g\left(\frac{t}{n}\right) = \alpha + \beta t + \sum_{j=1}^{2} \gamma_j \cos\left(\frac{2\pi \omega_j t}{n}\right) \)

and with \( u_t = \phi u_{t-1} + \varepsilon_t, \varepsilon_t \sim N\left(0, \sigma^2_t\right) \). For the trend function \( g\left(\frac{t}{n}\right) \) the critical parameters determining the degree of smoothness (and the complexity of the curve) are the frequencies \( f_j = \omega_j / n \); higher values decrease smoothness - see Figure 1 for an illustration (the black line corresponds to the less smooth trend, the red (or gray) line to the more smooth trend).

For the final series \( x_t \), the critical parameters are \( \left[ \phi, \sigma^2_t / \left(1 - \phi^2\right) \right] \), the persistence and the variance of the error term; higher values make it more difficult to separate the trend from the errors. In the end, consider the following combinations for the parameters:

\[
\alpha = 0, \beta = 2, \gamma_1 = 0.50, \gamma_2 = -0.25, \\
\omega_1 = 2, \omega_2 = \{5,10\}, \phi = \{0.0,0.8\}, \\
\sigma^2_t = 0.2^2
\]

For the second DGP consider the well known form of signal-plus-noise or local level model as:

\[
\text{DGP II: } x_t = g\left(\alpha, S_t\right) + u_t \tag{8}
\]

with \( g\left(\alpha, S_t\right) = \alpha + S_t \), where \( \alpha \) is the drift parameter, \( S_t = \sum_{j=1}^{t} \varepsilon_j \) is the random walk component of the series with normally
distributed errors $\varepsilon_i \sim N\left(0, \sigma^2_\varepsilon\right)$, and where $u_i$ are the added errors that have either a normal or a $t$-distribution, that is, $u_i \sim N\left(0, \sigma^2_u\right)$ or $t_{\nu}$, where $\nu = \nu(6)$. The drift parameter is set to $\alpha = 0.1$, and the variance of the normally distributed $u_i$ is set to $\sigma^2_u = 0.6^2$ (the later corresponds to a 1:3 signal-to-noise ratio). Typical sequences from the DGP of equation (8) are shown in Figure 2 (the black (upper) line corresponds to normally distributed additive errors, the red (lower) line to additive $t$-distributed errors).

From each DGP, $r = 1, 2 \ldots 400$ realizations of sizes $n = \{200, 400\}$ were simulated and for each realization the full sample and the sub-sampling approach was used, the latter with $m = \{50, 100\}$ for the corresponding sample sizes, to compute the optimal value of the smoothing parameters of each method. The ranges of plausible values for minimizing the MRS were set to the following:

- For the symmetric MA and SSA methods that use integer values: $K = \{2k + 1 | k = 1, 2, \ldots, 11, 12\}$
- For the local linear smoothing that uses real values for the bandwidth: $K = \{1.5^{k-12} s_x | k = 1, 2, \ldots, 11, 12\}$, where $s_x$ denotes the standard deviation of the data.
- For the smoothing splines that use real values for the smoothing parameter $K = \{k | k = 0.00, 0.14, \ldots, 1.36, 1.50\}$, a sequence of 12 values in the interval $[0.0, 1.50]$.

With the selected $k^*$, as computed either with the full sample or the sub-sampling approach, the mean absolute deviation of the true trend component from the estimated trend component is computed for each replication, that is $m_{r}^{y_{ts}} = n^{-1} \sum_{r=1}^{n} \|g_{r}(\cdot) - s_{r_{ts}}\|$. Finally, the

![Figure 2: Sample Realization of Stochastic Trend from DGP II of Equation (8)](image-url)
average, \( m^{k^*} = 400^{-1} \sum_{r=1}^{400} m_r^{k^*} \), was computed as well as the optimal values of \( k^* \) from all 400 replications (note that the reported replication averages for integer \( k^* \) will not necessarily be odd numbers). These measures are reported in Tables 1 and 2 which show the results on the simulations for DGP I, and in Table 3 which shows the results for DGP II.

Discussion of Simulation Results: DGP I

For the smaller sample size of \( n = 200 \) (see Table 1), the discussion can be separated into two cases: one for \( \phi = 0 \) and the other for \( \phi = 0.8 \); for the first case also note some small differences depending on the value of \( \omega^2 \). Thus, for the parameter combination \( \phi = 0, \omega^2 = 5 \) the performance of the sub-sampling approach is improving the average accuracy in extracting the trend for the MA and SS methods. There is no change for the local linear smoother. Note that only the moving average coupled with sub-sampling performs on par with the GCV-based and plug-in approaches but this is an important result: the smoothing spline and local linear methods have their own approaches (GCV and plug-in) for selecting the degree of smoothing while the for a moving average there is no such existing method.

For the parameter combination \( \phi = 0, \omega^2 = 10 \) however the results are much less satisfactory since no alternative beats the GCV-based and plug-in-based approaches. Turning next to the parameter combinations where \( \phi = 0.8 \) a much improved picture results in terms of the performance of the proposed methodology and the use of moving averages.

Table 1: Average Absolute Deviation of True from Estimated Trend & Optimal Degrees of Smoothing; Simulations from DGP I and Sample Size \( n = 200 \)

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m^{k^*} )</td>
<td>0.09</td>
<td>0.06</td>
<td>0.13</td>
<td>0.08</td>
<td>0.05</td>
<td>0.13</td>
<td>0.16</td>
<td>0.05</td>
</tr>
<tr>
<td>( k^* )</td>
<td>10</td>
<td>8</td>
<td>0.48</td>
<td>0.33</td>
<td>0.58</td>
<td>0.09</td>
<td>0.09</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m^{k^*} )</td>
<td>0.19</td>
<td>0.19</td>
<td>0.24</td>
<td>0.23</td>
<td>0.24</td>
<td>0.21</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td>( k^* )</td>
<td>8</td>
<td>7</td>
<td>0.45</td>
<td>0.31</td>
<td>0.22</td>
<td>0.08</td>
<td>0.08</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m^{k^*} )</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.09</td>
<td>0.06</td>
<td>0.17</td>
<td>0.17</td>
<td>0.07</td>
</tr>
<tr>
<td>( k^* )</td>
<td>10</td>
<td>8</td>
<td>0.50</td>
<td>0.30</td>
<td>0.44</td>
<td>0.11</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m^{k^*} )</td>
<td>0.21</td>
<td>0.21</td>
<td>0.24</td>
<td>0.23</td>
<td>0.24</td>
<td>0.22</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td>( k^* )</td>
<td>9</td>
<td>7</td>
<td>0.44</td>
<td>0.32</td>
<td>0.22</td>
<td>0.08</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Here the use of either the full or sub-sampling approaches coupled with a moving average produces is better (when $\omega_2 = 5$) or on par (when $\omega_2 = 10$) with the alternative methods.

When the sample size increases to $n = 400$ (see Table 2) further improvements are observed in performance from the use of the proposed methodology – especially from the use of moving averages. Specifically, in all four parameter combinations considered, a moving average coupled with sub-sampling performs on par or better than GCV-based and plug-in-based approaches. Note that this improvement is more pronounced in some cases and is worth elaborating about. For example, in the case where $\phi = 0.8, \omega_2 = 5$ the moving average performs on par with the local linear smoother with plug-in selection of bandwidth; the smoothing splines do not perform as well. In the case where $\phi = 0.8, \omega_2 = 10$, the moving average with sub-sampling outperforms the local linear smoother. Finally, the smoothing splines with sub-sampling now perform on par with the GCV-based smoothing splines.

The results from the DGP I simulations show that the proposed methodology can be competitive to existing methods, by either: (1) assisting less sophisticated methods, such as moving averages, to perform well in smoothing and trend extraction, and/or (2) producing results using other methods, such as smoothing splines that are equivalent to the more sophisticated GCV or plug-in approaches.

Simulation Results Discussion: DGP II

Recall that the simulations of the second DGP of equation (8) do not have an underlying deterministic smooth function that serves as the trend component, but rather have a stochastic trend that is masked by additive errors. This type of DGP has a corresponding optimal smoother,

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^k$</td>
<td>0.05</td>
<td>0.04</td>
<td>0.13</td>
<td>0.06</td>
<td>0.03</td>
<td>0.12</td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>$k^*$</td>
<td>11</td>
<td>10</td>
<td>0.55</td>
<td>0.40</td>
<td>0.59</td>
<td>0.10</td>
<td>0.09</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^k$</td>
<td>0.17</td>
<td>0.17</td>
<td>0.22</td>
<td>0.20</td>
<td>0.24</td>
<td>0.20</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>$k^*$</td>
<td>10</td>
<td>9</td>
<td>0.46</td>
<td>0.36</td>
<td>0.15</td>
<td>0.10</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^k$</td>
<td>0.10</td>
<td>0.07</td>
<td>0.14</td>
<td>0.07</td>
<td>0.05</td>
<td>0.15</td>
<td>0.17</td>
<td>0.05</td>
</tr>
<tr>
<td>$k^*$</td>
<td>11</td>
<td>10</td>
<td>0.58</td>
<td>0.38</td>
<td>0.45</td>
<td>0.11</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SS-full</th>
<th>SS-sub</th>
<th>SS-GCV</th>
<th>LL-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^k$</td>
<td>0.19</td>
<td>0.17</td>
<td>0.22</td>
<td>0.21</td>
<td>0.24</td>
<td>0.20</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>$k^*$</td>
<td>10</td>
<td>9</td>
<td>0.50</td>
<td>0.34</td>
<td>0.15</td>
<td>0.09</td>
<td>0.07</td>
<td>0.01</td>
</tr>
</tbody>
</table>

150
based on the state space representation of equation (8), the Kalman fixed point smoother. Results can thus be compared to this natural benchmark. Here the results are much more uniform across sample sizes and distributions and highly encouraging. For all cases considered in Table 3 there is at least one instance of either the MA or the SSA smoother, with sub-sampling, that

From the above discussion it is clear that a carefully, data-based, selected MA or SSA smoother can potentially perform as well or better than more sophisticated methods when extracting a stochastic trend from underlying additive errors. Note that the simplicity of these methods is important in the context of this discussion: they require no assumptions about the DGP of the problem to be made and can thus be applied universally.

Real Series: The U.S. GDP
An interesting series, for which the current methodology is relevant, is that of the United States real Gross Domestic Product (GDP - series GDPC96 from the Federal Reserve Bank of St. Louis online database). This analysis includes the last 200 available quarters for the years 1958 to 2008.

This series is the main economic indicator for the United States and from it the so-called output gap and the growth rate of the economy is computed. The logarithm of this series is plotted in Figure 3 which shows its salient characteristics, namely that it appears to be quite smooth and that it contains a trend component, which corresponds to the long-run (low frequency) movement of the economy.

Considerable literature exists in economics related to which type of stochastic process is best suited for describing the series.

### Table 3: Average Absolute Deviation of True from Estimated Trend & Optimal Degrees of Smoothing; Simulations from DGP II and Sample Sizes $n = \{200, 400\}$

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SSA-full</th>
<th>SSA-sub</th>
<th>SS-GCV</th>
<th>KF-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^\kappa$</td>
<td>0.24</td>
<td>0.22</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
<td>0.20</td>
<td>0.32</td>
<td>0.21</td>
</tr>
<tr>
<td>$k^\gamma$</td>
<td>9</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>0.56</td>
<td>n.a.</td>
<td>0.12</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SSA-full</th>
<th>SSA-sub</th>
<th>SS-GCV</th>
<th>KF-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^\kappa$</td>
<td>0.25</td>
<td>0.24</td>
<td>0.22</td>
<td>0.21</td>
<td>0.20</td>
<td>0.20</td>
<td>0.52</td>
<td>0.22</td>
</tr>
<tr>
<td>$k^\gamma$</td>
<td>11</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>0.42</td>
<td>n.a.</td>
<td>0.19</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SSA-full</th>
<th>SSA-sub</th>
<th>SS-GCV</th>
<th>KF-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^\kappa$</td>
<td>0.33</td>
<td>0.32</td>
<td>0.36</td>
<td>0.34</td>
<td>0.30</td>
<td>0.30</td>
<td>0.37</td>
<td>0.30</td>
</tr>
<tr>
<td>$k^\gamma$</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>0.75</td>
<td>n.a.</td>
<td>0.13</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoother</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SSA-full</th>
<th>SSA-sub</th>
<th>SS-GCV</th>
<th>KF-full</th>
<th>LL-sub</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m^\kappa$</td>
<td>0.33</td>
<td>0.31</td>
<td>0.35</td>
<td>0.32</td>
<td>0.30</td>
<td>0.29</td>
<td>0.52</td>
<td>0.30</td>
</tr>
<tr>
<td>$k^\gamma$</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>0.59</td>
<td>n.a.</td>
<td>0.18</td>
<td>0.03</td>
</tr>
</tbody>
</table>
However, no claims as to which process is indeed appropriate are put forth herein. Despite the visual proximity, it is not clear if a global deterministic trend is observed or a particular manifestation of a stochastic trend $g_t(\alpha_t, S_t)$ with structural changes. No definite answer has emerged from the related literature but the consensus agrees that a deterministic linear trend will be a poor approximation both because its shape does not agree with the underlying economic intuition and because it is not expected that such a global structure will remain stable over long periods of time. Therefore alternative ways of extracting the trend component by filtering or smoothing must be considered.

The most popular smoother, in this and related macroeconomic contexts is the Hodrick and Prescott (1997) or HP smoother. Note that this smoother is only optimal under specific conditions for the DGP (see for example Dermoune, et al., 2007). Nevertheless, it is so frequently used that its merits as an accurate representation of the DGP are not further discussed. The performance of the HP smoother and its degree of smoothing being selected by various methods are compared with the performance of the other smoothers we considered previously.

The potential differences from the application of different smoothing methods in the GDP series can only be assessed indirectly because there is no true trend component with which to compare results. Thus, the residuals after smoothing - the output gap - are considered as the variable of interest on which performance comparisons can be made.

The full and sub-sampling approaches have been applied to the MA, SSA and HP smoothers. In addition, the GCV-based smoothing splines were considered along with the plug-in based local linear smoother and the HP filter with an optimally selected value for the degree of smoothness (Dermoune, et al., 2007).

Denote by $u_{t}^{k,j}$ the residuals obtained
from the $j$th method and by $u_{i, \text{HP-opt}}^k$ the residuals obtained using the HP smoother with an optimally selected degree of smoothing. For each of these series we report their sample standard deviation and a Kolmogorov-Smirnov type test for the differences in the empirical cumulative distribution between $u_{i,j}^k$ and $u_{i, \text{HP-opt}}^k$. To compute the latter test the following steps are used:

1. Compute the empirical distributions of $u_{i,j}^k$ and $u_{i, \text{HP-opt}}^k$, for example, $F_n^j(u)$ and $F_n^{\text{HP-opt}}(u)$, over a grid of values, for example, $u \in U \subset R$.

2. Compute the Kolmogorov-Smirnov statistic $D_n = \sup_{u \in U} \sqrt{n} | F_n^j(u) - F_n^{\text{HP-opt}}(u) |$ for testing the equality of the underlying distributions.

3. Obtain an appropriate critical value for the test in the above step using the bootstrap – the stationary bootstrap (see Politis & Romano, 1994) was used in this study.

A number of interesting results are summarized and can be read from Table 4. Immediately it is observed that the hypothesis of equal distributions for the output gap between the HP smoother and all the other smoothers is not rejected. Therefore, in terms of the distribution of the residuals, all smoothers are essentially equivalent.

In addition there are a number of other interesting results that can be deduced from Table 4. First, note that the standard deviation of the residuals for the HP-based methods is practically the same irrespective of whether one uses the optimally selected degree of smoothing, as in Dermoune, et al. (2007), or uses the full or the sub-sampling methodology proposed herein.

Second, the MA and SSA smoothers produce residuals with larger standard deviation than the previous HP smoothers but which are on par with the standard deviation of the residuals obtained when the HP smoother is applied with the default degree of smoothing (equal to 1,600) as originally recommended by Hodrick and Prescott (1997). That value of the standard deviation was found to be 0.015. Finally, as shown in Figures 4, 5 and 6, the smoothers can be clustered together based on the standard deviations of their residuals to visualize their similarities and differences.

In Figure 4 the residuals from the three HP smoothers are plotted as in Table 4 plus the GCV-based smoothing spline smoother; it may be observed that the series are practically identical and this lends considerable support to the methodology proposed in this article as the residual series of the optimal HP smoother is able to be reproduced using both the full and sub-sampling approaches in minimizing the mean residual sign.

In Figure 5 the residuals from the MA (full and sub-sampling), the SSA (full only) and the default HP smoothers are plotted. Again a remarkable degree of closeness in the shape and magnitude of the four series is observed, especially of the moving average with sub-sampling and the default HP smoother.

Finally, Figure 6 plots the residuals from the singular spectrum analysis smoother with sub-sampling and the local linear smoother with plug-in bandwidth and, again, the series look practically identical.

<table>
<thead>
<tr>
<th>Smoother</th>
<th>HP-opt</th>
<th>HP-full</th>
<th>HP-sub</th>
<th>MA-full</th>
<th>MA-sub</th>
<th>SSA-full</th>
<th>SSA-sub</th>
<th>SS-GCV</th>
<th>LL-plug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SD$ of Residuals</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.018</td>
<td>0.015</td>
<td>0.018</td>
<td>0.007</td>
<td>0.003</td>
<td>0.006</td>
</tr>
<tr>
<td>$p$-value</td>
<td>n.a.</td>
<td>0.922</td>
<td>0.822</td>
<td>0.962</td>
<td>0.902</td>
<td>0.972</td>
<td>0.717</td>
<td>0.800</td>
<td>0.825</td>
</tr>
</tbody>
</table>
Figure 4: Residuals from Three HP Smoothers and the Smoothing Spline

Figure 5: Residuals from MA, SSA and Default HP Smoothers
It is evident from the figures that both similarities and differences exist among the smoothers and this is due to both their underlying filters and to the way the optimal degree of smoothing is selected. To explain the results consider the fact that the HP smoother is the optimal smoother for a stochastic process that is stationary in second differences. Therefore its application will necessarily lead to excess differencing if the true DGP becomes stationary after first differencing. Because the first differences of the GDP series are probably stationary (see Figure 7), then the HP smoother will remove a broader band of frequency components than the one corresponding to the trend of the series. The same holds true for the GCV-based smoothing splines smoother. To visualize this observe the shapes of the series in Figures 4 to 6; it can also be judged from the shapes of their corresponding autocorrelation or spectral density functions.

In Figures 8, 9 and 10 the spectral densities of the series that correspond to Figures 4, 5 and 6 are presented, these figures reinforce the idea of optimal smoothing for trend extraction. In all plots in these two figures the spectral densities have a single clear peak at frequencies corresponding to about 20 quarters (Figure 9) and 12 quarters (Figure 10) respectively. Both of these numbers fall within the range of the business cycles frequencies noted above. In fact, the peak of 12 quarters obtained by the smoothers in Figure 10 is almost the mid-range of the business cycles frequencies. Either the SSA smoother with sub-sampling or
Figure 7: Autocorrelation Function of Quarterly U.S. Real GDP Growth

Figure 8: Spectral Densities of Residual Series of Figure 4
Figure 9: Spectral Densities of Residual Series of Figure 5

Figure 10: Spectral Densities of Residual Series of Figure 6
the local linear smoother with the plug-in bandwidth appear to be a reasonable, economically viable compromise as those smoothers that capture the essence of the trend in U.S. output.

Based on the above discussion findings from this study may be summarized as follows:
• The proposed methodology can be used to achieve the same degree of smoothing for the HP smoother as that implied by other, more sophisticated, approaches.
• A number of alternative smoothers lead to the same shape and properties for the output gap as the HP smoother and these smoothers can be clustered together based on the shape of the series and their corresponding spectral densities.
Combining MA or SSA smoothers with subsampling leads to essentially the same results as the ones obtained by the default HP smoother.

Analyses herein illustrate a high potential for the application of less sophisticated, universally applicable, smoothing methods in trend extraction. This article proposes a simple, intuitive and immediately applicable method for selecting the degree of smoothing for such methods. One of the advantages of the having such methods available is that they can be used for benchmarks against which other, more sophisticated methods, can be compared.

Combining MA or SSA smoothers with subsampling leads to essentially the same results as the ones obtained by the default HP smoother.

Conclusions
This article proposed a new methodology for selecting the degree of smoothing in problems of trend extraction. The method uses an alternative to, mean squared error, proximity criterion which is minimized for selecting the required value of the smoothing parameter. This criterion is based on the average sign of the residuals obtained after smoothing and its minimization implies a probabilistic symmetrization of the residuals: it was shown that the minimizing value implies that the resulting residuals have a zero median.

The viability and usefulness of the proposed method is illustrated using simulations where the underlying type of trend is known. The results from these simulations are suggestive that the method is competitive because it can perform on par with - or better than - existing methods. In particular, it was shown that less sophisticated smoothing methods, like the moving average, for which no formal method for selecting the degree of smoothing exist, can be made to perform on par with more sophisticated methods. The use of sub-sampling can also help in improving performance.

A number of extensions can be undertaken based on the current work include the following:
• Consider the construction of confidence bands around the trend; since the method of this paper results in residuals with zero median such confidence bands can be based on the quantiles of the residual distribution.
• Consider a more systematic, expanded comparison between smoothing methods and approaches for selecting the optimal degree of smoothing.
• Apply the method of this article in the context of non-parametric autoregressive models and examine whether it can successfully be used in selecting both the degree of smoothing and the order of the model.

References


Can Specification Searches Be Useful for Hypothesis Generation?

Samuel B. Green  Marilyn S. Thompson
Arizona State University

Previous studies suggest that results from specification searches, as typically employed in structural equation modeling, should not be used to reach strong research conclusions due to their poor reliability. Analyses of computer generated data indicate that search results can be sufficiently reliable for exploratory purposes with properly designed and analyzed studies.

Key words: Structural equation modeling, specification searches, Lagrange multiplier test, modification indices.

Introduction

In specification searches, researchers seek to modify their hypothesized model by freeing or imposing model constraints. Particular constraints on model parameters are freed to maximize improvement in model fit or imposed to yield a more parsimonious model while minimizing loss of fit. The emphasis in searches is generally on freeing model constraints because researchers seek a better fitting model after determining their hypothesized model failed to closely fit the data. Accordingly, we focus our study on specification searches that relax model constraints.

Researchers may revise a model as a result of a specification search and, upon achieving good fit, describe strong conclusions about an obtained model or portray changes to the model as hypothesis generation. A specification search may be regarded as leading to strong conclusions if the results are used to create a single model and presented as a validation of that model. In contrast, a specification search is conducted for exploratory purposes if the results are used to generate one or more models that are presented as potential alternatives to the originally postulated model(s), and these alternative models are described as requiring validation in future research.

Search methods used to reach strong conclusions should have to meet very stringent psychometric standards. By contrast, standards for hypothesis generation might be more relaxed, but should be sufficiently rigorous to prevent researchers from wasting time and energy investigating models based on non-replicable specification searches. Previous research suggests that search procedures are inadequate to reach strong conclusions (MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992; Silvia & MacCallum, 1988). However, it is unclear whether searches can be useful for hypothesis generation in that standards have not been used that are consistent with this objective. The purpose of the study is to evaluate whether specification search methods can yield sufficiently accurate results to be used for the purpose of hypothesis generation.

This study considers only sequential specification searches, those that relax constraints one at a time, rather than nonsequential searches, such as Tabu, which attempt to determine combinations of parameters that would maximize model fit (e.g., Marcoulides, Drezner, & Schumacker, 1998). Sequential search methods are used almost exclusively in practice. In addition, this study evaluates searches that involve only adding
parameters to models that were constrained in
the original model to be equal zero. Thus, search
methods in this paper are discussed in the
context of adding parameters to models rather
than relaxing linear constraints in general.

Specification Search Methods

Researchers must define implicitly or
explicitly a search family of parameters that
potentially could be added to an initial model.
Parameters should be included in a search family
only if they are interpretable within the context
of the research study. Specification searches that
add parameters from the search family involve
forward selection, backward selection, or a
combination of forward and backward selection.

Forward Selection

The most popular forward selection
methods are based on the Lagrange multiplier
(LM) test (Chou & Bentler, 1990; Jöreskog &
Sörbom, 1989; Sörbom, 1989). The LM test
statistic evaluates whether one or more
constraints imposed on parameters should be
maintained and is asymptotically distributed as a
\( \chi^2 \), with degrees of freedom equal to the number
of constraints being evaluated. The LM test is
also called a modification index if it is used to
evaluate a single constraint.

Two sequential approaches are generally
available that use the LM test: the LM-
respecified and the LM-incremental
method. In the first step of the LM-respecified
method, the parameter in the search family with
the largest univariate LM \( \chi^2 \) statistic is selected.
At the second step, the model is respecified to
include this parameter, and then, among the
remaining parameters in the search family, the
parameter is selected with the largest univariate
LM \( \chi^2 \) statistic. At the third step, the model is
respecified to include the parameter selected at
the previous step, and the process continues. At
any step, the search stops when the \( p \) value for
the largest LM \( \chi^2 \) statistic is greater than the
prescribed alpha (e.g., .05).

The LM-incremental method is similar
to the LM-respecified method in that the parameter from the search family is added at
each step that maximally increases model fit
according to univariate LM tests. However, with
this approach, the univariate LM \( \chi^2 \) statistics are
incremental and are obtained by partitioning a
multivariate LM \( \chi^2 \) statistic into single-df \( \chi^2 \)
statistics. At step 1, the LM-incremental and the
LM-respecified methods are identical. However,
at step 2, the model is not respecified. Instead,
multivariate LM \( \chi^2 \) statistics are computed for
the addition of two parameters to the model: the
parameter selected at step 1 plus each of the
remaining parameters in the search family. An
incremental univariate \( \chi^2 \) statistic can now be
computed at step 2 for each of these remaining
parameters: the multivariate LM \( \chi^2 \) statistic for
the parameter selected at step 1 plus a remaining
parameter in the search family minus the largest
LM \( \chi^2 \) statistic from step 1. The parameter is
selected with the largest univariate incremental
LM \( \chi^2 \) statistic. The process continues until the \( p \)
value for the largest LM univariate incremental
\( \chi^2 \) statistic is greater than a prescribed alpha
(e.g., .05).

Backward Selection

Alternatively, stepwise backward
approaches may be applied using the Wald test
(Bentler, 1995; Chou & Bentler, 2002; Satorra,
1989), which is asymptotically distributed as a
\( \chi^2 \). With these approaches, all parameters in the
search family are added to a model at the
beginning of the search process. Then,
parameters in the search family are deleted one
at a time such that loss of model fit is
minimized.

Two backward selection methods are the
Wald-respecified and the Wald-incremental
methods. The distinctions between these
approaches are similar to those between the LM-
respecified method and the LM-incremental
method. With the Wald-respecified method, the
parameter with the smallest Wald \( \chi^2 \) statistic is
selected at each step, and then prior to the next
step, the model is respecified to exclude the
selected parameter. With the Wald-incremental
method, the model does not have to be
respecified at each step in that univariate Wald
tests are obtained by partitioning a multivariate
\( \chi^2 \) statistic into single-df \( \chi^2 \) statistics.

Combination Forward-Backward Selection

Other sequential procedures might
involve both forward and backward searches
(Green, Thompson, & Poirier, 1999). Analogous
to stepwise regression analysis, parameters from the search family could be added and deleted at each step in the search process. However, such an approach might be considered too cumbersome by researchers unless automated by a software package. Alternatively, a two-stage search process could be employed. In the first stage, parameters in the search family are sequentially added based on the LM test; then, in the second stage, the added parameters from the first stage are sequentially deleted based on the Wald test. A two-stage search process could be used as an alternative to a backward search approach if the latter approach is not possible because the model is underidentified when all parameters in the search family are added to the model before backward deletion.

Methods to Minimize Errors in Specification Searches

Traditionally $\chi^2$ statistics for individual LM and Wald tests have been evaluated at the .05 level in sequential searches, disregarding the number of conducted tests. Green, Thompson, and Babyak (1998) and Hancock (1999) have offered methods for controlling for Type I errors with multiple LM tests, while Green, Thompson, and Poirer (2001) have suggested a method with Wald tests.

The suggested methods are adaptations of either the Roy union-intersection method (Roy, 1953) or Holm’s sequential Bonferroni method (Holm, 1979). However, controlling the probability of committing Type I errors across tests is problematic in that it increases the probability of committing Type II errors, failure to add parameters to the model that should be included. Consequently, methods for controlling Type I error rates are more likely to produce accurate results if large samples are employed to minimize Type II error rates.

Errors due to misspecification occur if a parameter improves model fit at a step in a search process, but would fail to improve fit if the model had been correctly specified. This type of error occurs even when a search is conducted on the population and, therefore, is distinguishable from an error due to sampling fluctuation. Based on past research, it is known that misspecification errors are less likely to occur if researchers have a relatively well-specified hypothesized model, have few irrelevant and many relevant parameters in their search family, and have large samples (MacCallum, 1986; MacCallum, Roznowski, Necowitz, 1992; Silvia & MacCallum, 1988). Accordingly, researchers must carefully construct not only the hypothesized model, but also the search family, based on best available theory and a thorough understanding of the empirical research literature. Errors due to misspecification also are less likely to occur with large samples in that the search process is more likely to avoid errors due to sampling fluctuation and thus yield better specified models. In addition, choice of search methods may have an effect on the likelihood of committing misspecification errors. Forward sequential search methods are by far most popular; however, backward sequential methods might yield better results (Green, Thompson, & Poirer, 2001; Chou & Bentler, 2002) in that the model is initially respecified to include all parameters in the search, reducing the likelihood of misspecification errors.

Purpose of the Study

MacCallum, Roznowski, and Necowitz (1992) demonstrated convincingly that forward search methods are not sufficiently replicable to be useful for reaching strong conclusions. As part of a larger analysis, they conducted a number of searches on data collected from 3,694 research participants. They initially determined the first four parameters that were added to a model for the total sample. For the purpose of their analyses, the total sample could be considered the population. Next, for 8 different sample sizes, they determined the parameters added in the first 4 steps of a specification search for 10 replicate samples drawn from the total sample. They reported the added parameters for all searches in tabular form. For each sample size, the percent of times the same 4 modifications were made in the 10 replicate samples drawn from the total sample. They reported the added parameters for all searches in tabular form. For each sample size, the percent of times the same 4 modifications were made in the 10 replicate samples and the total sample was presented. As shown in Table 1, zero percent of the searches with $N$ of 100, 150, 200, and 250 yielded the same 4 modifications as were made in the total sample. Even for the largest sample size ($N = 1,200$), only 60% of the searches produced the same 4 modifications.
Based on these results and others, methodologists tend to view specification searches skeptically (e.g., Boosma, 2000; Breckler, 1990; MacCallum, et al., 1992). MacCallum (1995) offered the following advice about searches:

…researchers must be concerned about use of the model generation strategy in practice. Users of this strategy must acknowledge that they are engaging in exploratory model development. There is not necessarily anything wrong with exploratory model development as long as it is acknowledged in practice that that is what is being done and that the outcome is a model that cannot be supported without being evaluated using new data. Serious problems arise when the model generation strategy is applied without any effort to attach substantive meaning to model modifications and when the resulting model is treated as if it has been confirmed because it fits the observed sample data well. The model generation strategy is a legitimate approach to model development if it is used responsibly, but such use seems to be the exception rather than the rule in much of the applied literature. (p. 34)

Searches should not have to meet as stringent criteria to be used for hypothesis generation, but nevertheless should demonstrate adequate psychometric quality. Based on the results presented by MacCallum, et al. (1992), we computed match statistics using a less stringent criterion. Matching was assessed for each selected parameter in a replicate sample rather than for the set of all four selected parameters. In other words, a match occurred if a parameter selected in a sample was the same as one of the four parameters selected in the total sample. For any sample size, the maximum number of possible hits using this definition of a match is 40 (= 4 parameters x 10 replicates) rather than 10.

As shown in Table 1, the hit percentages were 30, 38, 40, and 55 for sample sizes of 100, 150, 200, and 250, respectively. The percent of hits continued to increase from 60 to 82 as sample size increased from 325 to 1,200. These results suggest that specification searches may be insufficiently accurate even for exploratory analyses at smaller sample sizes (perhaps less than a sample size of 250 based on these results), but acceptable for this purpose at larger sample sizes.

Table 1: Summary of Results from MacCallum, et al. (1992) Study: Match Statistics between Parameters Selected in the 10 Replicate Samples and the First Four Selected Parameters in the Total Sample

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Percent of 4:4 All Matches</th>
<th>Mean Percent of 4:4 Any Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>150</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>200</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>250</td>
<td>0</td>
<td>55</td>
</tr>
<tr>
<td>325</td>
<td>10</td>
<td>60</td>
</tr>
<tr>
<td>400</td>
<td>20</td>
<td>65</td>
</tr>
<tr>
<td>800</td>
<td>40</td>
<td>68</td>
</tr>
<tr>
<td>1,200</td>
<td>60</td>
<td>82</td>
</tr>
</tbody>
</table>

*aAll 4 parameters selected in sample match all 4 parameters selected in population. These percentages were presented by MacCallum, et al. (1992); bAny of 4 parameters selected in sample match any of 4 parameters selected in population. We calculated these mean percentages based on the results of the searches reported by MacCallum, et al. (1992).
reaching strong conclusions (MacCallum, 1986; MacCallum, Roznowski, Necowitz, 1992; Silvia & MacCallum, 1988). However, it is less clear whether searches can be used to meet exploratory goals. This study analyzed the covariance matrix examined by MacCallum, et al. (1992) in their classic study of specification searches and, by drawing samples based on this covariance matrix, investigated whether search results can be sufficiently accurate to warrant their use for hypothesis generation.

The strength of the approach in this study is that the examined covariance matrix was based on real data, and this matrix was investigated in a well known study that led to negative conclusions about specification searches. Given these negative results, it is important to establish that specification search methods can be useful for exploratory purposes. If this can be established, then researchers might be encouraged to conduct further research on these popular search methods, even though this research requires methods for tracking all possible additions to models and thus is time-consuming.

A number of authors (Bentler & Chou, 1993; Chou & Bentler, 1993; Kaplan, 1990; Sörbom, 1989) have convincingly argued that researchers should evaluate not only the $\chi^2$ values associated with tests in searches, but also statistics that assess expected change in a parameter when that parameter is freed. In this paper we focus on the $\chi^2$ values and do not attempt to address the broader and more complex issue of combining the results of significance tests with the expected change parameter statistic.

Methodology
Initial Model and Data
Searches were conducted using the hypothesized model and covariance matrix examined by MacCallum, et al. (1992) in their study of specification searches. The heuristic model (presented in Figure 1) is of employee responses to affect (Hulin, Roznowski, & Hachiya, 1985) and includes 21 measured variables and 7 factors. The data are based on a questionnaire administered to 3,694 employees from two hospitals (see MacCallum, et al. (1992) for a more detailed description of the model and the data.) The raw data were not available. Consequently, the covariance matrix was treated as the population covariance matrix, and samples were generated based on this matrix using the multivariate normal generator available in EQS (Bentler, 1995).

One thousand replicate samples were generated for each of four sample sizes: 200, 500, 800 and 1,200. If difficulties emerged in the estimation process (e.g., iterative process failed to converge; parameter estimates were out of bounds), additional replicate samples were generated to yield 1,000 replicates. The sample sizes were similar to those explored by MacCallum et al. (1992), which varied in size from 100 to 1,200. Sample sizes less than 200 were excluded because past research indicates that these sample sizes are inadequate for many applications of SEM (Tomarken & Waller, 2005) and, in particular, have been shown to be insufficient for specification searches (MacCallum, 1986; MacCallum et al., 1992; Silvia & MacCallum, 1988).

Forward Searches
The LM-incremental method was used to conduct the searches. This method was chosen for four reasons: (a) the LM-incremental method is available within EQS (Bentler, 1995) and, accordingly, is presumably a relatively popular approach; (b) forward search procedures, such as the LM-incremental method, are most frequently applied and, in this sense, are more interesting to explore; (c) the LM-incremental method is efficient in that it does not require model respecification; and (d) the belief that the replicability of search results using the LM-incremental test should be similar to those using the LM-respecified method, the alternative forward approach for conducting specification searches.

For each sample size, two sizes of search families were considered to assess whether a decrease in family size would increase the replicability of specification search results. The small family contained 69 parameters. These parameters included (a) 16 paths between the exogenous factors $F_1$ through $F_3$ and their indicators $V_1$ through $V_8$, (b) 39 paths between the endogenous factors $F_4$ through $F_7$ and their indicators $V_9$ through $V_{21}$, (c) 5 paths between
the exogenous factor $F_1$ through $F_3$ and the endogenous factors $F_4$ through $F_7$, and (d) 9 paths among the endogenous factors $F_4$ through $F_7$, excluding the path from $F_4$ to $F_5$ to avoid an underidentified model if selected. Covariances among residuals were ignored in the small family because they are likely in many applications to be included in models to improve model fit without conceptual justification (Hoyle & Panter, 1995). The large family contained 178 parameters and included (a) the parameters in the small family, (b) 27 covariances among the indicator errors $E_1$ through $E_8$, (c) 78 covariances among the indicator errors $E_9$.

Factors 1, 2, and 3 are work satisfaction, pay satisfaction, and perceptions about physical working conditions. Factors 4, 5, 6, and 7 are change to improve work conditions, citizenship such that individuals volunteer and display extra-role behaviors in the work place, psychological or passive withdrawal of individuals from the workplace, and thoughts and intentions about physical withdrawal from the organization in the future.
through $E_{21}$, and (d) 4 covariances among disturbances $D_4$ through $D_7$ associated with the endogenous factors. The covariance between $E_7$ and $E_8$, the covariance between $D_4$ and $D_5$, and the covariance between $D_6$ and $D_7$ were not part of the family in that freeing these parameters would have produced an underidentified model. The large family is similar in size to the one used by MacCallum, et al. (1992) in their searches.

MacCallum, et al. (1992) reported results for the replicability of the first four parameters added to the initial model based on searches. For their searches, all added parameters were significant at the .01 level or lower. This study examined both the first four parameters as well as the first eight parameters added to the model in sample data. At any step within a search, a parameter was not added to the model and the search was terminated if the parameter was not significant at the .05 level. The first four added parameters always were significant at the .05 and, most often, at a much lower value. For the next four added parameters, the search was terminated for some replicate samples with a sample size of 200 and the small family size. For this combination of conditions, the search was discontinued 37 times out of 1,000 replicate samples: 1 time at step 6, 7 times at step 7, and 29 times at step 8.

Assessment of Replicability

Replicability was assessed by computing the extent to which the results of the LM-incremental method in the sample matched those in the population. An average match rate across the 1,000 replications was computed, with four different definitions for a match.

In order to reach strong conclusions based on specification searches, it would be ideal if all parameters selected to be added to the model in the sample to match those found in the population. To assess the accuracy of searches for this purpose, a stringent criterion was defined: the 4:4 All Match. For this criterion, all four parameters selected first in the sample had to be the same as all four parameters selected first in the population. For any one replication, the selected parameters either matched or failed to match. The percent of 4:4 all matches was computed across the 1,000 replications. This definition of percent of matches is the same as the one presented by MacCallum, et al. (1992).

If search methods are not used to reach strong conclusions, but rather to generate hypotheses, less stringent criteria can be used for matches. Three less stringent criteria were defined as follows:

- **4:4 Any Match**: Any 1 parameter of the first 4 parameters selected in the sample matches any 1 parameter of the first 4 parameters selected in the population. The maximum number of matches for a replication was 4 and occurred if all 4 of the sample parameters matched all 4 of parameters selected in the population.

- **8:4 Any Match**: Any 1 parameter of the first 8 parameters selected in the sample matches any 1 parameter of the first 4 parameters selected in the population. The maximum number of matches for a replication was 4 and occurred if any 4 of the 8 parameters selected in the sample matched the first 4 parameters selected in the population.

- **4:12 Any Match**: Any 1 parameter of the first 4 parameters selected in the sample matches any 1 parameter of the first 12 parameters selected in the population. The maximum number of matches for a replication was 4 and occurred if the 4 parameters selected in the sample matches any 4 of the 12 parameters selected in the population.

For each of these three criteria, the percent of matches was computed for each replication (0%, 25%, 50%, 75%, or 100%), and the mean percent was computed across the 1,000 replications.

From a hypothesis generating perspective, the 4:4 any match (as well as the 4:4 all match) might be considered too stringent in comparison with the 8:4 any match. Researchers can choose to conduct more steps in the search process than the number of parameters that they actually add to their model. For example, they might continue the search process through the first 8 steps rather than the first 4 steps. By conducting this deeper search, they are more likely to find the first four parameters added in the population. To the
extent that the mean percentage for 8:4 any matches exceeds the mean percentage for 4:4 any matches, deeper searches are recommended. The 8:4 any match is consistent with searches for hypothesis generation in that researchers are not likely to include all new parameters indicated by a search. Even if parameters are included in a search family only if they are conceptually meaningful, some combinations of parameters are more meaningful than others and thus make for better generated hypotheses.

Three of the four matching criteria assessed matches to the first four parameters added to the model at the population level. However, the choice of the number of added parameters was arbitrary. Even with 12 parameters added to the model, the fit was not perfect in the population. Accordingly, a criterion was included in the analyses matching sample searches to the first 12 parameters added to the model at the population level.

Search Results at the Population Level

An appropriate initial model for this study should not fit so well that a search is unnecessary, but not so badly that a search would be fruitless because the model is severely misspecified. For the original sample data explored by MacCallum, et al. (1992; \(N = 3,694\)), the model (i.e., Figure 1) fit the data adequately, but not as close as desired by many researchers, \(\chi^2 (177) = 3215.44, p < .01, CFI = .89, RMSEA = .068\) (90% CI of .066 to .070). The CFI is less than either the traditional cutoff of .90 or the cutoff of .95 recommended by Hu and Bentler (1999), while the RMSEA is slightly greater than the cutoff of .06 (Hu & Bentler, 1999). Consequently, a search would be considered by many researchers.

The model fit much better after the first four parameters were added based on LM-incremental method for either the small or the large search family (same four added parameters for both families), \(\chi^2 (173) = 1825.06, p < .01, CFI = .94, RMSEA = .051\) (90% CI of .049 to .053); the model shows further improvement in fit if more parameters are added. Model fit improved when 4 more parameters were added to the model based on the LM-incremental method for the small search family (total of 8 added parameters): \(\chi^2 (165) = 1396.68, p < .01, CFI = .96, RMSEA = .044\) (90% CI of .042 to .047). It increased even further with 4 more added parameters (total of 12 added parameters): \(\chi^2 (165) = 1177.21, p < .01, CFI = .96, RMSEA = .041\) (90% CI of .039 to .043).

To the extent that the use of the LM-incremental and the LM-respecified methods yielded very different sets of added parameters, there is less confidence that the results of this study based on the LM-incremental method would generalize to those using the LM-respecified method. Of course, even if the results were identical, one could not be confident that the findings of this study would generalize to those based on the LM-respecification method. As shown in Table 2, the two methods produced similar – but not identical – results at the population level for both the small and large search families.

Results

Percentages Based on 4:4 All Matches

As shown in Table 3, the percentages for 4:4 all matches were uniformly low for the large search family, with the percentages varying from 6% for a sample size of 200 to 39% for a sample size of 1,200. The percentages for 4:4 all matches were higher for the small search family, but did not exceed 50% for sample sizes of 500 or less. The percentages approached 80% for only the largest sample size of 1,200.

Percentages Based on Alternative Definitions of Matches

The percentages improved considerably when a match did not require all of the first four parameters in the sample to match all of the first four parameters in the population. For 4:4 any matches, the means for the percent of matches for the small search family ranged from 65% for an \(N\) of 200 to 94% for an \(N\) of 1,200. The means for the percent of matches were 12 to 13 percentage points lower for the large search family, ranging from 52% for an \(N\) of 200 to 81% for an \(N\) of 1,200.

For more liberal definitions of a match, the match percentages were higher under comparable conditions. The means for the percent of 8:4 any matches always exceeded 90% except with \(Ns\) of 200. The means for the
USEFULNESS OF SEARCHES

Table 2: Ranks of Entry into Model Based on Incremental LM Test and LM Test with Model Respecification in the Population

<table>
<thead>
<tr>
<th>Search Family Size of 69 Parameters</th>
<th>Search Family Size of 178 Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>V17, F5</td>
<td>1</td>
</tr>
<tr>
<td>V19, F6</td>
<td>2</td>
</tr>
<tr>
<td>V20, F6</td>
<td>3</td>
</tr>
<tr>
<td>V11, F7</td>
<td>4</td>
</tr>
<tr>
<td>V3, F3</td>
<td>5</td>
</tr>
<tr>
<td>F5, F7</td>
<td>6</td>
</tr>
<tr>
<td>V12, F4</td>
<td>7</td>
</tr>
<tr>
<td>V20, F4</td>
<td>8</td>
</tr>
<tr>
<td>V15, F5</td>
<td>9</td>
</tr>
<tr>
<td>V16, F7</td>
<td>10</td>
</tr>
</tbody>
</table>

Note: For path parameters, the second variable affects the first variable. Parameters are shown if they were selected in the first 10 steps using the LM-incremental method. Ranks are not presented if the rank of entry into the model exceeded 15 steps.

Table 3: Match Statistics between Parameters Selected in Sample and Population for Different Sizes of Samples and Search Families

<table>
<thead>
<tr>
<th>Mean Match Statistics</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Smaller Search Family Size: 69 Parameters</td>
<td></td>
</tr>
<tr>
<td>Percent of 4:4 all matches (All 4 parameters</td>
<td>17</td>
</tr>
<tr>
<td>selected in sample match all 4 parameters</td>
<td></td>
</tr>
<tr>
<td>selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Mean percent of 4:4 any matches (Any of 4</td>
<td>65</td>
</tr>
<tr>
<td>parameters selected in sample match any of</td>
<td></td>
</tr>
<tr>
<td>4 parameters selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Mean percent of 8:4 any matches (Any of 8</td>
<td>82</td>
</tr>
<tr>
<td>parameters selected in sample match any of</td>
<td></td>
</tr>
<tr>
<td>4 parameters selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Mean percent of 4:12 any matches (Any of 4</td>
<td>80</td>
</tr>
<tr>
<td>parameters selected in sample match any of</td>
<td></td>
</tr>
<tr>
<td>12 parameters selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Larger Search Family Size: 178 Parameters</td>
<td></td>
</tr>
<tr>
<td>Percent of 4:4 all matches (All 4 parameters</td>
<td>6</td>
</tr>
<tr>
<td>selected in sample match all 4 parameters</td>
<td></td>
</tr>
<tr>
<td>selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Mean percent of 4:4 any matches (Any of 4</td>
<td>52</td>
</tr>
<tr>
<td>parameters selected in sample match any of 4</td>
<td></td>
</tr>
<tr>
<td>parameters selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Mean percent of 8:4 any matches (Any of 8</td>
<td>69</td>
</tr>
<tr>
<td>parameters selected in sample match any of 4</td>
<td></td>
</tr>
<tr>
<td>parameters selected in population.)</td>
<td></td>
</tr>
<tr>
<td>Mean percent of 4:12 any matches (Any of 4</td>
<td>63</td>
</tr>
<tr>
<td>parameters selected in sample match any of</td>
<td></td>
</tr>
<tr>
<td>12 parameters selected in population.)</td>
<td></td>
</tr>
</tbody>
</table>
percent of 4:12 any matches were relatively high when the family size was small; that is, they exceeded 90% as long as $N$ was greater than 200. However, the means for the percent of 4:12 any matches were generally lower for the large family size, ranging from 63% to 86%.

Conclusion
The results based on matching all of the first four parameters added in the sample with all of the first four parameters added in the population (i.e., 4:4 all matches) are consistent with those of MacCallum, et al. (1994) and do not support the use of search methods for reaching strong conclusions unless the size of the search family is quite restricted and sample size is very large. For the model in this study, the percent of matches approached 80% only if the sample size was 1,200 and the family size was small. Conversely, the results based on more liberal match criteria support the use of search methods for exploratory purposes with moderately large sample sizes and a small search family. The mean of the percentages for 8:4 any matches and 4:12 any matches were generally satisfactory (i.e., approximately 80% or greater) if sample size was 500 or greater. The mean percentages were always higher for the smaller search family.

Results demonstrate that specification searches can produce replicable results for exploratory purposes. However, the reported percentages are specific to the model and the data set examined. Although the sample size of 500 was adequate in a number of conditions for the searches conducted in our study, smaller sample sizes might be satisfactory if the initial model fails to include a minimal number of relevant parameters and the search is conducted on a very restricted search family. By contrast, larger sample sizes might be required if the initial model excludes a large number of relevant parameters and the search family is very large. It is suggested that other studies using real data sets and generated data be conducted to assess the generalizability of our results.

The findings in this study are consistent with previous recommendations about conducting specification searches (Green et al., 1999; Kaplan, 1990; MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992; Schumacker & Lomax, 1996; Silvia & MacCallum, 1988). Researchers should (a) narrow the number of parameters in searches based on their conceptual understanding of the substantive area and the methods employed; (b) conduct searches on large samples; and (c) carry out deep searches in that parameters generated later in a search might be appropriate to add to an initial model. Researchers may be hesitant to carry out searches with small search families in that modified models resulting from searches are likely to yield better fit if more parameters are included in search families. However, this improvement in fit is likely to be illusory, resulting from overfitting a model to the peculiar characteristics of a specific sample; that is, the improved fit is unlikely to hold up in replicate samples.

Based on our results, specification searches can be appropriate for exploratory purposes if used judiciously and can be reported as potentially a valid method for hypothesis generation. In agreement with others who have written about specification searches (e.g., Boomsma, 2000; MacCallum, 1995), researchers should describe explicitly in their research publication their initial model, the search family, the search method, and the conceptual meaning of the added parameters so that readers can evaluate appropriately the meaningfulness of their results. In addition, they need to indicate the importance of validating in new samples the models that are generated through the searches. Because some researchers are hesitant to report the search process involved in conducting their SEM analyses, reviewers of manuscript in which SEM is applied should ask authors to describe their initial model and to delineate the search family and the search method employed to obtain their final model(s).

We suspect the results in this study would have been similar if we had investigated the LM-respecified method and linear constraints other than model parameters being equal to zero. However, future research should investigate this hypothesis. It would also be useful to assess other search methods (e.g., backward selection) as well as the effect of controlling for Type I error rate across multiple tests.
Acknowledgement
The authors would like to extend our appreciation to Dr. R. C. MacCallum for sharing the covariance matrix used in his 1992 study with us.

References


Measuring Openness

Gaetano Ferrieri
Studi Interdisciplinari, Italy

A method for measuring international openness is elaborated. This synthetic indicator measures the capacity of countries for a given phenomenon adjusted for their weight in the same phenomenon. The method implemented and applied to international trade and illustrated here as a case study in merchandise exports, has a wide range of applications in the socio-economic field.

Key words: Index, International Openness, International Trade.

Introduction

An innovative method for measuring openness as applied to international migration (Ferrieri, 2006) was recently proposed. In this article, the method is further implemented and applied to international trade (e.g., merchandise exports). The research focus is on the top world economies (by GDP share). Sample calculations based on empirical data and some simulations are provided in order to better understand the methodology and to evaluate the analytical properties of the index proposed.

Methodology

The method for measuring openness involves two steps. First, the values of the basic indicator (e.g., exports-to-GDP ratio), collected for a number of countries, are divided by their highest value in order to obtain index values on a scale referring to one. In this work, the benchmark is the maximum value at the current year. However, it is preferable to fix as benchmark the highest value observed over time in order to make time comparisons. In the second phase, the index values so normalized are adjusted for the weight of the same countries in the world aggregate considered (e.g., merchandise exports). The share (not in percentage terms) is expressed as distance from one (total observed or assumed) and is raised to a constant factor for all countries. This factor should take into account the importance (in terms of dispersion or variability) of the aggregate that figures as denominator of the basic indicator (e.g., the Gross Domestic Product).

The Index is formulated as follows:

\[
Index = \left( \frac{V_i}{V_{\text{MAX}}} \right)^{(1-\pi)^\kappa}
\]

in which: \(V_i\) is the value of the basic indicator (e.g., EXP–to-GDP ratio) for each country in a given year; \(V_{\text{MAX}}\) is the maximum value of the basic indicator (i.e. EXP–to-GDP ratio) across the countries; \(\pi\) is the share of each country in the world aggregate considered (e.g., merchandise exports) in the given year; \(\kappa\) is the constant factor for all countries (e.g., the coefficient of variation of the gross domestic product of the countries analysed, not expressed in percentage terms).

Thus, two different effects determine the Index value and can be defined as:

(1) Capacity effect, given by the expression \(\frac{V_i}{V_{\text{MAX}}}\), and
(2) Size effect, given by the expression \((1-\pi)^\kappa\), with the exponent \(\kappa = \text{constant}\).
The value of the Index is basically determined by the capacity effect, while the size effect produces a growth in value for all countries, and particularly for those with a high share in the phenomenon concerned.

It is not difficult to observe how the maximum value of the Index (unity) can be obtained not only in terms of best capacity but also in terms of best size. However, the probability of this latter happening is very remote and even in this hypothesis the Index is mathematically valid and consistent. In this hypothesis, the value of one could be achieved (absurdly) by monopolising the phenomenon analysed (country’s weight = 1). Mathematically, the overall result is one. In fact: 1−1=0; zero raised to any number (different from zero) is equal to zero; any number raised to an exponent of zero is equal to one.

Results

The Index applied to merchandise exports (conventionally defined IOEXP) is calculated for the top 40 economies by share in the world current gross domestic product (GDP). Data for elaboration refer to merchandise exports (F.O.B. valuation) and GDP expressed in US dollars at market exchange rates in 2004. Table 1 illustrates some sample calculations. The samples refer to Malaysia, Germany and the United States. Results for all countries are shown in Table 2.

The value of the EXP-to-GDP ratio of Malaysia is 106.92 percent, the highest value among the 40 countries analysed (see also Table 2). The value of the same indicator is equal to 33.12 percent for Germany and 6.98 percent for the United States that rank respectively 12th and 40th among the countries analysed. However,

Table 1: Index of Openness to (Merchandise) Exports (IOEXP). Sample Calculations: Malaysia, Germany and the United States*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Malaysia</th>
<th>Germany</th>
<th>United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exports-to-GDP ratio (Vi)</td>
<td>106.92 %</td>
<td>33.12 %</td>
<td>6.98 %</td>
</tr>
<tr>
<td>Share in world merchandise exports (π)</td>
<td>0.014</td>
<td>0.103</td>
<td>0.093</td>
</tr>
<tr>
<td>Constant = coefficient of variation of GDP (κ) (a)</td>
<td>2.04</td>
<td>2.04</td>
<td>2.04</td>
</tr>
<tr>
<td>Indicator value normalised (IVN = Vi / VMAX) (b)</td>
<td>1.000</td>
<td>0.310</td>
<td>0.065</td>
</tr>
<tr>
<td>IVN (or Vi) rank</td>
<td>1</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>Index of Openness (IOEXP)</td>
<td>1.000</td>
<td>0.391</td>
<td>0.107</td>
</tr>
<tr>
<td>Index (IOEXP) rank</td>
<td>1</td>
<td>9</td>
<td>39</td>
</tr>
<tr>
<td>Difference between IOEXP and IVN value</td>
<td>-</td>
<td>26.38 %</td>
<td>63.60 %</td>
</tr>
<tr>
<td>Difference between IOEXP and IVN rank</td>
<td>-</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>


(a) The coefficient of variation of GDP (κ) is here calculated over the 40 world economies analysed.

(b) VMAX is the maximum value of Vi across the countries analysed and is equal to 106.92% (Malaysia).
Table 2: Index of Openness to (Merchandise) Exports (IOEXP) in the Top 40 World Economies

<table>
<thead>
<tr>
<th>Country</th>
<th>EXP/GDP% Vi</th>
<th>Share in World EXP</th>
<th>Value</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>IVN IOEXP</td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>106.92</td>
<td>0.014</td>
<td>1.000 1.000</td>
<td>1 1</td>
</tr>
<tr>
<td>Singapore(a)</td>
<td>91.70</td>
<td>0.011</td>
<td>0.858 0.861</td>
<td>2 2</td>
</tr>
<tr>
<td>Belgium</td>
<td>85.75</td>
<td>0.035</td>
<td>0.802 0.814</td>
<td>3 3</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>63.75</td>
<td>0.008</td>
<td>0.596 0.601</td>
<td>4 4</td>
</tr>
<tr>
<td>Netherlands</td>
<td>58.96</td>
<td>0.041</td>
<td>0.551 0.579</td>
<td>5 5</td>
</tr>
<tr>
<td>Thailand</td>
<td>60.25</td>
<td>0.011</td>
<td>0.563 0.571</td>
<td>5 6</td>
</tr>
<tr>
<td>Ireland</td>
<td>56.46</td>
<td>0.012</td>
<td>0.528 0.536</td>
<td>7 7</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>50.31</td>
<td>0.014</td>
<td>0.471 0.481</td>
<td>8 8</td>
</tr>
<tr>
<td>Germany</td>
<td>33.12</td>
<td>0.103</td>
<td>0.310 0.391</td>
<td>12 9</td>
</tr>
<tr>
<td>Austria</td>
<td>39.84</td>
<td>0.013</td>
<td>0.373 0.383</td>
<td>9 10</td>
</tr>
<tr>
<td>Korea</td>
<td>37.35</td>
<td>0.029</td>
<td>0.349 0.371</td>
<td>10 11</td>
</tr>
<tr>
<td>China</td>
<td>30.72</td>
<td>0.067</td>
<td>0.287 0.339</td>
<td>21 12</td>
</tr>
<tr>
<td>Sweden</td>
<td>34.94</td>
<td>0.014</td>
<td>0.327 0.337</td>
<td>11 13</td>
</tr>
<tr>
<td>Canada</td>
<td>31.86</td>
<td>0.036</td>
<td>0.298 0.325</td>
<td>17 14</td>
</tr>
<tr>
<td>Switzerland</td>
<td>33.01</td>
<td>0.013</td>
<td>0.309 0.319</td>
<td>13 15</td>
</tr>
<tr>
<td>Finland</td>
<td>32.95</td>
<td>0.007</td>
<td>0.308 0.313</td>
<td>15 16</td>
</tr>
<tr>
<td>Israel</td>
<td>32.95</td>
<td>0.004</td>
<td>0.308 0.311</td>
<td>14 17</td>
</tr>
<tr>
<td>Norway</td>
<td>32.05</td>
<td>0.009</td>
<td>0.300 0.307</td>
<td>16 18</td>
</tr>
<tr>
<td>Russia</td>
<td>31.06</td>
<td>0.021</td>
<td>0.290 0.306</td>
<td>20 19</td>
</tr>
<tr>
<td>Denmark</td>
<td>31.33</td>
<td>0.009</td>
<td>0.293 0.299</td>
<td>18 20</td>
</tr>
<tr>
<td>Venezuela</td>
<td>31.07</td>
<td>0.004</td>
<td>0.291 0.293</td>
<td>19 21</td>
</tr>
<tr>
<td>Poland</td>
<td>29.66</td>
<td>0.008</td>
<td>0.277 0.284</td>
<td>22 22</td>
</tr>
<tr>
<td>Mexico</td>
<td>27.66</td>
<td>0.021</td>
<td>0.259 0.274</td>
<td>24 23</td>
</tr>
<tr>
<td>Indonesia</td>
<td>28.42</td>
<td>0.008</td>
<td>0.266 0.272</td>
<td>23 24</td>
</tr>
<tr>
<td>Iran, Islamic Republic of</td>
<td>27.52</td>
<td>0.005</td>
<td>0.257 0.261</td>
<td>25 25</td>
</tr>
<tr>
<td>France</td>
<td>21.94</td>
<td>0.051</td>
<td>0.205 0.241</td>
<td>27 26</td>
</tr>
<tr>
<td>Italy</td>
<td>20.24</td>
<td>0.040</td>
<td>0.189 0.216</td>
<td>30 27</td>
</tr>
<tr>
<td>Argentina</td>
<td>22.67</td>
<td>0.004</td>
<td>0.212 0.215</td>
<td>26 28</td>
</tr>
<tr>
<td>South Africa</td>
<td>21.41</td>
<td>0.005</td>
<td>0.200 0.204</td>
<td>28 29</td>
</tr>
<tr>
<td>Turkey</td>
<td>20.86</td>
<td>0.007</td>
<td>0.195 0.200</td>
<td>29 30</td>
</tr>
<tr>
<td>Portugal</td>
<td>20.14</td>
<td>0.004</td>
<td>0.188 0.191</td>
<td>31 31</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>16.26</td>
<td>0.039</td>
<td>0.152 0.176</td>
<td>33 32</td>
</tr>
<tr>
<td>Spain</td>
<td>17.16</td>
<td>0.020</td>
<td>0.160 0.173</td>
<td>32 33</td>
</tr>
<tr>
<td>Brazil</td>
<td>15.98</td>
<td>0.011</td>
<td>0.149 0.156</td>
<td>34 34</td>
</tr>
<tr>
<td>Japan</td>
<td>12.33</td>
<td>0.064</td>
<td>0.115 0.152</td>
<td>36 35</td>
</tr>
<tr>
<td>Australia</td>
<td>13.56</td>
<td>0.010</td>
<td>0.127 0.132</td>
<td>35 36</td>
</tr>
<tr>
<td>Hong Kong SAR(a)</td>
<td>12.05</td>
<td>0.002</td>
<td>0.113 0.114</td>
<td>37 37</td>
</tr>
<tr>
<td>India</td>
<td>11.35</td>
<td>0.009</td>
<td>0.106 0.110</td>
<td>38 38</td>
</tr>
<tr>
<td>United States</td>
<td>6.98</td>
<td>0.093</td>
<td>0.065 0.107</td>
<td>40 39</td>
</tr>
<tr>
<td>Greece</td>
<td>7.31</td>
<td>0.002</td>
<td>0.068 0.069</td>
<td>39 40</td>
</tr>
</tbody>
</table>


(a)Exports data for Hong Kong and Singapore do not include re-exports, but only refer to domestic exports. All figures (values, ranks) result from electronic calculations. Data for elaboration are drawn from WTO (2006) and IMF (2006).
Germany presents the highest share in the world merchandise exports (10.3 percent), followed by the United States (9.3 percent). Owing to their economic size, it is not thinkable that Germany and the United States, as well as other great economies, can have much higher (basic) indicator values. The size effect can increase the index value of Germany and the United States and also their rank, without undermining the performance of Malaysia that ranks steadily first thanks to its capacity effect.

As shown in Table 2 (last two columns), for 12 countries the rank by index value is the same as that by indicator value; for 17 countries there is a fall and for 11 countries an increase. In other terms, 70 percent of the countries analysed show a change in rank. For example, the top four countries by EXP-to-GDP ratio (Malaysia, Singapore, Belgium and Czech Republic) maintain their positions in the IOEXP ranking. The Netherlands and Thailand invert their rank: there is not a considerable difference between their indicator values, and the size effect can improve the position of the European country.

As observed, the size effect can help in particular those countries with a high share in the phenomenon analysed. However, the size effect cannot work wonders without an adequate capacity effect that remains the fundamental base of the Index (see also Ferrieri 2006, p. 249).

The United States, for example, presents the lowest EXP-to-GDP ratio and though the size effect increases its index by around 64 percent, the latter remains very low (0.107) as compared to the other countries analysed and the US advances just one position (from 40th to 39th). The EXP-to-GDP ratio of China is equal to 30.72 percent. Thanks to the size effect the country grows by about 18 percent in terms of value and advances nine positions (from 21st to 12th). In the case of China, the EXP-to-GDP ratio is not low (it is above the average of the 40 countries) and is close to other countries that have a lower weight in the phenomenon analysed. So the size effect can significantly help China.

In this cross-country perspective, it is interesting to analyze those economies with a very similar indicator value. This is the case for Finland and Israel: both countries lose in terms of rank, but due to the size effect Finland surpasses Israel. Venezuela and Russia rank respectively 19th and 20th by indicator value, but due to the size effect Russia gains one place (ranking 19th) and Venezuela loses two positions (ranking 21st).

Further Considerations and Simulations on the Index

Due to its mathematical formulation, the Index is able to reconcile capacity with size as it preserves the role of capacity. In this regard, it is important to remark that the Index is basically expressed by the indicator value. The size effect, as said, can help those countries with a high size and particularly, among them, those having good performance in terms of capacity. For instance, the difference between Germany and the United States (see Table 1) is that the indicator value of the first country is notably higher than that of the US, while the difference in terms of weight between the two countries is not remarkable. So the index model allows more gains for Germany than for the United States.

The aggregates that determine the index value are assumed as those observed across countries at a given time. With respect to the benchmark (indicator value), it is necessary to fix an optimum over time in order to make time comparisons. As for size, shares are simply those referring to the time or period of analysis (a given year; three-year average, etc), being one the total of the phenomenon concerned. Obviously, as observed, it is really unlikely that a country can monopolise a given phenomenon.

On the other hand, the need to individuate a maximum weight, theoretically acceptable, depends on the importance and the meaning that one wants to give to the Index: to emphasize capacity by introducing a simple adjustment for size (as shown in the model proposed and calculations) or to better balance capacity and size (it is possible in the model proposed, as well). In the latter hypothesis, it is possible to fix as maximum share (one), for example, 25 percent of the (observed) world exports value instead of 100 percent. This seems to be a maximum weight theoretically acceptable. Some simulations are given in Table 3.
# Measuring Openness

Table 3: Index of Openness to (Merchandise) Exports
Values and Ranks Given the Assumptions of a Maximum Size in the World Exports Equivalent to 100% and 25%

<table>
<thead>
<tr>
<th>Country</th>
<th>Maximum Size = 100%</th>
<th>Maximum Size = 25%</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Share in Total EXP</td>
<td>IOEXP Rank</td>
<td>Share in Total EXP</td>
</tr>
<tr>
<td>Argentina</td>
<td>0.004</td>
<td>0.215</td>
<td>28</td>
</tr>
<tr>
<td>Australia</td>
<td>0.010</td>
<td>0.132</td>
<td>36</td>
</tr>
<tr>
<td>Austria</td>
<td>0.013</td>
<td>0.383</td>
<td>10</td>
</tr>
<tr>
<td>Belgium</td>
<td>0.035</td>
<td>0.814</td>
<td>3</td>
</tr>
<tr>
<td>Brazil</td>
<td>0.011</td>
<td>0.156</td>
<td>34</td>
</tr>
<tr>
<td>Canada</td>
<td>0.036</td>
<td>0.325</td>
<td>14</td>
</tr>
<tr>
<td>China</td>
<td>0.067</td>
<td>0.339</td>
<td>12</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>0.008</td>
<td>0.601</td>
<td>4</td>
</tr>
<tr>
<td>Denmark</td>
<td>0.009</td>
<td>0.299</td>
<td>20</td>
</tr>
<tr>
<td>Finland</td>
<td>0.007</td>
<td>0.313</td>
<td>16</td>
</tr>
<tr>
<td>France</td>
<td>0.051</td>
<td>0.241</td>
<td>26</td>
</tr>
<tr>
<td>Germany</td>
<td>0.103</td>
<td>0.391</td>
<td>9</td>
</tr>
<tr>
<td>Greece</td>
<td>0.002</td>
<td>0.069</td>
<td>40</td>
</tr>
<tr>
<td>Hong Kong SAR</td>
<td>0.002</td>
<td>0.114</td>
<td>37</td>
</tr>
<tr>
<td>India</td>
<td>0.009</td>
<td>0.110</td>
<td>38</td>
</tr>
<tr>
<td>Indonesia</td>
<td>0.008</td>
<td>0.272</td>
<td>24</td>
</tr>
<tr>
<td>Iran, Islamic Republic of</td>
<td>0.005</td>
<td>0.261</td>
<td>25</td>
</tr>
<tr>
<td>Ireland</td>
<td>0.012</td>
<td>0.536</td>
<td>7</td>
</tr>
<tr>
<td>Israel</td>
<td>0.004</td>
<td>0.311</td>
<td>17</td>
</tr>
<tr>
<td>Italy</td>
<td>0.040</td>
<td>0.216</td>
<td>27</td>
</tr>
<tr>
<td>Japan</td>
<td>0.064</td>
<td>0.152</td>
<td>35</td>
</tr>
<tr>
<td>Korea</td>
<td>0.029</td>
<td>0.371</td>
<td>11</td>
</tr>
<tr>
<td>Malaysia</td>
<td>0.014</td>
<td>1.000</td>
<td>1</td>
</tr>
<tr>
<td>Mexico</td>
<td>0.021</td>
<td>0.274</td>
<td>23</td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.041</td>
<td>0.579</td>
<td>5</td>
</tr>
<tr>
<td>Norway</td>
<td>0.009</td>
<td>0.307</td>
<td>18</td>
</tr>
<tr>
<td>Poland</td>
<td>0.008</td>
<td>0.284</td>
<td>22</td>
</tr>
<tr>
<td>Portugal</td>
<td>0.004</td>
<td>0.191</td>
<td>31</td>
</tr>
<tr>
<td>Russia</td>
<td>0.021</td>
<td>0.306</td>
<td>19</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>0.014</td>
<td>0.481</td>
<td>8</td>
</tr>
<tr>
<td>Singapore</td>
<td>0.011</td>
<td>0.861</td>
<td>2</td>
</tr>
<tr>
<td>South Africa</td>
<td>0.005</td>
<td>0.204</td>
<td>29</td>
</tr>
<tr>
<td>Spain</td>
<td>0.020</td>
<td>0.173</td>
<td>33</td>
</tr>
<tr>
<td>Sweden</td>
<td>0.014</td>
<td>0.337</td>
<td>13</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0.013</td>
<td>0.319</td>
<td>15</td>
</tr>
<tr>
<td>Thailand</td>
<td>0.011</td>
<td>0.571</td>
<td>6</td>
</tr>
<tr>
<td>Turkey</td>
<td>0.007</td>
<td>0.200</td>
<td>30</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>0.039</td>
<td>0.176</td>
<td>32</td>
</tr>
<tr>
<td>United States</td>
<td>0.093</td>
<td>0.107</td>
<td>39</td>
</tr>
<tr>
<td>Venezuela</td>
<td>0.004</td>
<td>0.293</td>
<td>21</td>
</tr>
</tbody>
</table>

(*) Total amounting to 8,826,396 million US$. (**) Total amounting to 2,220,699 million US$. The gross domestic product (GDP) values are unchanged as well as the CV ($\kappa = 2.04$).
As observed in Table 3, in the new assumption (25%) compared to the previous one (100%), the main gainers are those countries with the highest export shares: the United States, France, Japan, Germany, etc. The index value of the United States, for example, grows by more than 200 percent and the country gains 21 positions as compared to the previous assumption. The top ranked countries – those in the first three places – remain unchanged: Malaysia, Singapore and Belgium.

Impact Analysis of Change in Export Value: Case One

What happens if the export value of a country rises without any change in the world export value? Suppose that Singapore records an increase in its export value (for example, the value of merchandise exports of Singapore grows from the observed value of 98,576 million US$ to 120,000 million US$ corresponding exactly to the decrease in the same aggregate of Malaysia (-21,424 million US$). Results from this scenario are illustrated in Table 4a, assuming all other factors remain the same.

The indicator value of Malaysia decreases from 106.92 to 88.81 percent. The size in the world export of the same country decreases from 0.014 to 0.012. The decrease is -16.94 percent as compared to the previous situation (given in Table 2). The index value of Malaysia in terms of capacity becomes 0.796, and with the size effect increases to 0.800. The loss in terms of sole capacity is somewhat higher than that in global terms (capacity and size): -20.44 percent compared to -20.00 percent. In both terms, Malaysia ranks 2nd behind Singapore. Due to its growth in size, Singapore gains correspondingly in terms of indicator value (+21.73%). The index value (capacity effect) of Singapore rises by 16.60 percent while its global index (capacity + size) grows by 16.19 percent. As for all other countries, obviously, even if their indicator values and size remain unchanged, there is an equal variation (-4.22%) in the index value (capacity effect: IVN) due to the new higher benchmark (Singapore: 111.63%); and a different decrease in global terms (IOEXP).

Impact Analysis of Change in Export Value: Case Two

What happens if the export value of a country rises with a corresponding change in the total world export value? Suppose that the same increase in the export value of Singapore (+21.424 million US$) translates into a corresponding growth in the world total exports (+0.24%). All other values (for all countries) remain the same; results are shown in Table 4b.

As in the previous case, the indicator value of Singapore is the highest and rises in the same measure (111.63%; +21.73%). The size of Singapore in the world total exports rises by 21.44 percent (as compared to 21.73% of the previous case). As the increase in the export value of Singapore is assumed to produce coherently a growth in the world total export value (+0.24%), it results that all other countries register a corresponding decrease in their share. Malaysia maintains the same indicator value (106.92%) but, as in the previous case, loses one position in the ranking. The capacity decrease for Malaysia is the same as that of all other countries (-4.22%), while that in global terms (IOEXP) is -4.10 percent (compared to -20.00% of the previous case). Excluding Singapore and Malaysia, the decrease in global terms (IOEXP) for all countries is higher than in the previous case. However, in both assumptions, the decrease in global terms is lower than that in terms of sole capacity. This is due to the size effect.

Another parameter is considered in the expression of the size effect: the constant κ. The parameter κ is assumed as a simple measure of variability of the phenomenon at the denominator of the indicator value. In the case study, the gross domestic product (GDP) represents the denominator of the basic indicator and its importance in determining the same indicator value is not irrelevant: it has been considered useful to introduce a measure of variability of this indicator in the index model. Given its formulation and meaning, the coefficient of variation (CV) is a suitable indicator in this regard. A higher CV means a higher variability of a given phenomenon (GDP) relative to its mean, and (in the case study) this would be in favour of countries having a high economic size.
### Table 4a: Index of Openness to (Merchandise) Exports

Values and Changes in Percentage Terms (compared to values in Table 2) Due to a Simulated Increase in the Export Value of Singapore (*) without a Corresponding Increase in the World Exports

<table>
<thead>
<tr>
<th>Country</th>
<th>EXP/GDP% (Vi)</th>
<th>Share in World EXP</th>
<th>IVN</th>
<th>IOEXP</th>
<th>EXP/GDP% (Vi)</th>
<th>Share in World EXP</th>
<th>IVN</th>
<th>IOEXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>22.67</td>
<td>0.004</td>
<td>0.203</td>
<td>0.206</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Australia</td>
<td>13.56</td>
<td>0.010</td>
<td>0.121</td>
<td>0.127</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Austria</td>
<td>39.84</td>
<td>0.013</td>
<td>0.357</td>
<td>0.367</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Belgium</td>
<td>85.75</td>
<td>0.035</td>
<td>0.768</td>
<td>0.782</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Brazil</td>
<td>15.98</td>
<td>0.011</td>
<td>0.143</td>
<td>0.149</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Canada</td>
<td>31.86</td>
<td>0.036</td>
<td>0.285</td>
<td>0.312</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>China</td>
<td>30.72</td>
<td>0.067</td>
<td>0.275</td>
<td>0.326</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>63.75</td>
<td>0.008</td>
<td>0.571</td>
<td>0.576</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Denmark</td>
<td>31.33</td>
<td>0.007</td>
<td>0.281</td>
<td>0.287</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Finland</td>
<td>32.95</td>
<td>0.007</td>
<td>0.295</td>
<td>0.300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>France</td>
<td>21.94</td>
<td>0.051</td>
<td>0.197</td>
<td>0.232</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Germany</td>
<td>33.12</td>
<td>0.103</td>
<td>0.297</td>
<td>0.378</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Greece</td>
<td>7.31</td>
<td>0.002</td>
<td>0.066</td>
<td>0.066</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Hong Kong SAR</td>
<td>12.05</td>
<td>0.002</td>
<td>0.108</td>
<td>0.109</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>India</td>
<td>11.35</td>
<td>0.009</td>
<td>0.102</td>
<td>0.106</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Indonesia</td>
<td>28.42</td>
<td>0.008</td>
<td>0.255</td>
<td>0.261</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Iran, Islamic Republic of</td>
<td>27.52</td>
<td>0.005</td>
<td>0.247</td>
<td>0.250</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Ireland</td>
<td>56.46</td>
<td>0.012</td>
<td>0.506</td>
<td>0.514</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Israel</td>
<td>32.95</td>
<td>0.004</td>
<td>0.295</td>
<td>0.298</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Italy</td>
<td>20.24</td>
<td>0.040</td>
<td>0.181</td>
<td>0.208</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Japan</td>
<td>12.33</td>
<td>0.064</td>
<td>0.111</td>
<td>0.146</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Korea</td>
<td>37.35</td>
<td>0.029</td>
<td>0.335</td>
<td>0.356</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Malaysia</td>
<td>88.81</td>
<td>0.012</td>
<td>0.796</td>
<td>0.800</td>
<td>-16.94</td>
<td>-16.94</td>
<td>-20.44</td>
<td>-20.00</td>
</tr>
<tr>
<td>Mexico</td>
<td>27.66</td>
<td>0.021</td>
<td>0.248</td>
<td>0.263</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Netherlands</td>
<td>58.96</td>
<td>0.041</td>
<td>0.528</td>
<td>0.556</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Norway</td>
<td>32.05</td>
<td>0.009</td>
<td>0.287</td>
<td>0.294</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Poland</td>
<td>29.66</td>
<td>0.008</td>
<td>0.266</td>
<td>0.272</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Portugal</td>
<td>20.14</td>
<td>0.004</td>
<td>0.180</td>
<td>0.183</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Russia</td>
<td>31.06</td>
<td>0.021</td>
<td>0.278</td>
<td>0.294</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>50.31</td>
<td>0.014</td>
<td>0.451</td>
<td>0.461</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Singapore</td>
<td>111.63</td>
<td>0.014</td>
<td>1.000</td>
<td>1.000</td>
<td>21.73</td>
<td>21.73</td>
<td>16.60</td>
<td>16.19</td>
</tr>
<tr>
<td>South Africa</td>
<td>21.41</td>
<td>0.005</td>
<td>0.192</td>
<td>0.195</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Spain</td>
<td>17.16</td>
<td>0.020</td>
<td>0.154</td>
<td>0.166</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Sweden</td>
<td>34.94</td>
<td>0.014</td>
<td>0.313</td>
<td>0.323</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Switzerland</td>
<td>33.01</td>
<td>0.013</td>
<td>0.296</td>
<td>0.306</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Thailand</td>
<td>60.25</td>
<td>0.011</td>
<td>0.540</td>
<td>0.547</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Turkey</td>
<td>20.86</td>
<td>0.007</td>
<td>0.187</td>
<td>0.192</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>16.26</td>
<td>0.039</td>
<td>0.146</td>
<td>0.169</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>United States</td>
<td>6.98</td>
<td>0.093</td>
<td>0.063</td>
<td>0.103</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
<tr>
<td>Venezuela</td>
<td>31.07</td>
<td>0.004</td>
<td>0.278</td>
<td>0.281</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-4.22</td>
</tr>
</tbody>
</table>

EXP: Merchandise Exports. IVN: Indicator value normalised: \( \frac{Vi}{V_{MAX}} = 111.63 \). IOEXP: Index of Openness to (Merchandise) Exports. The gross domestic product (GDP) values are unchanged. (*) The growth in the export value of Singapore (+21,424 million US$) corresponds to an equivalent decrease in that of Malaysia.
Table 4b: Index of Openness to (Merchandise) Exports
Values and Changes in Percentage Terms (compared to values in Table 2) Due to a Simulated Increase in the Export Value of Singapore (*) with a Corresponding Increase in the World Exports.

<table>
<thead>
<tr>
<th>Country</th>
<th>EXP/GDP (%) (Vi)</th>
<th>Share in World EXP</th>
<th>IVN</th>
<th>IOEXP</th>
<th>EXP/GDP% (% (Vi))</th>
<th>Share in World EXP</th>
<th>IVN</th>
<th>IOEXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>22.67</td>
<td>0.004</td>
<td>0.203</td>
<td>0.206</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.19</td>
</tr>
<tr>
<td>Australia</td>
<td>13.56</td>
<td>0.010</td>
<td>0.121</td>
<td>0.127</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Austria</td>
<td>39.84</td>
<td>0.013</td>
<td>0.357</td>
<td>0.367</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.11</td>
</tr>
<tr>
<td>Belgium</td>
<td>85.75</td>
<td>0.035</td>
<td>0.768</td>
<td>0.782</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.93</td>
</tr>
<tr>
<td>Brazil</td>
<td>15.98</td>
<td>0.011</td>
<td>0.143</td>
<td>0.149</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.14</td>
</tr>
<tr>
<td>Canada</td>
<td>31.86</td>
<td>0.036</td>
<td>0.285</td>
<td>0.312</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.94</td>
</tr>
<tr>
<td>China</td>
<td>30.72</td>
<td>0.067</td>
<td>0.275</td>
<td>0.326</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.71</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>63.75</td>
<td>0.008</td>
<td>0.571</td>
<td>0.576</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Denmark</td>
<td>31.33</td>
<td>0.009</td>
<td>0.281</td>
<td>0.287</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Finland</td>
<td>32.95</td>
<td>0.007</td>
<td>0.295</td>
<td>0.300</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.16</td>
</tr>
<tr>
<td>France</td>
<td>21.94</td>
<td>0.051</td>
<td>0.197</td>
<td>0.232</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.84</td>
</tr>
<tr>
<td>Germany</td>
<td>33.12</td>
<td>0.103</td>
<td>0.297</td>
<td>0.378</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.44</td>
</tr>
<tr>
<td>Greece</td>
<td>7.31</td>
<td>0.002</td>
<td>0.066</td>
<td>0.066</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.21</td>
</tr>
<tr>
<td>Hong Kong SAR</td>
<td>12.05</td>
<td>0.002</td>
<td>0.108</td>
<td>0.109</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.20</td>
</tr>
<tr>
<td>Ireland</td>
<td>11.35</td>
<td>0.009</td>
<td>0.102</td>
<td>0.106</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Indonesia</td>
<td>28.42</td>
<td>0.008</td>
<td>0.255</td>
<td>0.260</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Iran, Islamic Republic of</td>
<td>27.52</td>
<td>0.005</td>
<td>0.247</td>
<td>0.250</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.18</td>
</tr>
<tr>
<td>Ireland</td>
<td>56.46</td>
<td>0.012</td>
<td>0.506</td>
<td>0.514</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.12</td>
</tr>
<tr>
<td>Israel</td>
<td>32.95</td>
<td>0.004</td>
<td>0.295</td>
<td>0.298</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.18</td>
</tr>
<tr>
<td>Italy</td>
<td>20.24</td>
<td>0.039</td>
<td>0.181</td>
<td>0.208</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.92</td>
</tr>
<tr>
<td>Japan</td>
<td>12.33</td>
<td>0.064</td>
<td>0.111</td>
<td>0.146</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.76</td>
</tr>
<tr>
<td>Korea</td>
<td>37.35</td>
<td>0.029</td>
<td>0.335</td>
<td>0.356</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.99</td>
</tr>
<tr>
<td>Malaysia</td>
<td>106.92</td>
<td>0.014</td>
<td>0.958</td>
<td>0.959</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.10</td>
</tr>
<tr>
<td>Mexico</td>
<td>27.66</td>
<td>0.021</td>
<td>0.248</td>
<td>0.263</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.05</td>
</tr>
<tr>
<td>Netherlands</td>
<td>58.96</td>
<td>0.040</td>
<td>0.528</td>
<td>0.556</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.89</td>
</tr>
<tr>
<td>Norway</td>
<td>32.05</td>
<td>0.009</td>
<td>0.287</td>
<td>0.294</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Poland</td>
<td>29.66</td>
<td>0.008</td>
<td>0.266</td>
<td>0.272</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.15</td>
</tr>
<tr>
<td>Portugal</td>
<td>20.14</td>
<td>0.004</td>
<td>0.180</td>
<td>0.183</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.19</td>
</tr>
<tr>
<td>Russia</td>
<td>31.06</td>
<td>0.021</td>
<td>0.278</td>
<td>0.294</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.06</td>
</tr>
<tr>
<td>Saudi Arabia</td>
<td>50.31</td>
<td>0.014</td>
<td>0.451</td>
<td>0.461</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.10</td>
</tr>
<tr>
<td>Singapore</td>
<td>111.63</td>
<td>0.014</td>
<td>1.000</td>
<td>1.000</td>
<td>21.73</td>
<td>21.44</td>
<td>16.60</td>
<td>16.19</td>
</tr>
<tr>
<td>South Africa</td>
<td>21.41</td>
<td>0.005</td>
<td>0.192</td>
<td>0.195</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.18</td>
</tr>
<tr>
<td>Spain</td>
<td>17.16</td>
<td>0.020</td>
<td>0.154</td>
<td>0.166</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.07</td>
</tr>
<tr>
<td>Sweden</td>
<td>34.94</td>
<td>0.014</td>
<td>0.313</td>
<td>0.323</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.11</td>
</tr>
<tr>
<td>Switzerland</td>
<td>33.01</td>
<td>0.013</td>
<td>0.296</td>
<td>0.306</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.11</td>
</tr>
<tr>
<td>Thailand</td>
<td>60.25</td>
<td>0.011</td>
<td>0.540</td>
<td>0.547</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.13</td>
</tr>
<tr>
<td>Turkey</td>
<td>20.86</td>
<td>0.007</td>
<td>0.187</td>
<td>0.192</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.16</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>16.26</td>
<td>0.039</td>
<td>0.146</td>
<td>0.169</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.93</td>
</tr>
<tr>
<td>United States</td>
<td>6.98</td>
<td>0.093</td>
<td>0.063</td>
<td>0.103</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-3.58</td>
</tr>
<tr>
<td>Venezuela</td>
<td>31.07</td>
<td>0.004</td>
<td>0.278</td>
<td>0.281</td>
<td>-</td>
<td>-0.24</td>
<td>-4.22</td>
<td>-4.19</td>
</tr>
</tbody>
</table>

EXP: Merchandise Exports. IVN: Indicator value normalised: Vi/V MAX = 111.63. IOEXP: Index of Openness to (Merchandise) Exports. The gross domestic product (GDP) values are unchanged. (*) The growth in the export value of Singapore is the same as in the previous table: +21,424 million US$. 

FERRIERI
It is also important to remark that $\kappa$ is constant. So, theoretically, all countries could take advantage of a growth in $\kappa$, but notably (and coherently) those with a high size in the phenomenon concerned. Any change in the $\kappa$ value does imply no change in the index value for the first ranked country, whose performance is determined exclusively by the indicator value. Finally, there is no need to comment that if $\kappa = 1$ the index value for all countries is determined by their “capacity effect” and a size effect based only on their share in the phenomenon concerned. If $\kappa = 0$ is equal to zero the index value is given by the sole capacity effect.

Conclusion
The proposed method for measuring international openness takes into account not only the capacity of countries for a given phenomenon but also their weight in the same. The values of the proposed index can be compared in both static and dynamic perspectives. In the latter approach (time comparisons) it is necessary to set a suitable benchmark for measuring the capacity effect, such as the maximum value observed across countries over time. It is also suitable to calculate the exponent factor ($\kappa$) over time in order to appreciate changes affecting the indicator value and the share in the phenomenon (a) concerned for each country. This method could be applied to many transferable phenomena, expressed in terms of flow and/or stock, such as international migration, foreign direct investment and many others.

References
Another Look at Resampling:
Replenishing Small Samples with Virtual Data through S-SMART

Haiyan Bai  
University of Central Florida

Wei Pan  
Leigh Lihshing Wang  
Phillip Neal Ritchey  
University of Cincinnati

A new resampling method is introduced to generate virtual data through a smoothing technique for replenishing small samples. The replenished analyzable sample retains the statistical properties of the original small sample, has small standard errors and possesses adequate statistical power.

Key words: Small sample, resampling, smoothing, bootstrap.

Introduction
Drawing statistical inferences based upon small samples has long been a concern for researchers because small samples typically result in large sampling errors (Hansen, Madow, & Tepping, 1983; Lindley, 1997) and small statistical powers (Cohen, 1988); thus, sample size is essential to the generalization of the statistical findings and the quality of quantitative research (Lindley, 1997). Unfortunately, sample size is often constrained by the cost or the restrictions of data collection (Adcock, 1997), especially for research involving sensitive topics. Consequently, it is not unusual for researchers to use small samples in their quantitative studies. For example, in American Educational Research Journal and Journal of Consulting and Clinical Psychology, 155 out of 575 (27.0%) quantitative research articles published between 2003 and 2007 used relatively small sample sizes ranging from 10 to 100. Therefore, the need for studies addressing the problem of small samples is prominent.

With the development of modern computer science, four commonly used resampling methods emerged as revolutionary techniques to address small sample problems (Diaconis & Efron, 1983), these are:

(a) randomization test (Fisher, 1935);
(b) cross-validation (Kurtz, 1948);
(c) the jackknife (Quenouille, 1949; modified by Tukey, 1958); and
(d) the bootstrap (Efron, 1979, 1982).

The bootstrap is the most commonly-used resampling method in research with small samples (Efron & Tibshirani, 1998). Not only have resampling methods been applied to basic statistical estimation, such as estimation bias, standard errors, and confidence intervals, but they also have been applied to more advanced statistical modeling, such as regression (Efron, 1979, 1982; Hinkley, 1988; Stine, 1989; Wu, 1986), discriminant analysis (Efron,1979), time series (Hinkley, 1988), analyses with censored data (Efron, 1979; Efron & Gong 1983), missing data (Efron, 1994), and dependent data (Lahiri, 2003).

Existing resampling methods are very useful statistical tools for dealing with small sample problems; however, they have critical limitations (see Bai & Pan, 2008, for a review).
Specifically, the randomization test requires data that can be rearranged, the cross-validation splits an already small sample, and the jackknife usually needs a large sample to obtain good estimates (Davison & Hinkley, 1999). Bootstrap, the most commonly-used method, also has a few limitations. For example, the lack of independent observations is evident due to the possible duplications of the observations in the bootstrap samples that are randomly drawn from the same small samples with replacement (Efron & Tibshirani, 1998); and the bootstrap technique is potentially nonrobust to outliers because every observation in the small sample, including the outliers, has an equal chance of being selected (Troendle, Korn & McShane, 2004).

More importantly, all of the resampling methods have a common problem: the new resamples are limited to being selected from the same original small sample, which leaves a considerable number of data points unselected in the population. The problem with this common resampling process is that the resamples carry no more information than the original small sample. This common limitation of existing resampling methods potentially undermines the validity of the statistical inference. Therefore, if a small sample could be replenished with additional data points to capture more information in the population, it would significantly reduce the common limitation of the extant resampling methods.

The smoothing technique (Simonoff, 1996) made a breakthrough in lessening the limitation of sampling from the original small sample; however, the smoothing procedure in the smoothed bootstrap (de Angelis & Young, 1992, Hall, DiCiccio, & Romano 1989; Silverman & Young 1987) is so complicated that even statisticians ask how much smooth is optimal. In addition, the question of when to use smoothing technique still remains problematic for researchers (Silverman & Young, 1987). Therefore, developing an alternative but simpler resampling method for lessening the limitations of the existent resampling methods is imperative.

The purpose of this article is to develop a new resampling method, namely Sample Smoothing Amplification Resampling Technique (S-SMART), for generating virtual data to replenish a small sample based on the information provided by the small sample, both in its own right and as an ingredient of other statistical procedures. The S-SMART technique directly employs kernel smoothing procedures to the original small sample before resampling so that the information carried by the original small sample is well retained. Not only does the S-SMART procedure potentially lessen some limitations of the existing resampling methods, but also it reduces sampling errors and increases statistical power. Therefore, the use of the replenished sample through S-SMART provides more valid statistics for estimation and modeling than does the original small sample; and ultimately, it will improve the quality of research with small samples.

Specifically, this study aims to develop S-SMART as a distribution-free method to produce S-SMART samples which (a) have sufficient sample sizes to provide an adequate statistical power, (b) have stable sample statistics across different replenishing scenarios, (c) have smaller sampling errors, (d) are independent observations, (e) are robust to outliers, and (f) are easily obtained in a single simple operation. This study evaluated the accuracy of the statistics through both a simulation study and an empirical example.

Methodology

The S-SMART Procedure

The S-SMART procedure integrates smoothing technique and distribution theory into a new resampling method. The smoothing parameter, sample size, quantiles, and standard error of the mean of the original small sample are the simulation parameters for generating S-SMART samples.

Figure 1 shows a schematic diagram of the S-SMART procedure. First, the percentile range from 2.5% to 97.5% of the original small sample of size \( n \) is evenly divided into \( k \) equal intervals for obtaining the corresponding quantiles \( q_i, i = 0, 1, \ldots, k \); second, the quantiles are used as means for the small smoothing Gaussian kernels and the standard error of the mean of the small sample are used as a random noise for the Gaussian kernels; third, random samples of size \( s \) are generated from the Gaussian kernels as the S-SMART sub-samples.
to replenish the small sample; and last, a union of the \((k + 1)\) S-SMART sub-samples is taken to obtain an amplified S-SMART sample with enlarged sample size of \((k + 1)\times s\).

The following algorithm is a more explicit description of the S-SMART procedure for obtaining the amplified S-SMART sample with replenishing virtual data:

1. Let \(U\) be a population with an unknown underlying probability density function \(f(x)\).
2. Let \(X = \{x_1, x_2, \ldots, x_n\}\) be an independent identically distributed (i.i.d.) random small sample of size \(n\) drawn from \(U\).
3. Let \(f_n(x)\) be the empirical probability density function of \(X\).
4. Let \(Q = \{q_0, q_1, \ldots, q_k\}, \quad k \geq n\), be a set of quantiles whose corresponding percentile ranks equally divide the 95% range of \(X\)'s percentile ranks into \(k\) intervals (i.e., \(q_0 \sim 2.5\%, \ldots, q_k \sim 97.5\%)\).
5. Let \(f_k^{(i)}(x)\) be the empirical probability density function of \(Q\). By the probability distribution is the inverse of its cumulative distribution function; therefore, \(f_k^{(i)}(x)\) captures the shape of the distribution \(f_k(x)\) (Gilchrist, 2000). Using the 95% range of the percentile ranks, instead of using all the percentile ranks, is intended to eliminate possible outliers; and the equal division of the 95% percentile range complies with the principle of smoothing (Simonoff, 1996).
6. Let \(X_i^* = \{x_{i1}^*, x_{i2}^*, \ldots, x_{is}^*\}, \quad i = 0, 1, \ldots, k\), be an i.i.d. random sample of size \(s\) drawn from a Gaussian kernel:

\[
G_i(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{x - q_i}{h} \right)^2 \right);
\]

that is, \(x_{ij}^* = q_i + h\varepsilon_{ij}\), where \(\varepsilon_{ij} \sim N(q_i, h^2)\), \(j = 1, 2, \ldots, s; \quad i = 0, 1, \ldots, k\). The choice for the bandwidth \(h\) can be the standard error of \(\bar{X}\) or \(\bar{Q}\) as suggested by Hesterberg (1999). The kernel estimators center the kernel

Figure 1: A Schematic Diagram of the S-SMART Procedure

Notes: (1) \(q_0, q_1, \ldots, q_k\) are the quantiles whose corresponding percentiles evenly divide the middle 95% percentile range of the original small sample; (2) \(k\) is the number of intervals which determine \(k + 1\) Gaussian kernels, each uses \(q_i\) \((i = 0, 1, \ldots, k)\) as its mean and the standard error of the mean of the original small sample as its standard deviation.
function at each quantile point \( q_i \), which smooths out the contribution of each point over a local neighborhood of that point. \( X_i^* \), \( i = 0, 1, \ldots, k \), serve as the replenishing virtual data to the small sample \( X \).

7. Let \( S^* = \bigcup_{i=0}^{k} X_i^* \). Then, the empirical probability density function of \( S^* \) can be expressed as a weighted average over the Gaussian kernel functions (Parzen, 1962; Silverman, 1986):

\[
f_k^*(x) = \frac{1}{kh} \sum_{i=0}^{k} G_i(x) = \frac{1}{kh\sqrt{2\pi}} \sum_{i=0}^{k} \exp \left( -\frac{1}{2} \left( \frac{x - q_i}{h} \right)^2 \right).
\]

\( S^* \) is called as the amplified S-SMART sample with an enlarged sample size of \((k + 1)\times\). \( f_k^*(x) \) is a consistent estimate of \( f(x) \) because the Gaussian kernel \( G_i(x) \) is a weighting function and the bandwidth \( h \), specified as the standard error, satisfies the conditions \( \lim_{k \to \infty} h = 0 \) and \( \lim_{k \to \infty} kh = \infty \) (Parzen, 1962).

It is notable that the resamples of all the existing resampling methods are the replica of the original data points in the small sample which is in practice not always representative of the population through the randomization because “random sampling for representativeness is theoretically possible on a larger scale” (Cook & Campbell, 1979, p. 75). On the contrary, this newly-developed S-SMART method intends to recover the randomization through the random noise of the Gaussian kernels in the smoothing technique, rather than striving to achieve a one-mode estimated empirical distribution from the small sample, which is the common goal of smoothing.

Monte Carlo

As with all other resampling methods, the S-SMART method is bound to have some properties that are mathematically difficult to evaluate; therefore, a simulation study is conducted to provide additional knowledge and numerical evidence to address the questions about the method’s properties (Johnson, 1987). The statistical behaviors of the S-SMART samples from both the simulated data and empirical data are evaluated in terms of the first two generalized method of moments (GMM), mean and variance, which are commonly used to describe a set of data (Gravetter & Wallnau, 2004). The sample distributions, sampling distributions of the means, and confidence intervals of the means and standard deviations are also studied.

The evaluation of the S-SMART method is conducted for small samples from normal, exponential, and uniform distributions, representing the three families of continuous distributions, which demonstrates that the S-SMART procedure is a distribution-free technique. The Statistical Analysis System (SAS) (SAS Institute Inc., 2001) is employed to both generate the small samples and resample the replenishing virtual data.

To investigate the stability of the S-SMART samples, the first two generalized method of moments of the S-SMART samples amplified from the simulated random small samples were compared across different amplifying times for each different sample size of the small samples. The different sample sizes of the small samples are 10, 15, 20, 25, and 50, which were randomly generated from normal, exponential, and uniform distributions. The small sample sizes were determined according to a power analysis, that is, the smallest sample size for statistical analysis with an adequate power, such as the \( t \)-test, is 64 (Cohen, 1988). The corresponding amplified S-SMART samples were simulated with the amplified sample sizes as 10, 20, 50, and 100 times as the original small sample sizes; accordingly, the sample sizes of the S-SMART samples ranged from 100 to 50,000, providing adequate power for a variety of statistical analyses (Cohen, 1988).

Results

Evaluating the Sample Distributions

To understand the properties of the distribution of the S-SMART samples, histograms of the S-SMART samples and the corresponding original small samples were
compared. Three sets of five small samples of size 10, 15, 20, 25, and 50 were randomly generated from three different distributions: standard normal, exponential, and uniform, respectively; then, they were amplified 10, 20, 50, and 100 times, respectively, through the S-SMART procedure.

For illustration purpose, Figure 2 only shows the amplification results for the small sample of size 20 because the amplification results for the small samples of other sizes were the same. In Figure 2, the histograms on the left panel illustrate that at all levels of the amplifying times, the S-SMART samples imitated the original small-sample distribution. That is, all the S-SMART samples amplified from the normally-distributed small sample appear to be also normally distributed. The same phenomenon was also true for both the exponential distribution and uniform distribution (see the middle and right panels in Figure 2).

Evaluating the First Two Moments

To test the stability of the sample statistics of the S-SMART sample across the different S-SMART sample sizes, an ANOVA test for equal means and the Brown and Forsythe’s (1974) test for equal variances were conducted. The Brown and Forsythe’s test for homogeneity of variances is a more robust and powerful test for comparing several variances under heteroscedasticity and non-normality (O’Brien, 1992; Tabatabai, 1985). It is important to examine the stability of variances because the variance is another essential measure reflecting the statistical properties of data.

The ANOVA results (see Tables 1, 2, 3) indicate that the S-SMART samples statistically did not have mean differences across all levels of the S-SMART samples and also have equal means with those of the original samples with almost all the p-values larger than .90; therefore, they had stable means with different amplifying times. Thus, the S-SMART procedure cloned samples carrying the same most important statistical behavior of the original small sample.

The results of the Brown and Forsythe’s test showed that no significant differences existed among the variances across different amplifying times and the small samples from non-normal, exponential, and uniform distributions (Tables 2, 3). The phenomenon is also generally true for the normal distribution (Table 1) except for the last two cases (see last two p-values for the Brown and Forsythe’s test in Table 1) where the sample sizes are very large, which inevitably caused overpowered tests with biased small p-values for testing equal variances between groups.

The significance of the Brown and Forsythe’s test for equal variance were generated from the very large sample sizes with excessive power and extremely unbalanced sample sizes. Nonetheless, when all the variances in Tables 1, 2 and 3 were examined, it was found that the magnitudes of the variances did not differ substantially. Taking the most significant group in the three tables, for example, the small sample of size 50 in Table 1 with the p-value of .001, the relative differences (the absolute value of (S-SMART_SD – Small-Sample_SD)/ Small-Sample_SD) are only .033 for the S-SMART of size 500, .041 for the S-SMART of size 1000, and .091 for S-SMART of size 5000. According to a rule of thumb provided by Efron and Tibshirami (1993), if the relative difference is less than .25, the differences can be ignored. As results show, all the relative differences are smaller than .25.

Evaluating Confidence Intervals

To further investigate the properties of the S-SMART samples, three groups of 95% confidence intervals were estimated for comparing the S-SMART samples with the original small samples from the three representative distributions: normal, exponential, and uniform. The current intervals were constructed parametrically because the S-SMART sample sizes were large enough (all > 100) for the sampling distributions to be approximately normal according to the central limit theorem. Figure 3 shows fifty replications of the confidence intervals generated based on the original small samples of size 20 from standard normal, exponential, and uniform distributions. The longest bar for the interval corresponding to the label 50 represents the confidence interval of the original small sample from each distribution. The clustered, short intervals corresponding to the labels Li, i = 1, 2, ..., 50, represent the confidence intervals for the
Figure 2: Histograms of the Original Small Sample ($n = 20$) and the S-SMART Samples

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Original Small Sample</th>
<th>Amplified 10 Times</th>
<th>Amplified 20 Times</th>
<th>Amplified 50 Times</th>
<th>Amplified 100 Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Histograms showing data distributions for each category, amplified at different factors.]
Table 1: Statistics Stability Test Results for Normal Data

<table>
<thead>
<tr>
<th>Small Sample (n)</th>
<th>S-SMART (n)</th>
<th>M</th>
<th>SD</th>
<th>ANOVA F</th>
<th>p</th>
<th>Brown and Forsythe’s Test F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>-0.22349274</td>
<td>0.96181667</td>
<td>0.04</td>
<td>0.9971</td>
<td>0.24</td>
<td>0.9143</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-0.19069581</td>
<td>0.90544626</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>-0.23229805</td>
<td>0.93433640</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.21896046</td>
<td>0.89065269</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>—</td>
<td>0.10073266</td>
<td>1.07928841</td>
<td>0.09</td>
<td>0.9869</td>
<td>1.39</td>
<td>0.2363</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>0.10655489</td>
<td>1.16529037</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.09984589</td>
<td>1.10383676</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>0.09580675</td>
<td>1.04833640</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>0.10087356</td>
<td>1.02527119</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>—</td>
<td>0.05855377</td>
<td>1.00343478</td>
<td>0.13</td>
<td>0.9706</td>
<td>0.79</td>
<td>0.5317</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.08582172</td>
<td>1.04418680</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.06449656</td>
<td>0.98977889</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.04226237</td>
<td>0.94223102</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.04501257</td>
<td>0.93789847</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>—</td>
<td>0.06687785</td>
<td>0.95343286</td>
<td>0.21</td>
<td>0.9315</td>
<td>3.64</td>
<td>0.0058</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.06043331</td>
<td>1.04887182</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.05241659</td>
<td>0.97349343</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1250</td>
<td>0.08302555</td>
<td>0.89050723</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.08860332</td>
<td>0.88523717</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>—</td>
<td>-0.09631057</td>
<td>0.95323886</td>
<td>0.09</td>
<td>0.9855</td>
<td>4.64</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>-0.08740131</td>
<td>0.98512214</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.08806617</td>
<td>0.91383556</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>-0.07644308</td>
<td>0.87315825</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>-0.07321744</td>
<td>0.86062323</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Table 2: Statistics Stability Test Results for Exponential Data

<table>
<thead>
<tr>
<th>Small Sample (n)</th>
<th>S-SMART (n)</th>
<th>$M$</th>
<th>$SD$</th>
<th>ANOVA</th>
<th>Brown and Forsythe’s Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$F$</td>
<td>$p$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10.7014225</td>
<td>6.89706159</td>
<td>0.11</td>
<td>0.9795</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>11.1492282</td>
<td>6.82349786</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>10.6658978</td>
<td>6.84719726</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>10.7544083</td>
<td>6.64931520</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>10.7359794</td>
<td>6.56322669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>8.99177780</td>
<td>6.68143971</td>
<td>0.10</td>
<td>0.9840</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>9.29636554</td>
<td>6.81088054</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>9.16996847</td>
<td>6.55176270</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>9.03166915</td>
<td>6.43617622</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>9.02165396</td>
<td>6.41504747</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>8.48010059</td>
<td>6.72584611</td>
<td>0.05</td>
<td>0.9956</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>8.57848028</td>
<td>6.73279566</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>8.36743416</td>
<td>6.60753442</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>8.42812276</td>
<td>6.50174430</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>8.39118617</td>
<td>6.45948948</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>7.90545531</td>
<td>6.25441137</td>
<td>0.18</td>
<td>0.9465</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>8.11257241</td>
<td>6.56518979</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>7.95441471</td>
<td>6.19530349</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1250</td>
<td>7.82502044</td>
<td>6.03353920</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>7.82480295</td>
<td>5.94773232</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>8.49794161</td>
<td>7.36349688</td>
<td>0.31</td>
<td>0.8682</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>8.52943241</td>
<td>7.29392134</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>8.36130973</td>
<td>7.04365418</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>8.24090907</td>
<td>6.73168660</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>8.23225084</td>
<td>6.66201829</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Statistics Stability Test Results for Uniform Data

<table>
<thead>
<tr>
<th>Small Sample $(n)$</th>
<th>S-SMART $(n)$</th>
<th>$M$</th>
<th>$SD$</th>
<th>ANOVA $F$</th>
<th>$p$</th>
<th>Brown and Forsythe’s Test $F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>—</td>
<td>0.10577084</td>
<td>0.07134369</td>
<td>0.06</td>
<td>0.9924</td>
<td>0.21</td>
<td>0.9352</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.10730272</td>
<td>0.06681932</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.10555560</td>
<td>0.06843615</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.10687757</td>
<td>0.06975516</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.10521426</td>
<td>0.06763157</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>—</td>
<td>0.12313716</td>
<td>0.07294208</td>
<td>0.03</td>
<td>0.9988</td>
<td>0.09</td>
<td>0.9867</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>0.12102230</td>
<td>0.07446817</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.12148089</td>
<td>0.07131987</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>0.12254982</td>
<td>0.07204738</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>0.12231181</td>
<td>0.07118637</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>—</td>
<td>0.12922731</td>
<td>0.07208107</td>
<td>0.03</td>
<td>0.9987</td>
<td>0.19</td>
<td>0.9429</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.12916945</td>
<td>0.07348706</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>0.12800374</td>
<td>0.06998499</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>0.12892277</td>
<td>0.06975528</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>0.12920833</td>
<td>0.07004994</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>—</td>
<td>0.13314226</td>
<td>0.06703704</td>
<td>0.03</td>
<td>0.9980</td>
<td>1.24</td>
<td>0.2931</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.13349798</td>
<td>0.06986992</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.13219283</td>
<td>0.06673523</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1250</td>
<td>0.13271716</td>
<td>0.06498865</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.13314679</td>
<td>0.06537055</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>—</td>
<td>0.13108854</td>
<td>0.06864826</td>
<td>0.00</td>
<td>1.0000</td>
<td>1.81</td>
<td>0.1238</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.13114757</td>
<td>0.07018810</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.13115067</td>
<td>0.06762741</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>0.13112839</td>
<td>0.06676879</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.13109725</td>
<td>0.06599090</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
means and standard deviations of the S-SMART samples with fifty replications for each group. Fifty was the maximum number of replications that could provide a clear enough graphical presentation of the confidence intervals.

As shown in Figure 3, all means and standard deviations of the S-SMART samples were centered at the mean or standard deviation of the original small sample in each group. Furthermore, the confidence intervals of the S-SMART samples covered all the means and standard deviations of the original small samples, except for the standard deviation of the original small sample from the exponential distribution. This problem with the skewed distribution has nothing specifically to do with the S-SMART method. Even for the well-established bootstrap method, the interval estimation for the skewed data also needs adjustment to obtain a better estimation (Hesterberg, 1999).

Evaluating Sampling Distributions

The simulation study shows that the S-SMART sampling distributions closely followed the normal distribution as concluded in the central limit theorem. The histograms in Figures 4 display the sampling distributions of the S-SMART samples based on random small samples from standard normal, exponential, and uniform distributions. For illustration purposes, Figure 4 presents two samples for each distribution to show the property of the S-SMART sampling distributions. The two S-SMART samples were selected based on sample sizes of 10 and 100, representing the sampling distributions of samples with the largest difference among the small sample sizes.

An Empirical Example

S-SMART may be claimed as the easiest resampling method in application because it does not require researchers to learn
any computer programming or model modifications to obtain an adequate sample size to conduct statistical analysis. At the current stage, a SAS macro program is ready for researchers to directly plug in their small sample to get the amplified S-SMART samples. A stand-alone computer program will be available soon. In this article, the SAS macro is presented as an example for the application of the S-SMART method.

To use the SAS macro, researchers need only to specify five macro variables in a short SAS macro: %S_SMART (in=, k=, NUM=, Var=, outfile=) to invoke the S-SMART macro %macro S_SMART (in=, k=, NUM=, Var=, outfile=) which is available from the first author. Researchers simply plug in the small sample data file after in =, the small sample size after k =, the times to amplify the small sample after NUM =, the name of the variable in the small sample after var =, and the output file to save the amplified S-SMART sample after outfile =. After running the SAS macro program, the S-SMART sample is ready for statistical analysis.

To illustrate how to obtain an S-SMART sample from an empirical small sample using the SAS macro program, a random sample \((N = 269)\) was drawn from a real, large-scale study of education: the National Educational Longitudinal Study of 1988 (NELS:88) database. A small sample of 20 cases was further randomly drawn from the random sample. An achievement variable, \(bytxhstd\) (base year history/citizenship/geography standardized test score), was used and was renamed as \(BHIST20\) (see Table 4).

If it is desired to amplify the 20 cases into 200 \((10\) times\), the small sample data file is C:\NELS20hist.dat, the variable name is \(BHIST20\), and the amplified output data file is C:\NELS20_200hist.txt. At this point the amplified S-SMART data can be obtained by plugging in the five macro variables into the SAS macro as follows: %S_SMART (in = C:\NELS20hist.dat, k = 20, NUM = 10, Var = BHIST20, outfile = C:\NELS20_200hist.txt).

This SAS macro program invokes the S-SMART macro % macro S_SMART (in=, k=, NUM=, Var=, outfile=); then the amplified S-SMART data in a text file is saved as C:\NELS20_200hist.txt. To study the property of the sample distribution of the S-SMART sample from the empirical data, histograms of the S-SMART sample were compared with the

![Figure 4: Histograms of 1,000 S-SMART Samples Based on Small Samples Sizes of 10 and 100](image-url)
ANOTHER LOOK AT RESAMPLING

The shape of the small sample distribution (Figure 5, left panel) roughly reflected the random sample of NELS:88 (Figure 5, right panel), but it had a gap between the scores of 33 and 38; while the sample distribution of the S-SMART sample (Figure 5, middle panel) closely followed the shape of the small sample with a similar gap.

To compare the means and variances between the small sample of size 20 from the random sample of the NELS:88 dataset, the corresponding S-SMART sample, and the random sample of NELS:88, ANOVA test and the homogeneity test were conducted over the three sets of data. The test results are shown in Table 5. From the ANOVA results it was found that there were no mean differences among the three groups of data with a $p$-value of .189 even with the unbalanced group. With two random errors, the sampling error and the Monte Carlo error, the S-SMART sample still reflect the sample mean of the small sample and the mean of the random sample from which the small sample was selected. When the means of the amplified S-SMART sample were compared with the small sample using the $t$-test, a $p$-value of .993 resulted, thus revealing the equality of the two sample means.

The results of the Brown and Forsythe's test revealed that equal variances were assumed among the three groups of data with a $p$-value of .762. Under this condition, it was found that the mean standard error (.68) of the S-SMART sample is very close to that (.59) of the random sample, and as expected, it is much smaller than that (2.07) of the small sample. This finding demonstrates that the S-SMART method has the potential to reduce sampling errors while maintaining all other statistical properties carried by the small sample.

To explore the property of the interval estimation of the S-SMART sample for the empirical data, the interval estimation of the means among the small sample, the S-SMART sample, and the random sample were compared.

Table 4: The Small Sample of 20 Cases from the NELS:88 Database

<table>
<thead>
<tr>
<th>No.</th>
<th>BHIST20</th>
<th>No.</th>
<th>BHIST20</th>
<th>No.</th>
<th>BHIST20</th>
<th>No.</th>
<th>BHIST20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.508</td>
<td>6</td>
<td>59.132</td>
<td>11</td>
<td>52.907</td>
<td>16</td>
<td>47.009</td>
</tr>
<tr>
<td>2</td>
<td>67.761</td>
<td>7</td>
<td>57.385</td>
<td>12</td>
<td>52.907</td>
<td>17</td>
<td>46.026</td>
</tr>
<tr>
<td>3</td>
<td>64.266</td>
<td>8</td>
<td>56.402</td>
<td>13</td>
<td>50.396</td>
<td>18</td>
<td>46.026</td>
</tr>
<tr>
<td>4</td>
<td>60.770</td>
<td>9</td>
<td>55.747</td>
<td>14</td>
<td>47.883</td>
<td>19</td>
<td>40.020</td>
</tr>
<tr>
<td>5</td>
<td>59.460</td>
<td>10</td>
<td>53.781</td>
<td>15</td>
<td>47.883</td>
<td>20</td>
<td>31.392</td>
</tr>
</tbody>
</table>

Figure 5. Histogram of the Scores in Base Year History of 20 Cases from NELS:88 and S-SMART Samples
Table 6 shows that S-SMART interval estimation of the mean is much narrower than that of the small sample and very close to the random sample. This result further reveals that if a small sample is randomly selected from a population, the S-SMART method can replenish the small sample to obtain a larger sample with the same statistical properties as those of any random sample of a comparable sample size from the same population.

In sum, the application of the S-SMART method to real-world data demonstrated that the newly-developed resampling method can be utilized in the real-world settings. The evaluation on the quality of the S-SMART sample yielded the same results as those of from the simulated data. In other words, the S-SMART sample generated from the real-world data has the same sample distribution as that of the original small sample; furthermore, the S-SMART method can replenish an original real-world small sample to a larger sample with the same sample distribution, mean and standard deviation, while the standard error is reduced.

**Conclusion**

This study developed a new resampling method, S-SMART, which can replenish a small sample becoming a larger sample to meet the requirement of sample size for inferential statistics. It allows researchers to use familiar statistical analysis directly on the amplified S-SMART samples.

**S-SMART is a Distribution-Free Method**

According to the theoretical verification, simulation study, and empirical evidence of distributional characteristics of the amplified S-SMART samples, it was also demonstrated that S-SMART is a distribution-free method. From the distributional study it was found that the S-SMART method can amplify an original small sample from any distributions into a larger sample with the same distribution as that of the original small sample. Regardless of whether the sample distribution of the original small samples are symmetric or asymmetric, the sample distribution of the S-SMART sample follows the same distribution as does the original sample; and the sampling distributions of the S-SMART samples are normal. The use of the Gaussian kernel smoothing over the percentile range from 2.5% to 97.5% of the original small sample captured the shape of the original small-sample distribution.

It may be argued that the S-SMART sample copies the sampling bias caused by the small size of the original sample. However, if a

<table>
<thead>
<tr>
<th>Sample</th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
<th>Std. Error of Mean</th>
<th>ANOVA $F$</th>
<th>$P$</th>
<th>Brown and Forsythe’s Test $F$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Sample</td>
<td>20</td>
<td>53.3331</td>
<td>9.2380</td>
<td>2.0657</td>
<td>1.672</td>
<td>.189</td>
<td>.502</td>
<td>.762</td>
</tr>
<tr>
<td>S-SMART Sample</td>
<td>220</td>
<td>53.3136</td>
<td>10.1527</td>
<td>.6845</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Sample</td>
<td>269</td>
<td>51.7111</td>
<td>9.6217</td>
<td>.5911</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>$N$</th>
<th>$M$</th>
<th>$SD$</th>
<th>Std. Error of Mean</th>
<th>CI Lower</th>
<th>CI Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Sample</td>
<td>20</td>
<td>53.3331</td>
<td>9.2380</td>
<td>2.0657</td>
<td>49.0059</td>
<td>57.6566</td>
</tr>
<tr>
<td>S-SMART Sample</td>
<td>220</td>
<td>53.3136</td>
<td>10.1527</td>
<td>.6845</td>
<td>51.9645</td>
<td>54.6626</td>
</tr>
<tr>
<td>Random Sample</td>
<td>265</td>
<td>51.7111</td>
<td>9.6217</td>
<td>.5911</td>
<td>50.5474</td>
<td>52.8749</td>
</tr>
</tbody>
</table>
small sample is not representative of the population due to the sampling bias, the sample is not valid to be used for any statistical analysis or any other resampling methods; therefore, in the case of sampling bias, researchers’ judgment must be relied upon to assess the quality of the data collected.

S-SMART is a Robust Procedure

From the simulation study, it was found that the S-SMART samples based on the original small samples from various distributions are robust to outliers. By using the middle 95% of percentiles instead of all the data points in the original sample, the S-SMART technique can reduce the influence of the extreme values or potential outliers in the original small sample. At the same time, some beneficial information carried by the extreme values can be retained through the estimation of the percentiles from the original sample.

The S-SMART Sample Statistics Are Stable

The results of the $F$-tests in ANOVA of the simulation study have shown that the means of the S-SMART samples are statistically equal across all different sizes of replenished samples for each of the different sizes of the original small samples. The stability of the means of the S-SMART samples makes the method reliable in representing the mean values of the original small samples at any times of amplification. The $F$-tests also show that the homogenous S-SMART sample means are not significantly different from that of the original small sample; this property of imitating the mean of the original small sample makes the S-SMART sample reflect the essential statistic of the original small sample well.

As evident in the Brown and Forsythe’s Tests, with few reasonable exceptions, the S-SMART samples have equal variances across almost all different sizes of replenished samples for each of the different sizes of the original small samples; and the stable S-SMART sample variances are not significantly different from that of the original small sample. The similarity between the homogeneous sample variances of the S-SMART samples and the original small sample makes the S-SMART sample closely mirror the original small sample for its variability. The variation of the S-SMART sample come from two sources: one is from the random errors and the other from the simulation procedure. However, even with the two sources of variation, the S-SMART procedure still produces the amplified samples with the similar variation to that of the original small sample. The stability of the amplified sample together with its robustness to the influence of outliers makes the S-SMART sample representative of the population or local data from which the original small sample is drawn.

The S-SMART Sampling Distribution is Normal

The sampling distribution of the mean of the S-SMART sample is also examined through a series of histograms. The sampling distributions for the S-SMART data from both symmetric and asymmetric distributions are normal as expected from the central limit theorem. The means of the S-SMART samples distribute normally and center around the small sample means. Even though the shape and dispersion depend on the original sample, the variation from sample to sample is not noticeable.

The S-SMART Samples Produce Accurate Confidence Intervals

The confidence intervals for both the mean and standard deviation of the amplified samples produced by the S-SMART method have been studied. The findings in the confidence interval study are promising in that the confidence intervals of the S-SMART samples aptly covered the means. The confidence intervals for the mean of the S-SMART sample are more accurate with the narrower range than the confidence intervals of the original samples for the symmetric and asymmetric distributions. The confidence intervals for the standard deviation of the S-SMART sample are better than those of the original small sample, with the exception of the skewed distribution. It requires more research on the adjustment of the skewness in the S-SMART procedure to make a better estimation for the confidence intervals.
Implications

The findings suggest that the S-SMART procedure has potential to lessen some limitations of the existing resampling methods. First, S-SMART can reduce the influence of outliers, a problem from which the bootstrap method has long suffered. It is known that outliers can severely influence statistics such as mean and variance. Reducing outlier influence can greatly improve the validity of statistical inferences, thus improving the quality of quantitative studies with small samples. Second, the S-SMART sample is the union of the sub-samples randomly generated from the Gaussian kernels centered at the quantiles with a random noise instead of repeatedly selecting resamples from the same data points in the original small sample; therefore, it has independent observations conditionally on the original small sample. Third, the S-SMART procedure produces amplified samples with larger statistical power than the original small sample. As is known, small samples suffer from problems of small statistical power in detecting significant effects of interventions. When only small samples are available, researchers can directly apply the amplified S-SMART sample to statistical analysis in their research to draw more accurate statistical inferences than using the original small sample.

Some researchers may have a concern that the S-SMART method would produce samples with too large power. It is true that researchers can amplify their small sample as large as they wish. However, samples size for any statistical analysis should be determined by a statistical power analysis (Cohen, 1988). The S-SMART is the right tool to help researchers amplify their small samples as large as required by the corresponding statistical power.

In sum, because the S-SMART samples are unbiased in imitating the original sample in terms of distributions and statistical behaviors with less influence of outliers through its robust procedures, they can better represent the population or local data from which the small samples are drawn. With this property the S-SMART samples have the potential to be used for any kind of statistical analysis in quantitative research with small samples.

Limitations

Like all other resampling methods, the S-SMART method relies on how well the small sample represents its population. Because S-SMART produces amplified samples based on the original small samples, if the original small sample is randomly selected and represents its population, S-SMART can produce the corresponding amplified samples representing the population; however, if the original small sample is not randomly selected, the S-SMART can only produce the corresponding amplified samples similar to the original small sample in terms of distribution and other statistical behaviors locally. In this case, the statistical analysis using the S-SMART samples cannot provide reliable statistical inferences to be generalized to the population even though the sample size is amplified. With this concern, researchers should judge the quality of their samples to see if their original small samples are randomly selected so that the statistical results can be generalized to the population; otherwise, statistical results from either the original small sample or the amplified S-SMART sample are only valid locally.

It is also worth noting that the S-SMART method has some restrictions for the sample sizes when estimating the confidence interval of the standard deviation of the data amplified from a skewed population distribution. This limitation requires further investigations. In addition, again like all other resampling methods, the present research of the S-SMART method lacks in-depth mathematical derivations. Adding to the numerical evidence from the simulation study and empirical example, mathematical investigations on the equalities of the means and variances between the S-SMART samples and the corresponding original small sample would make the study of S-SMART stronger.

Further Research

Simulation study on the new resampling method S-SMART has produced promising findings; however, it is desirable to have more mathematical investigations on sample distributions, sampling distribution, sample means, standard errors, and sampling bias. This
study focused on the basic concept, simulation procedures, and verification of the newly-developed S-SMART method; therefore, the simulation study was only conducted over the small samples with univariate data. Besides the current univariate small sample simulation study, investigations with the S-SMART method to amplify multivariate data is in progress.

Future studies could also involve more real-world data to examine how to solve real research questions with the S-SMART samples and thereafter to compare the data analysis results from the S-SMART samples with the results from the other resampling methods. In addition to the above considerations for future studies, it is also desirable to compile a stand-alone computer program package with a user-friendly interface.

References


Nonlinear Parameterization in Bi-Criteria Sample Balancing

Stan Lipovetsky
GfK Custom Research North America

Sample balancing is widely used in applied research to adjust a sample data to achieve better correspondence to Census statistics. The classic Deming-Stephan iterative proportional approach finds the weights of observations by fitting the cross-tables of sample counts to known margins. This work considers a bi-criteria objective for finding weights with maximum possible effective base size. This approach is presented as a ridge regression with the exponential nonlinear parameterization that produces nonnegative weights for sample balancing.

Key words: Sample balance, ridge regression, nonlinear parameterization.

Introduction

Sample balance method was introduced by Deming and Stephan (1940). It is also known in terms of raking or post-stratification, and it is widely used in applied research to adjust sample data to the known proportions in the population. Chi-squared criterion is applied to adjust the counts’ contingency table to the needed margins (Stephan, 1942; Deming, 1964), which yields the weights for observations. The classic method has been developed in numerous approaches (Ireland & Kullback, 1968; Darroch & Ratcliffe, 1972; Holt & Smith, 1979; Feinberg & Meyer, 1983; Little & Wu, 1991; Conklin & Lipovetsky, 2001; Bosch & Wildner, 2003; Kozak & Verma, 2006). The original technique has been further extended, particularly, in calibration and generalized regression (GREG) estimations (Deville & Sarndal, 1992; Sarndal, et al., 1992; Deville, et al., 1993; Sarndal, 1996; Chambers, 1996; Yung & Rao, 2000; Zhang, 2000; Singh, 2003).

Making a sample closer to the required margins, the weighting simultaneously reduces the effective base size of the data. The farther the sample cross-table subtotals are from the margins, the smaller is the effective base in comparison with the original sample size. Decreased effective base produces worse statistical test values and wider confidence intervals around the estimates which can be incorrectly identified as being insignificant. A problem of simultaneous sample balancing with maximization of the effective base was considered in Lipovetsky (2007a), and the solution was obtained in a ridge regression approach (Hoerl & Kennard, 1970, 1988; Lipovetsky, 2006, 2010). Changing the profile ridge parameter yields a better fit of the margins, or a higher effective base, and the trade-off between them is needed: For small ridge parameters corresponding to a better margins fit, some weights could get negative values which are hardly acceptable for applied research.

This article shows how to improve the weights estimation and how to obtain always positive values via nonlinear parameterization for the weights. This approach is presented in the nonlinear optimizing technique for a complex objective and can be reduced to iteratively re-weighted Newton-Raphson procedure (Becker & Le Cun, 1988; Arminger, et al., 1995; Hastie & Tibshirani, 1997; McCullagh & Nelder, 1997; Bender, 2000; Lipovetsky, 2006, 2007b, 2009a,b). The exponential, quadratic and logit parameterizations of the weights are tried. The exponential function is the most convenient for obtaining always nonnegative weights.

Stan Lipovetsky is Senior Research Director at GfK Custom Research North America in Minneapolis, MN. Email him at: stan.lipovetsky@gfk.com.
Methodology

Sample Balancing and Maximizing Effective Base

Let the data be presented in a matrix $X$ of $N$ by $n$ order with elements $x_{ij}$ for an $i^{th}$ observation ($i = 1, 2, \ldots, N$ – number of observations) and a $j^{th}$ variable $x_j$ ($j = 1, 2, \ldots, n$ – number of variables). Besides the design matrix $X$, the required margins are given (census or other totals). Consider $k_j$ bins of given margins for each variable $x_j$, so all the margins can be presented in a vector $y$ of $m$th order, where

$$m = \sum_{j=1}^{n} k_j .$$

Let the variable $x_j$ be measured in the $k_j$ point scale, or the values of $x_j$ are segmented into $k_j$ bins corresponding to the given margins. Each $x_j$ can be categorized by $k_j$ levels, and presented by a set of $k_j$ binary variables. The whole set of these variables can be incorporated into a matrix $Z$ of $N$ by $m$ order. The columns of $Z$ present binary variables $z_{ip}$ with 0-1 values of the elements $z_{ip}$ ($p = 1, 2, \ldots, m$). The matrix $Z$ is singular, because the rank of a matrix of categorized binary variables is not higher than $m-n$.

Deming-Stephan sample balancing consists in fitting the counts $n_l$ in the cross-table (indexed as $l = 1, 2, \ldots, L$) of $Z$ matrix by the theoretical counts $v_l$ in Chi-squared criterion

$$\chi^2 = \sum_{l=1}^{L} \frac{(n_l - v_l)^2}{n_l} ,$$

restricted by the conditions of equivalence of the sample adjusted totals by each variable to the given margins. Adding these restrictions to the objective (1) and minimizing such a conditional objective by the theoretical parameters $v_l$ yields a solution for the weights $w_i$ which can be reached in the algorithm of iterative proportional fitting. Total of the weights equals the sample base, or the weights can be normalized by the relation:

$$\sum_{i=1}^{N} w_i = N .$$

With the weights $w_i$ obtained the Deming-Stephan sample balance procedure, the effective base size of the weighted sample is evaluated by the expression:

$$EB = \left( \sum_{i=1}^{N} w_i \right)^2 / \left( \sum_{i=1}^{N} w_i^2 \right) = N^2 / \left( \sum_{i=1}^{N} w_i^2 \right) ,$$

where the last equality holds only for the normalized weights (2). When the weights are distributed more evenly, closer to 1, the effective base is close to (but always below) the original sample size. Adding and subtracting the constant of the base size, the effective base for any set of weights can be represented as follows:

$$EB = N + \frac{\left( \sum_{i=1}^{N} w_i \right)^2}{\sum_{i=1}^{N} w_i^2} - N

= N - N \frac{1}{\sum_{i=1}^{N} w_i^2} \left( \sum_{i=1}^{N} w_i \right)^2

= N \left( 1 - \frac{\sum_{i=1}^{N} (w_i - \bar{w})^2}{\sum_{i=1}^{N} w_i^2} \right)$$

where $\bar{w}$ is the mean value of the weights. For all weights equal one their mean is $\bar{w} = 1$, so the effective base equals the sample size. Minimization of the weights deviation from their mean corresponds to finding the most effective base (4).

Sample Balancing with Maximum Effective Base

Based on Lipovetsky (2007a), the relation between the given vector of margins $y$ and theoretical $\hat{y}$ vector of margins is presented in a simple linear model:
\[ y = \hat{y} + \epsilon = Z'w + \epsilon. \]  

(5)

The theoretical vector \( \hat{y} = Z'w \) is estimated by the weighted binary variables (prime denotes transposition), where \( w \) is the \( N \)th order vector-column of unknown weights \( w_i \), and \( \epsilon \) is a vector of deviations between the given and theoretical margins. The model (5) reminds an ordinary linear regression – however, with the number \( N \) of the unknown coefficients \( w_i \) significantly larger than the number \( m \) of the values by the dependent variable of margins \( y \). Chi-squared criterion can be applied directly to minimizing the deviations \( \epsilon \) in (5) by fitting the given margins with the weighted binary data:

\[ \chi^2 = \sum_{p=1}^{m} \left( \frac{y_p - \hat{y}_p}{\tilde{z}_p} \right)^2 = \sum_{p=1}^{m} \left( \frac{1}{\tilde{z}_p} \right) \left( y_p - (Z'w)_p \right)^2. \]  

(6)

The notation \( \hat{y}_p \) is used for the elements of the theoretical vector \( \hat{y} \) (5), and \( \tilde{z}_p \) in the denominator (6) are the total counts of the binary variables in the columns of matrix \( Z \), so they are the elements of the vector of \( m \)th order \( \tilde{z} = Z'1_N \), where \( 1_N \) denotes a uniform vector-column of size \( N \).

Simultaneous minimization of the Chi-squared criterion (6) and the efficient variance of the weights in (4) can be achieved by the conditional objective:

\[ F = \chi^2 + q \text{ var}(w) \]

\[ = \sum_{p=1}^{m} \left( 1 \right) \left( y_p - \sum_{i=1}^{N} Z_{ip} w_i \right)^2 + q \sum_{i=1}^{N} \left( w_i - 1 \right)^2 \]

\[ = (y - Z'w)' D^{-1} (y - Z'w) + q (w - 1_N)' (w - 1_N), \]  

(7)

where \( q \) is Lagrange term, \( D \) and \( D^{-1} \) denote the \( m \)th order diagonal matrix and its inversion defined via the total counts:

\[ D = \text{diag}(\tilde{z}), \]

\[ D^{-1} = \text{diag}(1/\tilde{z}). \]  

(8)

The condition for minimization yields a system of linear equations:

\[ \frac{\partial F}{\partial w} = -2ZD^{-1}(y - Z'w) + 2q(w - 1_N) = 0 \]

which is a matrix equation:

\[ \left( ZD^{-1}Z' + qI_N \right)w = ZD^{-1}y + q1_N, \]  

(9)

For \( q \) close to zero this system corresponds to margins fit objective, and with \( q \) growing the main input comes from the efficient base objective with the solution of uniform weights. The equation (10) corresponds to the ridge regression system of equations with the profile parameter \( q \). The regularization item \( qI_N \) added to the diagonal of the matrix in the left-hand side (10) guarantees that it becomes non-singular and invertible.

Solution of the system (10) is given in the work (Lipovetsky, 2007a), and can be presented explicitly as follows:

\[ w = 1_N + Z(Z'Z + q \text{ diag}(\tilde{z}))^{-1}(y - \tilde{z}) \]  

(11)

Due to (11), the weights are distributed around 1, and depend on the difference of the given margins \( y \) and counts \( \tilde{z} = Z'1_N \) by the categorized variables. For \( y - \tilde{z} = 0 \) all the weights are \( w_i = 1 \). A unit change \( \Delta y_p = 1 \) in a \( p \)th component of the vector of margins leads to the weights change equal the elements \( Z(Z'Z + q \text{ diag}(\tilde{z}))^{-1} \) of the \( p \)th column of the transfer matrix, which shows the rate of relaxation of the closeness to the given margins.

Variation in the parameter \( q \) permits a trade-off between better correspondence to the given margins versus more efficient weights of the higher effective base. Dividing the expression (4) by \( N \) yields a quotient \( EB/N \) of the effective to sample base, which is defined as
one minus the ratio of the centered and non-centered weights’ second moments:

\[ R^2_{EB} = \frac{EB}{N} = 1 - \frac{\left( \sum_{i=1}^{N} (w_i - \bar{w})^2 \right)}{\left( \sum_{i=1}^{N} w_i^2 \right)}, \tag{12} \]

The expression (12) has a form of the coefficient of determination \( R^2 \) known in regression analysis, and demonstrates similar properties. If the residual sum of squares in the numerator at the right-hand side (12) is close to zero, \( R^2 \) is close to one, and the effective base reaches the sample base. It is convenient to introduce another coefficient of determination for the margins fitting in Chi-squared objective (6) which also is a weighted least squares objective:

\[ R^2_{\text{mrg}} = 1 - \frac{\chi^2}{\chi^2_{\text{orig}}} \]

\[ = 1 - \frac{\left( \sum_{p=1}^{m} \left( \frac{1}{\bar{z}_p} \right) (y_p - (Z'w)_p)^2 \right)}{\left( \sum_{p=1}^{m} \left( \frac{1}{\bar{z}_p} \right) (y_p - \bar{z}_p)^2 \right)}, \tag{13} \]

where the original value of the objective \( \chi^2_{\text{orig}} \) is taken using the sample counts \( \bar{z} \). Both coefficients \( R^2_{EB} \) and \( R^2_{\text{mrg}} \) can be profiled by the parameter \( q \) for finding an acceptable level of adjustment to margins at a sufficiently large effective base.

Nonlinear Parameterization for Finding Nonnegative Weights

In practice researchers often encounter with the sample total counts too different from the assigned Census margins. Such a discrepancy can easily produce weights with negative values. In these cases the linear ridge-regression solution (11) requires to increase the parameter \( q \) high enough to reach all the weights non-negative. In the ridge regression it is not a problem, but at a price of losing the needed level \( R^2_{\text{mrg}} \) of margins fitting. To obtain positive weights a special parameterization for the weights can be used. For example, the positive weights can be presented by the exponent

\[ w_i = \exp(v_i), \tag{14} \]

or the non-negative weights can be given by the quadratic dependence

\[ w_i = (v_i)^2, \tag{15} \]

where \( v_i \) are the unknown parameters. The logistic parameterization is:

\[ w_i = w_{\text{min}} + \Delta w \frac{1}{1 + \exp(-v_i)}, \tag{16} \]

where \( w_{\text{min}} \) and \( w_{\text{max}} \) are the given constants of the minimum and maximum values of the desired weights. For any \( v_i \), the weights \( w_i \) always belong to the range from \( w_{\text{min}} \) to \( w_{\text{max}} \).

Numerical minimization of the objective (7) by the parameters \( v_i \) of the positive weights can be efficiently performed by Newton-Raphson optimizing technique. Consider the Newton-Raphson algorithm for the objective (7) which can be approximated as:

\[ F(v) \approx F(v^{(0)}) + \frac{\partial F}{\partial v}(v - v^{(0)}), \tag{17} \]

where \( v^{(0)} \) is an initial approximation for the vector \( v \) which consists of the unknown parameters \( v_i \). An extreme value of a function can be found from the condition of the first derivative equals zero, thus taking the derivative of (17) yields:

\[ \frac{dF}{dv} = \frac{\partial^2 F}{\partial v^2}(v - v^{(0)}) + \frac{\partial F}{\partial v} = 0. \tag{18} \]
Solution of the equation (18) for the vector \( v \) is:

\[
v = v^{(0)} - \left( \frac{\partial^2 F}{\partial v^2} \right)^{-1} \left( \frac{\partial F}{\partial v} \right) = v^{(0)} - H^{-1} \nabla F ,
\]

(19)

where a matrix of the second derivatives, or Hessian, is denoted as \( H \), so \( H^{-1} \) is the inverted Hessian, and the vector of the first derivatives is the gradient \( \nabla F \). The obtained expression (19) is used in the iterations for finding each \( (t+1) \)-st approximation for the vector \( v^{(t+1)} \) via the previous vector \( v^{(t)} \) at the \( t \)-th step.

The first derivative of (7) by each parameter \( v_k \) is:

\[
\frac{\partial F}{\partial v_k} = -2 \sum_{p=1}^{m} \sum_{i=1}^{N} z_p^{-1} \left( y_p - \sum_{i=1}^{N} z_p w_i \right) \frac{d w_k}{d v_k} + 2q(w_k - 1),
\]

(20)

which corresponds to the derivative in matrix form (9) multiplied by the derivative of each weight by its parameter. The second derivative by any two parameters \( (r \text{ and } k) \), running by the observations \( i = 1, 2, \ldots, N \) is as follows:

\[
\frac{\partial^2 F}{\partial v_r \partial v_k} = 2 \left( \sum_{p=1}^{m} z_p^{-1} z_{rp} z_{kp} + q \delta_{rk} \right) \frac{d w_r}{d v_r} \frac{d w_k}{d v_k} + \left[ -2 \sum_{p=1}^{m} z_p^{-1} \left( y_p - \sum_{i=1}^{N} z_p w_i \right) z_{kp} + 2q(w_k - 1) \right] \frac{d^2 w_k}{d v_k^2} \delta_{rk} ,
\]

(21)

where \( \delta_{rk} \) is Kronecker delta. Hessian (21) in the braces contains an expression coinciding with that in braces of the first derivatives (20). The first derivative reaches zero at the optimum, therefore Hessian can be reduced to the first part (21) which in matrix notation is:

\[
H = 2G \left( zD^{-1} Z' + qI_N \right) G , \quad G = \text{diag} \left( \frac{d w_i}{d v_i} \right)
\]

(22)

All the notations in (22) are the same as in (5), (8)-(10), and \( G \) denotes the \( N \)-th order diagonal matrix of weight derivatives by the parameters. Vector of the first derivatives (20) can be also represented in matrix notation as:

\[
\nabla F = (-2)G \left( zD^{-1} (y - Z'w) + q(w - 1)I_N \right) .
\]

(23)

Substituting the expressions (22)-(23) into (19) yields the expression for minimization the objective (7):

\[
v = v^{(0)} + G^{-1} \left( zD^{-1} Z' + qI_N \right)^{-1} \left( zD^{-1} (y - Z'w) - q(w - 1)I_N \right)
\]

\[
= v^{(0)} + G^{-1} \left( zD^{-1} Z' + qI_N \right)^{-1} \left( zD^{-1} y + qI_N \right) - G^{-1}w .
\]

(24)

The second item in (24) contains the expression coinciding with the solution of the system (10) which can be denoted as linear solution, \( w_{lin} \), given in explicit form in (11). The recurrent equation (24) for a \( t \)-th and the next steps of approximation can be represented as:

\[
v^{(t+1)} = v^{(t)} + G^{-1} \left( w_{lin} - w^{(t)} \right) .
\]

(25)

Formula (25) presents the iteratively re-weighted Newton-Raphson procedure for minimizing the objective (7) in a nonlinear parameterization, and it usually quickly converges.

For the exponential function (14), the inverted matrix of derivatives (22) is:

\[
G^{-1} = \text{diag} \left( \exp(-v_i^{(t)}) \right) = \text{diag} \left( \frac{1}{w_i^{(t)}} \right) ,
\]

(26)

and for the quadratic function (15) it is:

\[
G^{-1} = \text{diag} \left( \frac{1}{2v_i^{(t)}} \right) = \text{diag} \left( \frac{1}{2 \sqrt{w_i^{(t)}}} \right) .
\]

(27)
For the logistic function (16) its diagonal matrix of the inverted derivatives is:

\[
G^{-1} = \text{diag}^{-1} \left( \Delta w \frac{\exp(-v_i^{(1)})}{(1+\exp(-v_i^{(1)}))^2} \right)
\]

\[
= \text{diag}^{-1} \left( \Delta w \frac{w_i^{(1)} - w_{\min}}{\Delta w} \left( 1 - \frac{w_i^{(1)} - w_{\min}}{\Delta w} \right) \right)
\]

\[
= \text{diag} \left( \frac{w_{\max} - w_{\min}}{(w_{\max} - w_i^{(1)})(w_i^{(1)} - w_{\min})} \right),
\]

(28)

where the constants \( w_{\min} \) and \( w_{\max} \) define the range \( \Delta w \) of the desired weights. With the initial parameters \( v_i^{(0)} = 1/N \), finding the initial weights by the formulae (14)-(16), and the related \( G^{-1} \) matrix by the corresponding formulae (26)-(28), and applying them in (25), it is easy to obtain the next approximation for the parameters, then the nonnegative weights, and to continue the process until it converges.

Numerical Example

Data from a marketing research project of six hundred observations contains variables of gender (two values), income (three levels), age group (three levels), and region (four levels) — these categories are given in the first columns of Table 1. The next two columns in Table 1 present the margins observed in the data and required by Census. Within each variable, a total of the observed or the required margins equal one. For example, the gender splits to 35% and 65%, while it should contain 40% and 60% of males and females, respectively. The largest difference of the sample and population values can be observed by the age groups of 18-34 and 54-65 years old respondents, and by Midwest and West regions.

The next column in Table 1 presents the results of the Deming-Stephan iterative proportional fitting (corresponds to the ridge parameter \( q = 0 \)). All proportions are reached, thus, the fitted margins coincide with the required ones in Table 1 and the coefficient of determination \( R^2_{EB} \) (12) for the effective sample size equals 0.15, so the effective base is reduced by 85% from the sample of 600 observations to the effective base of only 90 observations, which is somewhat low. Descriptive statistics for the obtained weights are given in the last three rows of this column: they show that the weights vary (around mean value equal one) in the wide range from the minimum (\( \min = -1.91 \)) to the maximum (\( \max = 18.29 \)), with the standard deviation (\( \text{std} = 2.42 \)). These results are poor and having negative weights is inconvenient in applied research (most of statistical software modules require the weights to be nonnegative).

Several other columns in Table 1 present the results of the linear ridge regression solutions (11) with the parameter \( q \) running by step 0.25 up to 2.25. Increasing \( q \) results in a loss on the margins adjustment, but a win on the effective sample size. Beginning from \( q = 0.75 \), all the weights become positive and distributed in the narrower range (the standard error reduces twice), and the effective base grows to \( R^2_{EB} = 0.38 \), so it becomes more than twice as large in comparison with the results of \( q = 0 \). Further increasing \( q \) to 1.75, the coefficient of determination for margins and for effective sample size becomes equal to 0.60.

Table 2 presents the results of the exponential parameterization (25)-(26) for the nonnegative weights (14). In difference to linear estimation, the nonlinear approach yields only nonnegative weights with similar characteristics of the quality of margins fit and effective base. The other nonnegative parameterizations (15)-(16) produce similar results to the exponential fitting. The outcomes in the considered example are typical for sample balance with maximizing effective size and nonnegative parameterization for weights.

As mentioned for the formulae (12)-(13), the coefficients of determination \( R^2_{EB} \) and \( R^2_{mrg} \) can be profiled by the growing parameter \( q \) for finding a point of intersection between the declining curve of margins adjustment \( R^2_{mrg} \) and the rising curve \( R^2_{EB} \) of the sufficiently effective base (see Figure 1). Comparison of the
coefficients of determination - $R^2_{EB}$ and $R^2_{mrg}$ - in Tables 1 and 2 and in Figure 1 show that the feasible solutions can be found in the range of $q$ from 0.75 to 1.75.

Table 1: Sample Balance with Maximum Effective Size: Linear Ridge Regression

<table>
<thead>
<tr>
<th>Variable Category</th>
<th>Margins</th>
<th>Ridge Profile Parameter $q$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed Census</td>
<td>0.00</td>
</tr>
<tr>
<td>Gender Male</td>
<td>0.35</td>
<td>0.40</td>
</tr>
<tr>
<td>Gender Female</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td>Income Low</td>
<td>0.44</td>
<td>0.48</td>
</tr>
<tr>
<td>Income Mid</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>Income High</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Age 18-34</td>
<td>0.04</td>
<td>0.32</td>
</tr>
<tr>
<td>Age 35-54</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>Age 54-65</td>
<td>0.55</td>
<td>0.28</td>
</tr>
<tr>
<td>Region Midwest</td>
<td>0.19</td>
<td>0.34</td>
</tr>
<tr>
<td>Region West</td>
<td>0.29</td>
<td>0.13</td>
</tr>
<tr>
<td>Region South</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>Region Northeast</td>
<td>0.17</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>$R^2_{mrg}$</th>
<th>1.00</th>
<th>0.95</th>
<th>0.88</th>
<th>0.81</th>
<th>0.74</th>
<th>0.69</th>
<th>0.64</th>
<th>0.60</th>
<th>0.56</th>
<th>0.53</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2_{EB}$</td>
<td>0.15</td>
<td>0.23</td>
<td>0.31</td>
<td>0.38</td>
<td>0.44</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td>Min</td>
<td>-1.91</td>
<td>-0.60</td>
<td>-0.15</td>
<td>0.09</td>
<td>0.24</td>
<td>0.35</td>
<td>0.43</td>
<td>0.49</td>
<td>0.54</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>Std</td>
<td>2.42</td>
<td>1.81</td>
<td>1.49</td>
<td>1.28</td>
<td>1.12</td>
<td>1.00</td>
<td>0.90</td>
<td>0.82</td>
<td>0.75</td>
<td>0.70</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Sample Balance with Maximum Effective Size: Linear Ridge Regression with Exponential Parameterization of the Coefficients

<table>
<thead>
<tr>
<th>Variable Category</th>
<th>Margins</th>
<th>Ridge Profile Parameter $q$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Census</td>
</tr>
<tr>
<td>Gender Male</td>
<td>0.35</td>
<td>0.40</td>
</tr>
<tr>
<td>Gender Female</td>
<td>0.65</td>
<td>0.60</td>
</tr>
<tr>
<td>Income Low</td>
<td>0.44</td>
<td>0.48</td>
</tr>
<tr>
<td>Income Mid</td>
<td>0.49</td>
<td>0.43</td>
</tr>
<tr>
<td>Income High</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Age 18-34</td>
<td>0.04</td>
<td>0.32</td>
</tr>
<tr>
<td>Age 35-54</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>Age 54-65</td>
<td>0.55</td>
<td>0.28</td>
</tr>
<tr>
<td>Region Midwest</td>
<td>0.19</td>
<td>0.34</td>
</tr>
<tr>
<td>Region West</td>
<td>0.29</td>
<td>0.13</td>
</tr>
<tr>
<td>Region South</td>
<td>0.35</td>
<td>0.33</td>
</tr>
<tr>
<td>Region Northeast</td>
<td>0.17</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Descriptive Statistics

<table>
<thead>
<tr>
<th></th>
<th>$R^2_{\text{mrg}}$</th>
<th>0.96</th>
<th>0.93</th>
<th>0.87</th>
<th>0.81</th>
<th>0.74</th>
<th>0.69</th>
<th>0.64</th>
<th>0.60</th>
<th>0.56</th>
<th>0.53</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2_{\text{EB}}$</td>
<td>0.21</td>
<td>0.25</td>
<td>0.32</td>
<td>0.38</td>
<td>0.44</td>
<td>0.50</td>
<td>0.55</td>
<td>0.60</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Min</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.24</td>
<td>0.35</td>
<td>0.43</td>
<td>0.49</td>
<td>0.54</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Max</td>
<td>15.48</td>
<td>13.02</td>
<td>11.15</td>
<td>9.72</td>
<td>8.60</td>
<td>7.78</td>
<td>7.13</td>
<td>6.59</td>
<td>6.14</td>
<td>5.75</td>
</tr>
<tr>
<td></td>
<td>Std</td>
<td>1.93</td>
<td>1.72</td>
<td>1.47</td>
<td>1.28</td>
<td>1.12</td>
<td>1.00</td>
<td>0.90</td>
<td>0.82</td>
<td>0.75</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Conclusion

This article considers a sample balancing procedure with simultaneous maximization of the effective base size and the parameterization which guarantees the nonnegative weights. A multiple criteria objective is reduced to a ridge regression model (10). The analytical linear solution for the weights (11) is generalized to the nonlinear parameterization of the weights by exponential and other functions (14)-(16). To obtain always nonnegative weights, solution of the nonlinear system of equations is considered in the Newton-Raphson iteratively re-weighted procedure (17)-(28). The suggested weighting scheme is optimal for finding the best margins adjustment with the best effective base size. With growth of the ridge profile parameter $q$, the margins fit (13) decreases while the effective base (12) increases, thus a trade-off between them is used. The suggested approach can serve in solving various practical and theoretical problems involving sample balance for nonnegative weights. For example, the described method can be applied to solving calibration problem for data obtained by different sources or in international market research. The data gathered in several countries by various attributes measured in ordinal scales can be skewed to higher or lower levels due to the cultural norms and specifics dissimilar in different countries. To render the data samples comparable for statistical research one country can be taken as a basic pattern, Census likewise, and its counts of the response distribution can be found by the attributes levels. Fitting each other country distribution to the basic one can be performed exactly by the sample balance procedure which yields a solution for weighting the adjusting data with positive weights.

References


Shrinkage Estimation in the Inverse Rayleigh Distribution

Gyan Prakash
S. N. Medical College, Agra, U. P., India

The properties of the shrinkage test–estimators of the parameter were studied for an inverse Rayleigh model under the asymmetric loss function. Both the single and double–stage shrinkage test–estimators are considered.

Key words: Shrinkage factor; Shrinkage test–estimator; Level of significance; Relative bias; Relative efficiency.

Introduction
If \( x \) is a random variable that follows the inverse Rayleigh distribution with the parameter \( \theta \), then it has the distribution function

\[
F(x; \theta) = \exp \left( -\frac{\theta}{x^2} \right); \ x > 0, \ \theta > 0. \quad (1.1)
\]

If \( x_1, x_2, \ldots, x_n \) is the \( n \) random observations drawn from model (1.1), then the maximum likelihood estimator (MLE) and the unbiased estimator of \( \theta \) are given respectively as

\[
\hat{\theta}_{ML} = \frac{n}{T} \quad \text{and} \quad \hat{\theta}_U = \frac{n - 1}{T}. \quad (1.2)
\]

Here, \( T = \sum_{i=1}^{n} \frac{1}{x_i^2} \) is a sufficient statistic for the parameter \( \theta \).

In the estimation problem when positive and negative errors have different consequences, the use of SELF (Squared error loss function) is not appropriate. Varian (1975) discussed an asymmetric loss function known as the LINEX loss function (LLF). This loss function is convex and its shape is determined by the value of its shape parameter. The positive (negative) values of the shape parameter, gives more weight to underestimation (overestimation). Further, the magnitude of the shape parameter reflects the degree of asymmetry. The invariant form of the LLF is defined as

\[
L(\Delta) = \left\{ e^{a\Delta} - a\Delta - 1 \right\}; \ a \neq 0
\]

and

\[
\Delta = \left( \frac{\hat{\theta}}{\theta} - 1 \right). \quad (1.3)
\]

Here 'a' is the shape parameter of the LLF and \( \hat{\theta} \) is any estimate of the parameter \( \theta \). When \( a > 0 \), the loss function increases almost exponentially for positive \( \Delta \) and almost linearly otherwise and overestimation is more heavily penalized than underestimation. When \( a < 0 \) the linear exponential rises are interchanged and underestimation is considered more costly than overestimation. The LINEX loss function may be considered a natural extension of SELF (for small values of 'a' (near to zero) the LINEX loss function tends to SELF). Srivastava and Tanna (2001), Xu and Shi (2004), Prakash and Singh (2006), Singh, et al. (2007), Prakash and Singh (2009) and others have discussed estimation procedures under LLF.

In many situations, the experimenter has some prior information about the parameter in the form of a point or guess value and it is recognized that a shrinkage estimator performs better if a guess value of the parameter is approximately the true value and the sample size

Gyan Prakash is in the Department of S. P. M. Email him at: ggyanji@yahoo.com.
is small. Thompson (1968), Mehta and Srinivasan (1971), Pandey and Singh (1977), Pandey (1979), Singh, et al. (1996), Singh, et al. (2007) and others have suggested shrinkage estimators utilizing the point guess value of the parameter.

The study is presented for the single and double stage shrinkage test–estimators for the parameter \( \theta \) under the LLF.

**Methodology**

Proposed Class of Estimator for the Parameter \( \theta \)

The proposed class of estimator for the parameter \( \theta \) is defined as

\[
\theta_c = C \hat{\theta}_U = C \frac{n-1}{T}; C \in \mathbb{R}^+.
\] (2.1)

The value of constant \( C = \hat{C} \) (for example), which minimizes the risk of \( \theta_c \) under the LLF, is obtained by solving the given equality numerically

\[
\frac{e^a}{n-1} = I \left( 0, \infty, \frac{1}{z} \left( a \frac{n-1}{z} \right) \right),
\] (2.2)

where

\[
I \left( p, q, \omega \right) = \frac{1}{\Gamma_n} \int_p^q (\omega) e^{-z} z^{n-1} \text{d} z
\]

and \( \omega \) is the function of \( z \).

Thus, the improved class of estimator of \( \theta \) in the class (2.1) is

\[
\hat{\theta}_c = \hat{C} \hat{\theta}_U
\] (2.3)

with the risk under the LLF

\[
R \left( \hat{\theta}_c \right) = e^{-a} I \left( 0, \infty, e^{s\Delta_0} \right) + a \left( 1 - \hat{C} \right) - 1;
\]

\[
\Delta_0 = \hat{C} \frac{n-1}{z}.
\] (2.4)

Proposed Shrinkage Estimator and its Properties

Following Thompson (1968), a shrinkage estimator for the parameter \( \theta \) when \( \theta_0 \), a guess value of \( \theta \) is available, is defined as

\[
\theta_{sh} = k_1 \left( \hat{\theta}_U - \theta_0 \right) + \theta_0; \ k_1 \in [0,1].
\] (3.1)

Depending on the guessed value \( \theta_0 \) used, a shrinkage factor \( k_1 \) is specified. The shrinkage procedure has been applied to a number of different problems, a few examples include: mean survival time in epidemiological studies (Harries & Shakarki, 1979), forecasting money supplies (Tso, 1990), estimating mortality rates (Marshall, 1991) and improving estimation in sample surveys (Wooff, 1985).

The risk under the LLF (1.3) for the shrinkage estimator \( \theta_{sh} \) is given by

\[
R \left( \hat{\theta}_{sh} \right) = e^{a(\delta-1)} I \left( 0, \infty, e^{s\Delta_1} \right) + a(1-\delta)
\]

\[
(1-k_1)-1,
\] (3.2)

where

\[
\Delta_1 = k_1 \left( \frac{n-1}{z} - \delta \right) \text{ and } \delta = \frac{\theta_0}{\theta}.
\]

The value of \( k_1 = k_2 \) (for example) that minimizes \( R \left( \hat{\theta}_{sh} \right) \), is also obtained by solving the given equality numerically:

\[
(1-\delta)e^{a(1-\delta)} = I \left( 0, \infty, \frac{\Delta_1}{k_1} e^{s\Delta_1} \right). \] (3.3)

Therefore, the improved shrinkage estimator for \( \theta \) in the class (3.1) is

\[
\hat{\theta}_{sh} = k_2 \left( \hat{\theta}_U - \theta_0 \right) + \theta_0.
\] (3.4)

The expressions of the relative bias and the risk under the LLF are obtained as

\[
\text{RB} \left( \hat{\theta}_{sh} \right) = \frac{1}{\theta} E \left( \hat{\theta}_{sh} \right) - 1 = (1-\delta)(k_1 - 1)
\] (3.5)

and
The expression of relative bias of $\hat{\theta}_{SH}$ clearly shows that the relative bias is zero at $\delta = 1.00$ and has a tendency of being negative for $0 < \delta < 1.00$ and positive otherwise.

The relative efficiency for the shrinkage estimator $\hat{\theta}_{SH}$ with respect to the improved estimator $\hat{\theta}_C$ is defined as

$$RE\left(\hat{\theta}_{SH}, \hat{\theta}_C\right) = \frac{R\left(\hat{\theta}_C\right)}{R\left(\hat{\theta}_{SH}\right)}.$$ 

The expression $RE\left(\hat{\theta}_{SH}, \hat{\theta}_C\right)$ involves $\delta$, $a$, and $n$. For the selected set of values of $\delta = 0.40$ (0.20) 1.60; $a = 0.25$, 0.50, 1.00 and $n = 04$, 08, 12, 15, the numerical findings of the relative efficiency are presented in Table 1 for $a = 0.25$, 0.50.

Based on the values in the table, it may be concluded that the shrinkage estimator $\hat{\theta}_{SH}$ performs better than the estimator $\hat{\theta}_C$ for the considered set of the parametric space and attains maximum efficiency at the point $\delta = 1.00$. Also, the efficiencies increase (decrease) for $a (n)$ increases when other parametric values are fixed (except $\delta = 1.00$).

### The Shrinkage Test–Estimators and their Properties

It has been shown that the shrinkage estimator $\hat{\theta}_{SH}$ has a lower risk than the improved estimator $\hat{\theta}_C$ when a guess value $\theta_0$ of $\theta$ is near to the true value of the parameter $\theta$.

Thus, the shrinkage test–estimator is proposed for testing the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ based on a given set of data. The test statistic $2 \theta_0^T \sim \chi^2_{(2n)}$ is used for testing $H_0$. If $\alpha$ is the level of significance then the null hypothesis $H_0$ is not rejected if $1 - \alpha = P \left[ m_1 \leq 2 \theta_0^T \leq m_2 \right]$.
Thus, the proposed shrinkage test–estimators are

\[
\hat{\theta}_{SHi} = \hat{\theta}_C + \left((1-k_i)\theta_0 + \left(k_i - \hat{C}\right)\hat{\theta}_U\right)I_{(t_i \leq T \leq t_2)};
\]

(4.1)

where \( I_{(A)} \) denotes the indicator of \( A \), \( t_i = \frac{m_i}{2\theta_0} \) and \( i = 1, 2 \). Here \( m_1 \) and \( m_2 \) are the values of the lower and upper 100\(\alpha/2\)% points of the Chi–square distribution with 2\(n\) degrees of freedom.

The expression of the relative bias is obtained as:

\[
\text{RB}(\hat{\theta}_{SHi}) = I(y_1, y_2, \Delta'') + \hat{C} - 1;
\]

(4.2)

where \( \Delta'' = (\Delta - \Delta_0 + \delta) \), \( y_i = \frac{m_i}{2\delta} \) and \( i = 1, 2 \).

Similarly, the expressions of the risk under the LLF for the proposed shrinkage test–estimators are

\[
\text{RE}(\hat{\theta}_{SHi}, \hat{\theta}_C) = \text{RE}(\hat{\theta}_C) / \text{RE}(\hat{\theta}_{SHi}) \;
\]

(4.3)

The value of \( k_i = k_3 \) (for example) that minimizes the risk of the shrinkage test–estimator \( \hat{\theta}_{SHi} \) may be obtained by solving the following equality

\[
e^{a(\delta)}I\left(y_1, y_2, \left(\Delta_{i} / k_i\right)\right) = I\left(y_1, y_2, \left(\Delta_{i} e^{a\delta} / k_i\right)\right).
\]

(4.4)

Hence, the improved shrinkage test–estimator is defined as

\[
\hat{\theta}_{SH3} = \hat{\theta}_C + \left((1-k_3)\theta_0 + \left(k_3 - \hat{C}\right)\hat{\theta}_U\right)I_{(t_1 \leq T \leq t_2)}
\]

(4.5)

The expressions of the relative bias and the risk under the LLF are given as

\[
\text{RB}(\hat{\theta}_{SH3}) = I(y_1, y_2, \Delta'') + \hat{C} - 1;
\]

\[
\Delta'' = (\Delta_3 - \Delta_0 + \delta)
\]

and

\[
\text{RE}(\hat{\theta}_{SH3}, \hat{\theta}_C) = e^{a(\delta)}I\left(y_1, y_2, e^{a\delta}\right)
\]

\[
- e^{-a}I\left(y_1, y_2, e^{a\delta}\right)
\]

\[
- aI\left(y_1, y_2, \Delta''\right)
\]

\[
+ e^{-a}I\left(0, \infty, e^{a\delta}\right) + a(1-\hat{C}) - 1.
\]

(4.6)

The relative efficiency of the shrinkage test–estimator \( \hat{\theta}_{SHi} \); \( i = 1, 2, 3 \), with respect to improved estimator \( \hat{\theta}_C \) is defined as

\[
\text{RE}(\hat{\theta}_{SHi}, \hat{\theta}_C) = \text{RE}(\hat{\theta}_C) / \text{RE}(\hat{\theta}_{SHi}); \; i = 1, 2, 3.
\]

The relative bias \( \text{RB}(\hat{\theta}_{SHi}) \) and the relative efficiency \( \text{RE}(\hat{\theta}_{SHi}, \hat{\theta}_C) \) are the functions of \( \delta, \; k, \; \alpha, \; a \) and \( n \). For a similar set of values as considered previously with \( k = 0.25, 0.50, 0.75 \) and \( \alpha = 0.01, 0.05 \), the relative bias (not presented) and the relative efficiency are presented in Table 2, for \( n = 08 \) and 12.

The relative biases are negligibly small and lie between -0.014 and 0.019. For small values of \( \delta \leq 1.00 \), the relative bias is negative but for large \( \delta \) it has a tendency to be positive. The value of the absolute relative bias (ARB) decreases as \( n \) increases for \( \delta \geq 1.00 \) when other parametric values are fixed. The ARB increases as \( a \) (\( \alpha \)) increases for small \( \delta \leq 1.00 \).
and decreases otherwise. In addition, the ARB decreases when \( k \) increases for the all considered values of \( \delta \) when other parametric values are fixed.

The shrinkage test–estimator \( \hat{\theta}_{SH1} \) has smaller risk than \( \hat{\theta}_C \) for the all considered values of the parametric space. The efficiency decreases as \( 'a' \) or \( k \) increases in the region \( 0.40 \leq \delta \leq 1.20 \) when other parametric values are fixed and the efficiency attains maximum at the point \( \delta = 1.00 \). In addition, as the level of significance \( \alpha \) increases, the efficiency decreases for the all considered values of \( \delta \).

The expressions of the relative bias and the relative efficiency for the test–estimator \( SH_i \theta \hat{} \), \( i = 2, 3 \) are the functions of \( \delta \), \( \alpha \), \( a \) and \( n \). For a similar set of values as considered earlier, the relative biases (not presented here) and the relative efficiencies are shown in Tables 3 and 4.

The relative biases of \( \hat{\theta}_{SH2} \) are negligibly small and lie between \(-0.017 \) and \( 0.029 \). For small values of \( \delta \leq 1.00 \), the relative bias is negative, but for large \( \delta \) it has a tendency of being positive. The ARB increases as \( a(\alpha) \) increases for small \( \delta \leq 1.00 \) and decreases otherwise. The relative biases of \( \hat{\theta}_{SH3} \) are also negligibly small and lie between \(-0.018 \) and \( 0.031 \). Other properties are similar to shrinkage test–estimator \( \hat{\theta}_{SH2} \).

The shrinkage test–estimator \( \hat{\theta}_{SH2} \) performs well with respect to \( \hat{\theta}_C \) for the all considered parametric values and attains maximum efficiency at the point \( \delta = 1.00 \) (Table 3). The efficiency decreases as \( 'a' \) increases when \( \delta \leq 1.00 \) for other fixed parametric values. This decreasing trend has also been observed when \( \alpha \) increases for all considered values of \( \delta \).

Table 4 shows that the shrinkage test–estimator \( \hat{\theta}_{SH3} \) performs uniformly well with respect to \( \hat{\theta}_C \) for the all considered parametric values. The efficiency decreases as \( n \) increases in the region \( 0.80 \leq \delta \leq 1.40 \) for other fixed parametric values. Other properties are observed to be similar to the shrinkage test–estimator \( \hat{\theta}_{SH2} \).

The Double–Stage Shrinkage Test–Estimator

A double–stage procedure using prior information in the form of an initial estimate or a guessed value has been considered by many authors (Katti, 1962; Shah, 1964; Waikar & Katti, 1971; Al–Bayyati & Arnold, 1972; Waikar, et al., 1984; Adke, et al., 1987). Arnold & Al–Bayyati (1970) considered the double–stage shrinkage estimator for the mean of a normal population when a prior guessed value of the mean is available. Pandey, et al. (1988) proposed some shrinkage estimators for the variance of a Normal distribution at double–stage under mean square error criterion.

Let \( x_{ji} (i = 1, 2, \ldots, n_j) ; j = 1, 2 \) be two random samples of size \( n_1 \) and \( n_2 \) respectively, drawn independently from the model (1.1) with the parameter \( \theta \). The pooled unbiased estimate of \( \theta \) based on two samples of size \( n_1 \) and \( n_2 \) is

\[
\hat{\theta}_p = \frac{(n_1 - 1)T_2 + (n_2 - 1)T_1}{2T_1T_2}.
\]

The proposed class of estimators for the pooled estimate of \( \theta \) is given by

\[
\theta_{PC} = l\theta_p ; l \in \mathbb{R}^+.
\]

The value of \( l = \hat{l} \) (for example), for which \( R(\theta_{PC}) \) is minimum is obtained by simplifying the given equality numerically

\[
2e^a = G \left( 0, \infty, 0, \infty, \left( \frac{\Delta'}{l} e^{a\delta_0} \right) \right),
\]

where
Table 2: Relative Efficiency for the Shrinkage Test-Estimator $\hat{\theta}_{SH1}$ with respect to $\hat{\theta}_C$

for $n = 08$ and $12$

<table>
<thead>
<tr>
<th>$n = 08$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$a$</td>
</tr>
<tr>
<td>0.01</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
</tr>
<tr>
<td>0.50</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
</tr>
</tbody>
</table>

| $n = 12$ |
|---|---|
| $\alpha$ | $a$ | $k$ | 0.40 | 0.60 | 0.80 | 1.00 | 1.20 | 1.40 | 1.60 |
| 0.01 | 0.25 | 0.25 | 1.6316 | 2.1191 | 2.3444 | 3.3353 | 2.5675 | 1.1708 | 1.1617 |
| | 0.50 | 1.6011 | 2.1061 | 2.0147 | 2.5170 | 2.2819 | 1.6603 | 1.0200 |
| | 0.75 | 1.4517 | 1.5849 | 1.4843 | 1.7736 | 1.6626 | 1.6518 | 1.4559 |
| 0.50 | 0.25 | 1.5354 | 1.6908 | 1.7535 | 2.8634 | 2.4524 | 1.1878 | 1.1736 |
| | 0.50 | 1.5087 | 1.6781 | 1.6401 | 2.2561 | 2.1817 | 1.6498 | 1.0498 |
| | 0.75 | 1.3727 | 1.1940 | 1.2765 | 1.6437 | 1.6092 | 1.6217 | 1.4366 |
| 0.05 | 0.25 | 1.4620 | 1.9624 | 2.2676 | 2.5836 | 1.8173 | 1.0414 | 1.1474 |
| | 0.50 | 1.4437 | 1.8656 | 2.0132 | 2.0955 | 1.6523 | 1.3561 | 1.0191 |
| | 0.75 | 1.3858 | 1.5181 | 1.4754 | 1.5833 | 1.3628 | 1.3994 | 1.3366 |
| 0.50 | 0.25 | 1.4319 | 1.6854 | 1.7453 | 2.0929 | 1.6666 | 1.0416 | 1.0582 |
| | 0.50 | 1.4145 | 1.6152 | 1.5948 | 1.7697 | 1.5237 | 1.3307 | 1.0478 |
| | 0.75 | 1.3557 | 1.1478 | 1.2402 | 1.3909 | 1.2701 | 1.3559 | 1.3211 |
Table 3: Relative efficiency for the Shrinkage Test-Estimator $\hat{\theta}_{SH2}$ with respect to $\hat{\theta}_C$

<table>
<thead>
<tr>
<th>n = 04</th>
<th>$\alpha$</th>
<th>$a$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1.6548</td>
<td>2.8665</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.5454</td>
<td>2.3551</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.5401</td>
<td>2.1247</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.25</td>
<td>1.5548</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.4454</td>
<td>1.6955</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.4402</td>
<td>1.6507</td>
</tr>
<tr>
<td>$n = 08$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1.6543</td>
<td>1.9973</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.4361</td>
<td>1.4759</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.3288</td>
<td>1.0766</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.25</td>
<td>1.5502</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.3626</td>
<td>1.4470</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.3191</td>
<td>1.0424</td>
</tr>
<tr>
<td>$n = 12$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1.4375</td>
<td>1.8921</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.3713</td>
<td>1.5575</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.3244</td>
<td>1.3924</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.25</td>
<td>1.3801</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.3551</td>
<td>1.5169</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.3159</td>
<td>1.3391</td>
</tr>
<tr>
<td>$n = 15$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>1.3716</td>
<td>1.7533</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.3628</td>
<td>1.6522</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.3114</td>
<td>1.3493</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.25</td>
<td>1.3486</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.3466</td>
<td>1.5130</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>1.2355</td>
<td>1.3321</td>
</tr>
</tbody>
</table>
## SHRINKAGE ESTIMATION IN THE INVERSE RAYLEIGH DISTRIBUTION

Table 4: Relative efficiency for the Shrinkage Test-Estimator $\hat{\theta}_{S13}$ with respect to $\hat{\theta}_C$

<table>
<thead>
<tr>
<th>n = 04</th>
<th>( \delta )</th>
<th>( \alpha )</th>
<th>a</th>
<th>( 0.40 )</th>
<th>( 0.60 )</th>
<th>( 0.80 )</th>
<th>( 1.00 )</th>
<th>( 1.20 )</th>
<th>( 1.40 )</th>
<th>( 1.60 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td></td>
<td>0.25</td>
<td>1.6548</td>
<td>2.9751</td>
<td>4.9163</td>
<td>12.671</td>
<td>6.9280</td>
<td>3.3047</td>
<td>1.9473</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.5454</td>
<td>2.4443</td>
<td>4.2197</td>
<td>10.744</td>
<td>3.7855</td>
<td>2.2037</td>
<td>1.7596</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.5401</td>
<td>2.2052</td>
<td>3.9810</td>
<td>10.433</td>
<td>6.2141</td>
<td>3.3629</td>
<td>1.8926</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>0.25</td>
<td>1.5548</td>
<td>1.8025</td>
<td>2.6512</td>
<td>5.3498</td>
<td>3.8165</td>
<td>2.6166</td>
<td>1.7863</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.4454</td>
<td>1.7597</td>
<td>2.3790</td>
<td>4.5038</td>
<td>2.5468</td>
<td>2.1138</td>
<td>1.7452</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.4402</td>
<td>1.7132</td>
<td>2.3442</td>
<td>4.3828</td>
<td>3.2998</td>
<td>3.1332</td>
<td>1.7381</td>
<td></td>
</tr>
<tr>
<td>n = 08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>0.25</td>
<td>1.6543</td>
<td>2.073</td>
<td>2.5476</td>
<td>6.5306</td>
<td>4.0401</td>
<td>2.1967</td>
<td>1.5681</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.4361</td>
<td>1.5318</td>
<td>2.3866</td>
<td>6.3226</td>
<td>2.9271</td>
<td>2.1739</td>
<td>1.6250</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.3288</td>
<td>1.1174</td>
<td>2.0052</td>
<td>6.0281</td>
<td>4.6093</td>
<td>2.9885</td>
<td>2.5328</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>0.25</td>
<td>1.5502</td>
<td>2.0625</td>
<td>2.3939</td>
<td>3.6278</td>
<td>2.5242</td>
<td>1.7899</td>
<td>1.4518</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.3626</td>
<td>1.5018</td>
<td>1.7946</td>
<td>3.3737</td>
<td>2.3620</td>
<td>1.8597</td>
<td>1.5200</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.3191</td>
<td>1.0819</td>
<td>1.6500</td>
<td>3.2957</td>
<td>2.4355</td>
<td>2.3905</td>
<td>1.1507</td>
<td></td>
</tr>
<tr>
<td>n = 12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>0.25</td>
<td>1.4375</td>
<td>1.9638</td>
<td>2.4524</td>
<td>4.1762</td>
<td>2.6656</td>
<td>1.8128</td>
<td>1.4662</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.3713</td>
<td>1.6165</td>
<td>1.8823</td>
<td>4.1094</td>
<td>2.5449</td>
<td>1.7856</td>
<td>1.4438</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.3244</td>
<td>1.4451</td>
<td>1.8094</td>
<td>3.9931</td>
<td>3.8208</td>
<td>2.4243</td>
<td>1.1731</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>0.25</td>
<td>1.3801</td>
<td>1.7792</td>
<td>2.3494</td>
<td>3.1007</td>
<td>1.8756</td>
<td>1.4805</td>
<td>1.3546</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.3551</td>
<td>1.5743</td>
<td>1.7784</td>
<td>2.9641</td>
<td>1.7227</td>
<td>1.4395</td>
<td>1.3322</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.3159</td>
<td>1.3898</td>
<td>1.5259</td>
<td>2.6987</td>
<td>2.3746</td>
<td>2.3645</td>
<td>1.1685</td>
<td></td>
</tr>
<tr>
<td>n = 15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td></td>
<td>0.25</td>
<td>1.3716</td>
<td>1.8197</td>
<td>2.3829</td>
<td>4.0996</td>
<td>2.1802</td>
<td>1.6663</td>
<td>1.4303</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.3628</td>
<td>1.7148</td>
<td>1.7636</td>
<td>4.0539</td>
<td>2.0954</td>
<td>1.6081</td>
<td>1.3781</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.3114</td>
<td>1.4004</td>
<td>1.6683</td>
<td>3.5093</td>
<td>3.2848</td>
<td>2.1105</td>
<td>1.2831</td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>0.25</td>
<td>1.3486</td>
<td>1.6060</td>
<td>2.3417</td>
<td>3.0588</td>
<td>1.6628</td>
<td>1.3632</td>
<td>1.3179</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
<td>1.3466</td>
<td>1.5703</td>
<td>1.7069</td>
<td>2.8050</td>
<td>1.5510</td>
<td>1.3023</td>
<td>1.2696</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.00</td>
<td>1.2355</td>
<td>1.3825</td>
<td>1.4490</td>
<td>2.6255</td>
<td>2.2484</td>
<td>2.1063</td>
<td>1.1815</td>
<td></td>
</tr>
</tbody>
</table>
PRAKASH

\[ G(p_1, q_1, p_2, q_2, \omega) = \frac{1}{\Gamma n_1 \Gamma n_2} \int_{z_1 \leq p_1} \int_{z_2 \leq p_2} (\omega) e^{-z_1} z_1^{n_1-1} e^{-z_2} z_2^{n_2-1} \, dz_1 \, dz_2 , \]

\[ \Delta' = \frac{1}{2} \left( \frac{n_1-1}{z_1} - \frac{n_2-1}{z_2} \right) \] and \( \omega \) may be the function of \( Z_1 \) and \( Z_2 \).

Thus, the improved pooled estimator among the class (5.2) is

\[ \hat{\theta}_{PC} = \hat{\theta}_p \] (5.4)

with the risk

\[ R\left( \hat{\theta}_{PC} \right) = e^{-\alpha} G(0, \infty, 0, \infty, e^{\alpha a_1}) + a \left( 1 - I \right) - 1, \] (5.5)

where

\[ \Delta = \frac{1}{2} \left( \frac{n_1-1}{z_1} - \frac{n_2-1}{z_2} \right). \]

The performances of the shrinkage test–estimator \( \hat{\theta}_{SH3} \) are better in terms of the magnitude of efficiency when they are compared with \( \hat{\theta}_{SH2} \). Hence, \( \hat{\theta}_{SH3} \) has been considered in double–stage setup. The proposed double–stage shrinkage test–estimator is given as

\[ \hat{\theta}_{DSH} = \hat{\theta}_{PC} + \left( 1 - k_3 \right) \hat{\theta}_0 + k_3 \hat{\theta}_{U1} - \hat{\theta}_{PC} \]

\[ I_{(t_1 \leq T \leq t_2)} ; \hat{\theta}_{U1} = \frac{n_1 - 1}{T_1} . \]

The proposed double–stage technique is to first obtain a sample size \( n_1 \) and compute \( \hat{\theta}_{U1} \). If \( \hat{\theta}_{U1} \) implies that the prior estimate \( \theta_0 \) was reasonable, the sampling is stopped and the parameter is estimated with the help of a shrinkage estimator. Otherwise, \( n_2 \) additional observations are obtained and used to improve the estimate based on all \((n_1 + n_2)\) observations. The risk under the LLF for \( \hat{\theta}_{DSH} \) is obtained as

\[ R\left( \hat{\theta}_{DSH} \right) = e^{-\alpha} I \left( y_1, y_2, e^{\alpha a_1} \right) - e^{-\alpha} G(0, \infty, 0, \infty, e^{\alpha a_1}) - a l \left( y_1, y_2, \Delta' \right) + e^{-\alpha} G(0, \infty, 0, \infty, e^{\alpha a_1}) + a \left( 1 - I \right) - 1, \]

where

\[ \Delta' = \frac{1}{2} \left( \frac{n_1-1}{z_1} + \frac{1}{z_1} \right) + \delta. \]

The problem is considered as a sequential estimation problem with stopping random variable \( N \) defined as

\[ \begin{cases} n_1 & \text{if } t_1 \leq T \leq t_2 \\ n_1 + n_2 & \text{otherwise.} \end{cases} \] (5.6)

If a cost \( d > 0 \) is introduced for each observation. Then the risk of \( \hat{\theta}_{DSH} \) is:

\[ \tilde{R}\left( \hat{\theta}_{DSH} \right) = R\left( \hat{\theta}_{DSH} \right) + d \, E(N) \]

Similarly the risk of \( \hat{\theta}_{PC} \) is:

\[ \tilde{R}\left( \hat{\theta}_{PC} \right) = R\left( \hat{\theta}_{PC} \right) + d (n_1 + n_2) \]

Therefore, the relative efficiency of \( \hat{\theta}_{DSH} \) with respect to \( \hat{\theta}_{PC} \) is given by:

\[ \text{RE}\left( \hat{\theta}_{DSH}, \hat{\theta}_{PC} \right) = \frac{\tilde{R}\left( \hat{\theta}_{PC} \right)}{R\left( \hat{\theta}_{DSH} \right)}. \]

The function of the relative efficiency involves \( n_1, n_2, \delta, a, \alpha \) and per unit cost \( d \). For a similar set of selected values as considered
previously with \( n_2 = 0.04, 0.08 \) and \( d = 0.50, 05, 10, 50 \), calculated relative efficiencies are presented for \( n_1 = 0.04, 0.08 \) and \( d = 0.50 \) in Table 5.

The double-stage shrinkage test–estimator \( \hat{\theta}_{\text{DSH}} \) performs well with respect to improved pooled estimator \( \hat{\theta}_{\text{PC}} \) for the all considered parametric set of values and attains maximum efficiency at the point \( \delta = 1.00 \). The efficiency decreases as \( \alpha(n) \) and increases for all \( \delta \) when other parametric values are fixed. The decreasing trend was observed when \( n_2 \) increased for all considered values of \( \delta \). The nominal loss was recorded when per unit cost increased but the effective interval did not alter.

Conclusion
Based on the data presented, the performances of both the shrinkage test–estimators are uniformly well respect to the improved estimator \( \hat{\theta}_{\text{C}} \) for the considered parametric set of values.

Based on the gain in efficiency, \( \hat{\theta}_{\text{SH2}} \) may be preferred over \( \hat{\theta}_{\text{SH1}} \) in the region \( 0.60 \leq \delta \leq 1.40 \). The double–stage shrinkage test–estimator \( \hat{\theta}_{\text{DSH}} \) performs well with respect to improved pooled estimator \( \hat{\theta}_{\text{PC}} \) for the all considered parametric set of values.

References


Table 5: Relative efficiency for the Shrinkage Test-Estimator $\hat{\theta}_{\text{DSH}}$ with respect to $\hat{\theta}_{\text{PC}}$
for $n_{1} = 0.4, 0.8$ and $d = 0.50$

<table>
<thead>
<tr>
<th>$n_{1} = 0.4$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$n_{2}$</td>
</tr>
<tr>
<td>0.01</td>
<td>04</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>08</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>04</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>08</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>$n_{1} = 0.8$</td>
<td>$\delta$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$n_{2}$</td>
</tr>
<tr>
<td>0.01</td>
<td>04</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>08</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>04</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>08</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Combining Independent Tests of Conditional Shifted Exponential Distribution

Abedel-Qader S. Al-Masri
Yarmouk University,
Irbid, Jordan

The problem of combining n independent tests as \( n \to \infty \) for testing that variables are uniformly distributed over the interval (0, 1) compared to their having a conditional shifted exponential distribution with probability density function \( f(x; \theta) = e^{-(x+\theta)}, \ x \geq \gamma \theta, \ \theta \in [a, \infty), a \geq 0 \) was studied. This was examined for the case where \( \theta_1, \theta_2, \ldots \) are distributed according to the distribution function (DF) F and when the DF is Gamma (1, 2). Six omnibus methods were compared via the Bahadur efficiency. It is shown that, as \( \gamma \to 0 \) and \( \gamma \to \infty \), the inverse normal method is the best among the methods studied.

Key words: Conditional shifted exponential, combining independent tests, omnibus methods, Bahadur efficiency.

Introduction
The combination of independent tests of hypothesis is an important and a popular statistical practice. Many methods are available to use to combine independent tests; these methods are compared by using different criteria including Exact Bahadur Slope (EBS), Approximate Bahadur Slope (ABS), Pitman Efficiency, Local Power, Admissibility and others. If \( H_0 \) is a simple hypothesis, Birnbaum (1955) showed that, for given any non-parametric combination method with a monotone increasing acceptance region, there exists a problem for which this method is most powerful against some alternative. Littell and Folks (1971) studied four methods of combining a finite number of independent tests. They found that the Fisher method is better than the inverse normal method, the minimum of p-value method and maximum of p-values via Bahadur efficiency. Later, Littell and Folks (1973) showed under mild conditions that the Fisher’s method is optimal among all methods for combining a finite number of independent tests. Brown, Cohen and Strawderman (1976) have shown that such tests form a complete class.

The Specific Problem
Consider n hypotheses of the form:

\[
H_0^{(i)} : \eta_i = \eta_o^{i} \quad \text{vs} \quad H_1^{(i)} : \eta_i \in \Omega_i - \{\eta_o^{i}\}
\]

such that \( H_0^{(i)} \) is rejected for large values, \( i = 1, 2, \ldots, n \) of some continuous random variable \( T^{(i)} \). The n hypotheses are combined into one as

\[
H_0 : (\eta_1, \ldots, \eta_n) = (\eta_o^1, \ldots, \eta_o^n) \quad \text{vs} \quad H_1 : (\eta_1, \ldots, \eta_n) \in \left\{ \prod_{i=1}^{n} \Omega_i - \{\eta_o^1, \ldots, \eta_o^n\} \right\}
\]

For \( i = 1, 2, \ldots, n \) the p-value of the \( i \)th test is given by

\[
P_i(t) = P_{H_0^{(i)}}(T^{(i)} > t) = 1 - F_i(t)
\]
COMBINING INDEPENDENT TESTS OF EXPONENTIAL DISTRIBUTION

where $F$ is the DF of $T^{(i)}$ under $H^{(i)}_o$. Note that under $H^{(i)}_o$ the random variable $P_i \sim U(0, 1)$ under $H^{(i)}_1$ has some distribution that is not $U(0, 1)$.

If considering the special case where $\eta_i = 0$ and $\eta_i = \theta_o$ for $i = 1, \ldots, n$, and also assume that $T^{(1)}, \ldots, T^{(n)}$ are independent, then (1) reduces to

\[ H_o : \theta = \theta_o \]

\[ \text{vs} \]

\[ H_1 : \theta \in \Omega \setminus \{\theta_o\}. \]

It follows that the p-values $P_1, \ldots, P_n$ are also independent identically distributed random variables that have a $U(0, 1)$ distribution under $H_0$, and under $H_1$ have a distribution whose support is a subset of the interval $(0, 1)$ and is not a $U(0, 1)$ distribution. Therefore, if $f$ is the probability density function (pdf) of $P$, then (4) is equivalent to

\[ H_o : P \sim U(0,1) \]

\[ \text{vs} \]

\[ H_1 : P \not\sim U(0,1) \]

where $P$ has a pdf $f$ with support a subset of the interval $(0, 1)$.

This study considers the case: $\eta_i = \gamma \theta_i$, $i = 1, \ldots, n$, where $\theta_1, \ldots, \theta_n$ are independent identically distributed with DF $F$ with support $[a, \infty)$, $a \geq 0$ and the following hypothesis is tested:

\[ H_o : \gamma = 0 \]

\[ \text{vs} \]

\[ H_1 : \gamma > 0 \]

where the $i^{th}$ problem is based on $T_1, \ldots, T_n$, which are independent random variables from a conditional shifted exponential with pdf $f(x|\theta) = e^{-(x-a)\theta}$, $x \geq \gamma \theta$ and $\theta_1, \ldots, \theta_n$ are independent identically distributed with DF $F$ with support $[a, \infty)$, $a \geq 0$. Six methods are compared, namely: maximum of p-values method, Tippett’s method, Fisher method, logistic method, inverse normal method, and the sum of p-values method. These methods reject $H_0$ for large values of

\[ -\max_{1 \leq i \leq n} (P_i) \] (Maximum of p-values),

\[ -\min_{1 \leq i \leq n} (P_i) \] (Tippett’s),

\[ -2 \sum_{i=1}^{n} \ln P_i \] (Fisher),

\[ -\sum_{i=1}^{n} \ln \left( \frac{P_i}{1-P_i} \right) \] (Logistic),

\[ -\sum_{i=1}^{n} \Phi^{-1}(P_i) \] (Inverse normal)

and

\[ -\sum_{i=1}^{n} \frac{P_i}{\sqrt{n}} \] (Sum of p-values).

Derivation of EBS

Let $X_1, \ldots, X_n$ be an independent identically distributed pdf with $f(x, \theta)$, the hypotheses test hypotheses are $H_0: \theta = \theta_o$ vs. $H_1: \theta \in \Omega - \{\theta_o\}$, $\{T_1^n\}$ and $\{T_2^n\}$ are two sequences of test statistics for testing $H_0$, and the p-value of $T^n_i$ is $P^n_i = 1 - F_i(T^n_i)$ where $F_i(t) = P_{H_o}(T^n_i \leq t_i)$ for $i = 1, 2$.

Under these assumptions, there usually exists a positive valued function $C_i(\theta)$, which is termed the EBS of the sequence $\{T^n_i\}$ at $\theta$. This EBS sequence has the property that $C_i(\theta) = \lim_{n \to \infty} -\frac{2}{n} \ln P^n_i$ w. p. 1 under $\theta$, and the Bahadur efficiency of $\{T^n_1\}$ relative to $\{T^n_2\}$ which is given by $\frac{C_i(\theta)}{C_2(\theta)}$. Therefore, comparing two tests via the Bahadur efficiency is equivalent to comparing their corresponding EBS’s. The following three theorems provide the EBS for tests based on sums of independent identically distributed random variables.
Theorem 1
Let \( X_1, \ldots, X_n \) be independent identically distributed random variables with DF \( F \) and \( S_n = \sum_{i=1}^{n} X_i \). Assume that the moment generating function \( M(z) = E_F e^{z\gamma} \) exists and is near zero. If \( m(t) = \inf_{z \geq 0} e^{-zt} M(z) \), then
\[
\lim_{n \to \infty} -\frac{2}{n} \ln P_F [S_n \geq nt] = -2 \ln m(t).
\]

Theorem 2
Let \( \{T_n\} \) be a sequence of test statistics satisfying the following conditions:
1. Under \( H_1: T_n \to b(\theta) \) w.p.1 under \( \theta \)
   where \( b(\theta) \) is a real function.
2. An open interval \( I \) containing \( \{b(\theta): \theta \in \Omega\} \) exists and a function \( g \) continuous on \( I \) such that
   \[
   \lim_{n \to \infty} -\frac{2}{n} \ln [1 - F_n(t\sqrt{n})] = g(t)
   \]
   where \( F_n \) is the DF of \( T_n \) under \( H_0 \).

Thus the EBS of \( \{T_n\} \) is \( C(\theta) = g(b(\theta)) \).

Theorem 3
Let \( U_1, U_2, \ldots \) be independent identically distributed random variables like \( U \) with pdf \( f \), if the test hypotheses are:
\( H_0: \) \( U_i \sim U(0, 1) \) vs \( H_1: \) \( U_i \sim f \) on \( (0, 1) \), which is not \( U(0, 1) \), then
1. \( C_{\max}(f) = -2 \ln(\text{ess.Sup}_f(u)) \), where \( \text{ess.Sup}_f(u) = \sup\{u: f(u) > 0\} \) w.p.1 under \( f \).
2. If \( t(\ln t)^2 f(t) \to 0 \) as \( t \to 0 \), then \( C_{\min}(f)=0 \).

Note that for testing problem (6), the \( i^{th} \) p-value is:
\[
P_i = P(X \geq x_i) = e^{-x_i} \quad (7)
\]

The next four lemmas give the EBS for Fisher (\( C_F \)), logistic (\( C_L \)), inverse normal (\( C_N \)), sum of p-values (\( C_S \)), Tippett’s (\( C_T \)) and maximum of p-values (\( C_{\max} \)) methods.

Lemma 1
\[
C_F(\gamma) = 2 \gamma E_F \theta - 2 \ln(1 + \gamma E_F \theta) \quad (8)
\]
\[
C_S(\gamma) = -2 \ln m_S \left( \frac{1}{2} E_F \left( e^{-\gamma \theta} \right) \right),
\]
where
\[
m_S(t) = \inf_{z \geq 0} \left\{ e^{-zt} \frac{1 - e^{-z}}{z} \right\} \quad (9)
\]
\[
C_L(\gamma) = -2 \ln m_L (b_L(\gamma)),
\]
where
\[
m_L(t) = \inf_{0 < z < 1} \left\{ e^{-b_L(\gamma)z} \pi C_S(z) \right\}
\]
and
\[
b_L(\gamma) = \gamma E_F \theta - E_F (e^{\gamma \theta} - 1) \ln(1 - e^{-\gamma \theta}) \quad (10)
\]
\[
C_N(\gamma) = \left( E_F \left[ e^{\gamma \theta} \phi(F^{-1}(e^{\gamma \theta})) \right] \right)^2 \quad (11)
\]
The proof for Lemma 1 follows from Theorems (1) and (2).

Lemma 2
Let \( U_1, U_2, \ldots \) be independent identically distributed like \( U \) with pdf \( f \), if the test hypotheses are:
\( H_0: \) \( U_i \sim U(0, 1) \) vs \( H_1: \) \( U_i \sim f \) on \( (0, 1) \) but not \( U(0, 1) \), then
\[
C_{\max}(f) = -2 \ln(\text{ess.sup}_f(U))
\]
Where \( \text{ess.Sup}_f U = \sup\{u: f(u) > 0\} \) w.p.1 under \( f \).

Proof: Lemma 2
Let \( G_0(t) \) be the DF of -max, \( U_i \) under \( H_0 \). Then for
\[
-1 < t < 0, 1 - G_0(t) = (-t)^n,
\]
which implies that
\[-\frac{2}{n} \ln[1 - G_a(\text{max}_i U_i)] = -2 \ln \text{max}_i U_i.\]

Therefore,
\[
C_{\text{max}}(f) = \lim_{p \to 0} \left[ -2 \ln \text{max}_i U_i \right] = -2 \ln \lim_{p \to 0} \text{max}_i U_i = -2 \ln(\text{ess sup}_{f} U).
\]

Lemma 3
\[
C_{\text{max}}(\gamma) = 2\gamma a \quad (12)
\]
Proof: Lemma 3
Assume \( g(\theta) \) is the pdf of the DF \( F \), then the joint pdf of \( x \) and \( \theta \) is
\[
h(x, \theta) = g(\theta) f(x | \theta),
\]
where \( f(x | \theta) = e^{-x a \gamma}, x > \gamma \theta \). Then the marginal pdf of \( x \) is
\[
f(x) = \int_{a}^{x/a} h(x, \theta) d\theta
\]
\[
= \int_{a}^{x/a} e^{-(x \gamma a)} g(\theta) d\theta, \quad x > a \gamma, a \geq 0
\]
\[
= e^{-x} \int_{a}^{x/a} e^{a \gamma} dF(\theta), \quad x > \gamma \theta
\]
The under \( \gamma \) the p-value is \( e^{-x} P \) satisfies \( 0 < P < e^{-(\gamma a)} \), then \( \text{ess.sup} \ P = e^{-(\gamma a)} \), which implies \( C_{\text{max}}(\gamma) = 2\gamma a \) by theorem (3).

Lemma 4
\[
C_{T}(\gamma) = 0 \quad (13)
\]
Proof: Lemma 4
\[
g(p) = \int_{\gamma}^{e^{p/\gamma}} e^{p \gamma} g(\theta) d\theta
\]
\[
= \lim_{p \to 0} \left( \frac{(lnp)^2}{p} \right) e^{p \gamma} g(\theta) d\theta
\]
\[
= \lim_{p \to 0} \left( \frac{2ln p}{p} \right) e^{p \gamma} g(\theta) d\theta + \left( \frac{(ln p)^2}{p} \right) e^{p \gamma} g(\theta) d\theta
\]
\[
= \lim_{p \to 0} 2ln p \int_{\gamma}^{e^{p/\gamma}} e^{p \gamma} g(\theta) d\theta = \lim_{p \to 0} 2ln p \int_{\gamma}^{e^{p/\gamma}} e^{p \gamma} g(\theta) d\theta = 0
\]
using L’hopital rule because \( g(\infty) = 0 \) and \( \lim_{p \to 0} (ln p)^2 = 0 \).

Results
First, it the limits of the ratio of any two methods under study were found as \( \gamma \to 0 \) and \( \gamma \to \infty \). This gives the following results.

Corollary 1
\[
\lim_{\gamma \to 0} \frac{C_{\text{max}}(\gamma)}{C_{a}(\gamma)} = 0,
\]
where
\[
C_{a}(\gamma) = \{C_{S}(\gamma), C_{N}(\gamma), C_{L}(\gamma), C_{F}(\gamma)\},
\]
and
\[
\lim_{\gamma \to 0} \frac{C_{F}(\gamma)}{C_{S}(\gamma)} = \frac{1}{3},
\]
\[
\lim_{\gamma \to 0} \frac{C_{S}(\gamma)}{C_{L}(\gamma)} = \lim_{\gamma \to 0} \frac{C_{S}(\gamma)}{C_{N}(\gamma)} = \lim_{\gamma \to 0} \frac{C_{F}(\gamma)}{C_{L}(\gamma)} = \lim_{\gamma \to 0} \frac{C_{L}(\gamma)}{C_{N}(\gamma)} = 0.
\]
Corollary 2
\[
\lim_{\gamma \to \infty} \frac{C_L(\gamma)}{C_F(\gamma)} = 1, \quad \lim_{\gamma \to \infty} \frac{C_F(\gamma)}{C_N(\gamma)} = \lim_{\gamma \to \infty} \frac{C_L(\gamma)}{C_N(\gamma)} = 0.
\]

Proof 1
By (8) and (10)
\[
\frac{C_L(\gamma)}{C_F(\gamma)} \leq \frac{-2\ln 2 - 2\ln b_L(\gamma) + 2b_L(\gamma)}{2\gamma E_F \theta - 2\ln(1 + 2\gamma E_F \theta)},
\]
and
\[
\lim_{\gamma \to \infty} \frac{C_L(\gamma)}{C_F(\gamma)} \leq \lim_{\gamma \to \infty} \frac{-2\ln 2 - 2\ln b_L(\gamma) + 2b_L(\gamma)}{2\gamma E_F \theta - 2\ln(1 + 2\gamma E_F \theta)} = \lim_{\gamma \to \infty} \frac{-2b_L'(\gamma) + 2b_L'(\gamma)}{2E_F \theta - 2\frac{E_F \theta}{1 + 2\gamma E_F \theta}} = 1
\]
by using the L’Hôpital rule.

Similarly, it can be shown that
\[
\lim_{\gamma \to \infty} \frac{C_L(\gamma)}{C_F(\gamma)} \geq 1.
\]

Regarding \(C_S(\gamma)\), nothing can be concluded about general prior \(F\) because \(\lim_{\gamma \to \infty} \frac{b_S'(\gamma)}{b_S(\gamma)}\) has an indeterminate for \((0/0)\), thus, only a certain prior - namely, \(G(\alpha, \beta)\) with \(\alpha = 1\) and \(\beta = 2\), is considered:
\[
\lim_{\gamma \to \infty} \frac{C_S(\gamma)}{C_F(\gamma)} = \lim_{\gamma \to \infty} \frac{b_S'(\gamma)}{b_S(\gamma)} = \lim_{\gamma \to \infty} \frac{b_S'(\gamma)}{E_F \theta}.
\]

Proof 2
By (8) and (10)
\[
\lim_{\gamma \to \infty} \frac{C_S(\gamma)}{C_F(\gamma)} \leq \lim_{\gamma \to \infty} \frac{-2\ln 2 - 2\ln b_S(\gamma)}{2\gamma E_F \theta - 2\ln(1 + \lambda E_F \theta)} = \lim_{\gamma \to \infty} \frac{-2b_S'(\gamma)}{b_S(\gamma)} = \lim_{\gamma \to \infty} \frac{-b_S'(\gamma)}{E_F \theta}.
\]

Similarly,
\[
\lim_{\gamma \to \infty} \frac{C_S(\gamma)}{C_F(\gamma)} \geq \lim_{\gamma \to \infty} \frac{b_S'(\gamma)}{E_F \theta},
\]
hence,
\[
\lim_{\gamma \to \infty} \frac{C_L(\gamma)}{C_F(\gamma)} = 1. \tag{14}
\]

From the above relations it may be concluded that locally as \(\gamma \to 0\)
\[
C_N(\gamma) > C_L(\gamma) > C_S(\gamma) > C_F(\gamma) > C_{\text{max}}(\gamma) > C_T(\gamma),
\]
but as \(\gamma \to \infty\)
\[
C_N(\gamma) > C_L(\gamma) > C_{\text{max}}(\gamma) > C_F(\gamma) > C_T(\gamma),
\]
The dominance of one method over the other in case \(\gamma > 0\) can be shown mathematically only in some cases. The proof is omitted because, although it is straightforward, it is lengthy; however, numerical comparison for all methods is shown in Table 1. It appears from Table 1 that
\[
\gamma = 0.05: C_N(\gamma) > C_L(\gamma) > C_S(\gamma) > C_F(\gamma);
\]
\[
\gamma \in [0.1, 0.5]: C_S(\gamma) > C_N(\gamma) > C_L(\gamma) > C_F(\gamma);
\]
\[
\gamma = 1.00: C_N(\gamma) > C_L(\gamma) > C_F(\gamma) > C_S(\gamma);
\]
\[
\gamma \in [2, 3]: C_L(\gamma) > C_F(\gamma) > C_N(\gamma) > C_S(\gamma);
\]
\[
\gamma \in [5, 8]: C_F(\gamma) > C_L(\gamma) > C_N(\gamma) > C_S(\gamma);
\]
\[
\gamma \in [10, 20]: C_N(\gamma) > C_F(\gamma) > C_L(\gamma) > C_S(\gamma).
\]
Table 1: The Exact Bahadur Slopes for Conditional Shifted Exponential with Prior G(1, 2)

<table>
<thead>
<tr>
<th>γ</th>
<th>C_S(γ)</th>
<th>C_L(γ)</th>
<th>C_F(γ)</th>
<th>C_N(γ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.050</td>
<td>0.0249</td>
<td>0.0298</td>
<td>0.0094</td>
<td>0.0320</td>
</tr>
<tr>
<td>0.100</td>
<td>0.0903</td>
<td>0.0818</td>
<td>0.0354</td>
<td>0.8447</td>
</tr>
<tr>
<td>0.200</td>
<td>0.2512</td>
<td>0.2168</td>
<td>0.1271</td>
<td>0.2323</td>
</tr>
<tr>
<td>0.300</td>
<td>0.4414</td>
<td>0.3796</td>
<td>0.2599</td>
<td>0.4096</td>
</tr>
<tr>
<td>0.400</td>
<td>0.6329</td>
<td>0.5633</td>
<td>0.4244</td>
<td>0.6059</td>
</tr>
<tr>
<td>0.500</td>
<td>0.8173</td>
<td>0.7644</td>
<td>0.6137</td>
<td>0.8153</td>
</tr>
<tr>
<td>1.000</td>
<td>1.5887</td>
<td>1.9598</td>
<td>1.8028</td>
<td>1.9829</td>
</tr>
<tr>
<td>2.000</td>
<td>2.6053</td>
<td>4.9002</td>
<td>4.7811</td>
<td>4.5961</td>
</tr>
<tr>
<td>3.000</td>
<td>3.2781</td>
<td>8.1865</td>
<td>8.1082</td>
<td>7.3718</td>
</tr>
<tr>
<td>5.000</td>
<td>4.1821</td>
<td>15.1632</td>
<td>15.2042</td>
<td>12.6267</td>
</tr>
<tr>
<td>8.000</td>
<td>5.0527</td>
<td>26.0621</td>
<td>26.3336</td>
<td>21.1932</td>
</tr>
<tr>
<td>10.00</td>
<td>5.4753</td>
<td>33.4909</td>
<td>33.9110</td>
<td>40.3568</td>
</tr>
<tr>
<td>20.00</td>
<td>6.8134</td>
<td>71.6401</td>
<td>72.5729</td>
<td>162.3284</td>
</tr>
</tbody>
</table>

References


A New Biased Estimator Derived from Principal Component Regression Estimator

Set Foong Ng  
Universiti Teknologi MARA,  
Malaysia

Heng Chin Low Soon Hoe Quah  
Universiti Sains Malaysia

A new biased estimator obtained by combining the Principal Component Regression Estimator and the special case of Liu-type estimator is proposed. The properties of the new estimator are derived and comparisons between the new estimator and other estimators in terms of mean squared error are presented.

Key words: Linear regression model, Principal Component Regression Estimator, special case of Liu-type estimator, mean squared error.

Introduction
Multicollinearity is one of the problems faced in linear regression models. When multicollinearity is detected in data and the regressors that caused it are identified, one solution is to eliminate the regressors that are causing the multicollinearity. However, deleting regressors is not a safe strategy: there is no warning for extrapolation and there is no data to support a prediction in the region away from the multicollinearity.

Principal component regression is an alternative to regression deletion. Principal component regression is one type of biased regression method and its purpose is to eliminate those dimensions (which usually correspond to very small eigenvalues) causing the multicollinearity problem. Principal component regression approaches the problem of multicollinearity by dropping the dimension defined by a linear combination of the independent variables but not by a single independent variable (Rawlings, et al., 1998). Thus, the Principal Component Regression Estimator (Massy, 1965; Marquardt, 1970; Hawkins, 1973; Greenberg, 1975) is a biased alternative to the unbiased Ordinary Least Squares Estimator in the presence of multicollinearity.

Motivated by the idea of combining different estimators that might produce a better estimator, the r-k Class Estimator was proposed by Baye & Parker (1984). It has been shown that theoretical gains exist from combining the principal component regression and the ridge regression techniques. In addition, Kaciranlar and Sakallioglu (2001) proposed the r-d Class Estimator by combining the Liu Estimator and the Principal Component Regression Estimator.

A linear regression model with a dependent variable and p independent variables is given by

\[ Y = Z\gamma + \varepsilon, \]  

(1.1)

where \( Y \) is an \( n \times 1 \) standardized observed random vector, \( Z \) is an \( n \times p \) standardized known matrix with \( p \) independent variables, \( \gamma \) is an \( p \times 1 \) vector of parameter and \( \varepsilon \) is an \( n \times 1 \) vector of errors such that \( \varepsilon \sim N(0, \sigma^2 I_n) \).

If the matrix \( \lambda \) is a \( p \times p \) diagonal matrix whose diagonal elements are eigenvalues of \( ZZ' \), where the eigenvalues of \( ZZ' \) are denoted by \( \lambda_1, \lambda_2, \ldots, \lambda_p \), and if the matrix \( T \) is
A BIASED ESTIMATOR DERIVED FROM PRINCIPAL COMPONENT REGRESSION

A $p \times p$ orthonormal matrix consisting of the $p$ eigenvectors of $ZZ'$, then the Principal Component Regression Estimator of parameter $\gamma$ is given by

$$\hat{\gamma}_r = T_r(T_r'ZT_r)^{-1}T_r'ZY,$$  \hspace{1cm} (1.2)

where $r < p$, $T_r$ are the remaining eigenvectors of $ZZ'$ after deleting $p-r$ of the columns of $T$. This satisfies $T_r'ZT_r = \lambda_r = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$.

From the Liu-type estimator proposed by Liu (2003), a special case of Liu-type estimator was derived by Ng, et al. (2007)

$$\hat{\gamma}_c = (Z'Z + cI)^{-1}(Z'Y + \hat{\gamma}),$$  \hspace{1cm} (1.3)

where $c > 1$, $\hat{\gamma} = (Z'Z)^{-1}ZY$ is the Ordinary Least Squares Estimator.

Methodology

In this article, a new estimator was derived by combining the advantage of the principal component regression, $\hat{\gamma}_r$, and the special case of the Liu-type Estimator, $\hat{\gamma}_c$. Here, the estimator, $\hat{\gamma}_c$, in $\hat{\gamma}_c = (Z'Z + cI)^{-1}(Z'Y + \hat{\gamma})$ was substituted by the Principal Component Regression Estimator, $\hat{\gamma}_r$. Hence, a new expression of $\hat{\gamma}_c$ was obtained, that is

$$\hat{\gamma}_c (\text{new}) = (Z'Z + cI)^{-1}(Z'Y + \hat{\gamma}_r).$$  \hspace{1cm} (1.4)

Note that the eigenvalues and the eigenvectors are ordered so that $\lambda_1 > \lambda_2 > ... > \lambda_p$. The purpose of principal component regression is to eliminate those dimensions that usually correspond to eigenvalues that are very small. Thus, the concept of principal component regression eliminates $p-r$ of the columns of $T$ which correspond to the smallest $p-r$ eigenvalues. Hence, $T_r = [t_1, t_2, ..., t_r]$ is the matrix of the remaining eigenvectors of $ZZ'$ while $\lambda_r = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$ is the matrix of the remaining eigenvalues of $Z'Z$ after deleting $p-r$ of the columns of $T$; once again, this satisfies $T_r'ZT_r = \lambda_r = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$.

By including the matrix $T_r$ in the new expression of $\hat{\gamma}_r$, $\hat{\gamma}_c (\text{new}) = (Z'Z + cI)^{-1}(Z'Y + \hat{\gamma}_r)$, a new biased estimator, $\hat{\gamma}_r(c)$, was obtained which is given by:

$$\hat{\gamma}_r(c) = T_r(T_r'ZT_r + cI)^{-1}(T_r'ZY + T_r\hat{\gamma}_r),$$  \hspace{1cm} (1.5)

where $c > 1$, $r < p$, $\hat{\gamma}_r = T_r(T_r'ZT_r)^{-1}T_r'ZY$, $T_r$ are the remaining eigenvectors of $ZZ'$ after deleting $p-r$ of the columns of $T$. This satisfies $T_r'ZT_r = \lambda_r = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_r)$.

Properties of the New Estimator

The properties of the proposed new estimator are as follows:

(1) Bias of $\hat{\gamma}_r(c)$:

$$\text{bias}(\hat{\gamma}_r(c)) = -T_{p-r}T_{p-r}^\prime \gamma - T_r(\lambda_r + cI_r)^{-1}(c-1)T_r^\prime \gamma$$  \hspace{1cm} (2.1)

(2) Variance-covariance matrix of $\hat{\gamma}_r(c)$:

$$\text{Var}(\hat{\gamma}_r(c)) =$$

$$[T_r(\lambda_r + cI_r)^{-1}(\lambda_r + cI_r)T_r^\prime]^{-1}T_r^\prime \lambda_r^{-1}T_r^\prime \sigma^2$$  \hspace{1cm} (2.2)

(3) Mean squared error of $\hat{\gamma}_r(c)$:

$$\text{mse}(\hat{\gamma}_r(c)) =$$

$$\sum_{j=1}^{p-r} \left[ \frac{(\lambda_j + 1)^2 \sigma^2}{\lambda_j (\lambda_j + c)^2} \right] + \sum_{j=r+1}^{p} \frac{\sigma^2}{\lambda_j (\lambda_j + c)^2}$$  \hspace{1cm} (2.3)

(4) When $r = p$, the new estimator, $\hat{\gamma}_r(c)$, is equal to the special case of Liu-type estimator, $\hat{\gamma}_c$, that is,
\( \hat{\gamma}_p(c) = \hat{\gamma}_c \). \hfill (2.4)

(5) When \( r = p \) and \( c = 1 \), the new estimator, \( \hat{\gamma}_r(c) \), is equal to the Ordinary Least Squares Estimator, \( \hat{\gamma} \), that is,

\( \hat{\gamma}_r(1) = \hat{\gamma} \). \hfill (2.5)

(6) When \( c = 1 \), the new estimator, \( \hat{\gamma}_r(c) \), is equal to the Principal Component Regression Estimator, \( \hat{\gamma}_r \), that is

\( \hat{\gamma}_r(1) = \hat{\gamma}_r \). \hfill (2.6)

Results

The new estimator, \( \hat{\gamma}_r(c) \), was compared with the special case of Liu-type estimator, \( \hat{\gamma}_c \), the Ordinary Least Squares Estimator, \( \hat{\gamma} \), and the Principal Component Regression Estimator, \( \hat{\gamma}_r \), in terms of mean squared error in order to evaluate the performance of the new estimator.

The mean squared errors \( \hat{\gamma}_c \), \( \hat{\gamma} \) and \( \hat{\gamma}_r \) are shown in Equations (3.1) to (3.3), respectively:

\[
\text{mse}(\hat{\gamma}_c) = \sum_{j=1}^{p} \frac{\sigma^2}{\lambda_j} + \frac{(c-1)^2 \gamma_j^2}{\lambda_j(c) + c^2} \hfill (3.1)
\]

\[
\text{mse}(\hat{\gamma}) = \sum_{j=1}^{p} \frac{\sigma^2}{\lambda_j} \hfill (3.2)
\]

\[
\text{mse}(\hat{\gamma}_r) = \sum_{j=1}^{r} \frac{\sigma^2}{\lambda_j} + \sum_{j=r+1}^{p} \gamma_j^2 \hfill (3.3)
\]

From the properties of the new estimator, \( \hat{\gamma}_r(c) \) is equivalent to \( \hat{\gamma}_c \) when \( r = p \).

Theorem 3.1: Comparison Between the \( \hat{\gamma}_r(c) \) and \( \hat{\gamma}_{r+1}(c) \)

If \( \sigma^2 > \gamma_{r+1}^2 \lambda_{r+1} \),

(a) \( \text{mse}(\hat{\gamma}_{r+1}(c)) > \text{mse}(\hat{\gamma}_r(c)) \) for \( 1 < c < a_1 \),

(b) \( \text{mse}(\hat{\gamma}_{r+1}(c)) < \text{mse}(\hat{\gamma}_r(c)) \) for \( c > a_1 \),

where

\[ a_1 = \frac{(\lambda_{r+1} + 1)^2 \sigma^2 + \lambda_{r+1} \gamma_{r+1}^2 (1- \lambda_{r+1}^2)}{2 \lambda_{r+1} \gamma_{r+1}^2 (1+ \lambda_{r+1})} \]

Theorem 3.1 Proof

Consider the difference between the mean squared errors of \( \hat{\gamma}_r(c) \) and \( \hat{\gamma}_{r+1}(c) \):

\[
D_1 = \text{mse}(\hat{\gamma}_{r+1}(c)) - \text{mse}(\hat{\gamma}_r(c))
\]

\[
= \sum_{j=1}^{r} \frac{(\lambda_j + 1)^2 \sigma^2}{\lambda_j(\lambda_j + c)^2} + \sum_{j=1}^{p} \frac{(c-1)^2 \gamma_j^2}{\lambda_j(c) + c^2} - \sum_{j=1}^{r} \frac{(\lambda_j + 1)^2 \sigma^2}{\lambda_j(c) + c^2} - \sum_{j=r+1}^{p} \gamma_j^2 - \sum_{j=1}^{r} \frac{(c-1)^2 \gamma_j^2}{\lambda_j(c) + c^2}
\]

\[
= \frac{(\lambda_{r+1} + 1)^2 \sigma^2}{\lambda_{r+1}(\lambda_{r+1} + c)^2} + \frac{(c-1)^2 \gamma_{r+1}^2}{\lambda_{r+1}(\lambda_{r+1} + c)^2} - \gamma_{r+1}^2
\]

\[
= \frac{(\lambda_{r+1} + 1)^2 \sigma^2 + \lambda_{r+1}(c-1)^2 \gamma_{r+1}^2 - \lambda_{r+1} \lambda_{r+1} + c)^2 \gamma_{r+1}^2}{\lambda_{r+1}(\lambda_{r+1} + c)^2}
\]

(3.4)

Thus, \( D_1 > 0 \) if and only if

\[
(\lambda_{r+1} + 1)^2 \sigma^2 + \lambda_{r+1}(c-1)^2 \gamma_{r+1}^2 - \lambda_{r+1} \lambda_{r+1} + c)^2 \gamma_{r+1}^2 > 0.
\]

(3.5)

The solution for the inequality (3.5) is

\[ 1 < c < a_1 \], \hfill (3.6)

where

\[ a_1 = \frac{(\lambda_{r+1} + 1)^2 \sigma^2 + \lambda_{r+1} \gamma_{r+1}^2 (1- \lambda_{r+1}^2)}{2 \lambda_{r+1} \gamma_{r+1}^2 (1+ \lambda_{r+1})} \]

Because \( c > 1 \), it requires \( a_1 > 1 \), that is
A BIASED ESTIMATOR DERIVED FROM PRINCIPAL COMPONENT REGRESSION

Thus, if $\sigma^2 > \gamma^2_p \lambda_j$, $\text{mse}(\hat{\gamma}_r(c)) > \text{mse}(\hat{\gamma}_r(c))$ for $1 < c < a_j$. Similarly, $\text{mse}(\hat{\gamma}_r(c)) < \text{mse}(\hat{\gamma}_r(c))$ for $c > a_j$. Hence, the proof for Theorem 3.1 is completed.

Theorem 3.2: Comparison between the New (Biased) Estimator and the Special Case of Liu-Type Estimator, $\hat{\gamma}_e$

If $\sigma^2 > \gamma^2_p \lambda_j$ for $j \in \{1, 2, \ldots, r\}$,
(a) $\text{mse}(\hat{\gamma}_e) > \text{mse}(\hat{\gamma}_r(c))$ for $1 < c < \min\{(a_j)\}$,
(b) $\text{mse}(\hat{\gamma}_e) < \text{mse}(\hat{\gamma}_r(c))$ for $c > \max\{(a_j)\}$,

where

$$(a_j) = \frac{(\lambda_j + 1) \sigma^2 + \lambda_j \gamma^2_p (1 - \lambda_j^2)}{2 \lambda_j \gamma^2_p (1 + \lambda_j)}$$

for $j \in \{1, 2, \ldots, p\}$. The proof for Theorem 3.2 is completed.

Theorem 3.3: Comparison between the New (Biased) Estimator and the Unbiased Ordinary Least Squares Estimator

(a) If $\sigma^2 > \gamma^2_p \lambda_j$ for all $j \in \{1, 2, \ldots, r\}$ and

$$\sum_{j=r+1}^p \frac{\gamma^2_j \lambda_j - \sigma^2}{\lambda_j} \leq 0,$$

$\text{mse}(\hat{\gamma}_r(c)) < \text{mse}(\hat{\gamma})$ for $c > 1$.

(b) If $\sigma^2 < \gamma^2_p \lambda_j$ for some $j \in \{1, 2, \ldots, r\}$ and

$$\sum_{j=r+1}^p \frac{\gamma^2_j \lambda_j - \sigma^2}{\lambda_j} \leq 0,$$

$\text{mse}(\hat{\gamma}_r(c)) < \text{mse}(\hat{\gamma})$ for $1 < c < \min\{(a_j)\}$, $\text{mse}(\hat{\gamma}_r(c)) > \text{mse}(\hat{\gamma})$ for $c > \max\{(a_j)\}$, where

$$(a_j) = \frac{\hat{\lambda}_j \gamma^2_p + 2 \lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma^2_p - \sigma^2}$$

for $j \in \{1, 2, \ldots, r\}$.

Theorem 3.3 Proof

Consider the difference between the mean squared errors of $\hat{\gamma}_r(c)$ and $\hat{\gamma}$:

$$D = \text{mse}(\hat{\gamma}_r(c)) - \text{mse}(\hat{\gamma})$$

$$= \sum_{j=1}^r \left[ \frac{(\lambda_j + 1) \sigma^2}{\lambda_j (\lambda_j + c)^2} \right] + \sum_{j=r+1}^p \gamma^2_j$$

$$+ \sum_{j=1}^r \left[ \frac{(c-1) \gamma^2_j}{(\lambda_j + c)^2} - \frac{\gamma^2_p \lambda_j - \sigma^2}{\lambda_j} \right]$$

$$= \sum_{j=1}^r \left[ \frac{(\lambda_j + 1) \sigma^2 + \lambda_j (c-1) \gamma^2_j - \sigma^2 (\lambda_j + c)^2}{\lambda_j (\lambda_j + c)^2} \right]$$

$$+ \sum_{j=r+1}^p \frac{\gamma^2_p \lambda_j - \sigma^2}{\lambda_j}$$

(3.8)
Thus, $D_z < 0$ if and only if
\[ \sum_{j=r+1}^{p} \frac{\gamma_j^2 \lambda_j - \sigma^2}{\lambda_j} \leq 0 \quad (3.9) \]
and
\[ (\lambda_j + 1)^2 \sigma^2 + \lambda_j (c - 1)^2 \gamma_j - \sigma^2 (\lambda_j + c)^2 < 0. \quad (3.10) \]

The inequality (3.10) can also be written as
\[ (\gamma_j^2 \lambda_j - \sigma^2) c^2 - 2(\gamma_j^2 \lambda_j + \lambda_j \sigma^2) c + (\gamma_j^2 \lambda_j + 2 \lambda_j \sigma^2 + \sigma^2) < 0 \quad (3.11) \]

Solving the equation
\[ (\gamma_j^2 \lambda_j - \sigma^2) c^2 - 2(\gamma_j^2 \lambda_j + \lambda_j \sigma^2) c + (\gamma_j^2 \lambda_j + 2 \lambda_j \sigma^2 + \sigma^2) - \lambda_j \gamma_j^2 - \sigma^2 = 0 \quad (3.12) \]
the solutions $c = \frac{\lambda_j \gamma_j^2 + 2 \lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma_j^2 - \sigma^2}$ and $c = 1$ are obtained.

Let $(a_j) = \frac{\lambda_j \gamma_j^2 + 2 \lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma_j^2 - \sigma^2}$ where the values of $(a_j)$ could be less than 1 or greater than 1. The condition requiring for $(a_j) < 1$ is given by
\[ (a_j) < 1 \]
\[ \frac{\lambda_j \gamma_j^2 + 2 \lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma_j^2 - \sigma^2} < 1 \]
\[ \frac{2 \lambda_j \sigma^2 + 2 \sigma^2}{\lambda_j \gamma_j^2 - \sigma^2} < 0 \]
\[ \lambda_j \gamma_j^2 - \sigma^2 < 0 \]
\[ \sigma^2 > \gamma_j^2 \lambda_j. \quad (3.13) \]

Similarly, the condition requiring for $(a_j) > 1$ is given by
\[ \sigma^2 < \gamma_j^2 \lambda_j. \quad (3.14) \]

For the first situation where $(a_j) < 1$ for all $j \in \{1, 2, \ldots, r\}$, the solution for the inequality (3.11) is
\[ c > 1 \text{ if } \sigma^2 > \gamma_j^2 \lambda_j \text{ for all } j \in \{1, 2, \ldots, r\}. \quad (3.15) \]

Thus, $\text{mse}(\hat{\gamma}_c(c)) < \text{mse}(\hat{\gamma})$ for $c > 1$ if $\sigma^2 > \gamma_j^2 \lambda_j$ for all $j \in \{1, 2, \ldots, r\}$ and
\[ \sum_{j=r+1}^{p} \frac{\gamma_j^2 \lambda_j - \sigma^2}{\lambda_j} \leq 0. \]

The proof for Theorem 3.3(a) is completed.

For the second situation where $(a_j) > 1$ for some $j \in \{1, 2, \ldots, r\}$, the solution for the inequality (3.11) is:
\[ 1 < c < \min \{ (a_j) \} \text{ if } \sigma^2 < \gamma_j^2 \lambda_j \text{ for some } j \in \{1, 2, \ldots, r\}, \]
where
\[ (a_j) = \frac{\lambda_j \gamma_j^2 + 2 \lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma_j^2 - \sigma^2} \text{ for } j \in \{1, 2, \ldots, r\}. \]

Thus, $\text{mse}(\hat{\gamma}_c(c)) < \text{mse}(\hat{\gamma})$ for $1 < c < \min \{ (a_j) \}$ and $\text{mse}(\hat{\gamma}_c(c)) > \text{mse}(\hat{\gamma})$ for $c > \max \{ (a_j) \}$ if $\sigma^2 < \gamma_j^2 \lambda_j$ for some $j \in \{1, 2, \ldots, r\}$ and
\[ \sum_{j=r+1}^{p} \frac{\gamma_j^2 \lambda_j - \sigma^2}{\lambda_j} \leq 0. \]

The proof for Theorem 3.3(b) is completed.

Theorem 3.4: Comparison between the New Estimator and the Principal Component Regression Estimator in terms of Mean Squared Error

(a) If $\sigma^2 > \gamma_j^2 \lambda_j$ for all $j \in \{1, 2, \ldots, r\}$, $\text{mse}(\hat{\gamma}_c(c)) < \text{mse}(\hat{\gamma})$ for $c > 1$.

(b) If $\sigma^2 < \gamma_j^2 \lambda_j$ for some $j \in \{1, 2, \ldots, r\}$, $\text{mse}(\hat{\gamma}_c(c)) < \text{mse}(\hat{\gamma})$ for $1 < c < \min \{ (a_j) \}$, $\text{mse}(\hat{\gamma}_c(c)) > \text{mse}(\hat{\gamma})$ for $c > \max \{ (a_j) \}$, where
A BIASED ESTIMATOR DERIVED FROM PRINCIPAL COMPONENT REGRESSION

\[(a_j) = \frac{\hat{\lambda}_j \gamma_j' + 2\lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma_j' - \sigma^2} \quad \text{for} \quad j \in \{1, 2, \cdots, r\}.\]

Theorem 3.4 Proof

The first derivative of \(\operatorname{mse}(\hat{\gamma}_r(c))\) with respect to \(c\) is given by

\[
\begin{align*}
\frac{d}{dc}[\operatorname{mse}(\hat{\gamma}_r(c))] &= \frac{d}{dc} \left[ \sum_{j=r}^p \left( \frac{(\lambda_j + 1)^2 \sigma^2}{\hat{\lambda}_j (\hat{\lambda}_j + c)} \right) + \sum_{j=r}^p \frac{(c-1) \gamma_j' \gamma_j}{(\hat{\lambda}_j + c)^2} \right] \\
&= 2 \sum_{j=r}^p \left[ -\sigma^2 (\lambda_j + 1)^2 \frac{1}{\hat{\lambda}_j (\hat{\lambda}_j + c)} + \frac{(c-1) (\lambda_j + 1) \gamma_j'}{\hat{\lambda}_j (\hat{\lambda}_j + c)^2} \right].
\end{align*}
\]

When \(c = 1\),

\[
\begin{align*}
\frac{d}{dc}[\operatorname{mse}(\hat{\gamma}_r(c))] &= -2\sigma^2 \sum_{j=r}^p \left[ \frac{1}{\hat{\lambda}_j (\hat{\lambda}_j + 1)} \right] < 0. \\
&= 2 \sum_{j=r}^p \left[ -\sigma^2 (\lambda_j + 1)^2 \frac{1}{\hat{\lambda}_j (\hat{\lambda}_j + c)} + \frac{(c-1) (\lambda_j + 1) \gamma_j'}{\hat{\lambda}_j (\hat{\lambda}_j + c)^2} \right].
\end{align*}
\]

It was found that the function of \(\operatorname{mse}(\hat{\gamma}_r(c))\) has a negative slope of approximately \(c = 1\). Furthermore, the new estimator, \(\hat{\gamma}_r(c)\), is equivalent to the Principal Component Regression Estimator, \(\hat{\gamma}_r\), when \(c = 1\). Thus, \(c > 1\) exists such that \(\operatorname{mse}(\hat{\gamma}_r(c)) < \operatorname{mse}(\hat{\gamma}_r)\).

Consider the difference between the mean squared errors of \(\hat{\gamma}_r(c)\) and \(\hat{\gamma}_r:\)

\[
D_j = \operatorname{mse}(\hat{\gamma}_r(c)) - \operatorname{mse}(\hat{\gamma}_r) \]

\[
= \sum_{j=r}^p \left[ \frac{(\lambda_j + 1)^2 \sigma^2}{\hat{\lambda}_j (\hat{\lambda}_j + c)} \right] + \sum_{j=r+1}^p \gamma_j' \gamma_j \\
+ \sum_{j=r}^p \left[ \frac{(c-1) \gamma_j'}{(\hat{\lambda}_j + c)^2} \right] - \sum_{j=r}^p \frac{\sigma^2}{\hat{\lambda}_j} - \sum_{j=r+1}^p \gamma_j^2 \\
= \sum_{j=r}^p \left[ \frac{(\lambda_j + 1)^2 \sigma^2 + \lambda_j (c-1) \gamma_j' - \sigma^2 (\lambda_j + c)}{\lambda_j (\lambda_j + c)^2} \right].
\]

Thus, \(D_j < 0\) if and only if

\[
(\lambda_j + 1)^2 \sigma^2 + \lambda_j (c-1) \gamma_j' - \sigma^2 (\lambda_j + c)^2 < 0.
\]

The inequality (3.19) can also be written as

\[
(\gamma_j' \lambda_j - \sigma^2) c^2 - 2(\gamma_j' \lambda_j + \lambda_j \sigma^2) c + (\gamma_j' \lambda_j + 2\lambda_j \sigma^2 + \sigma^2) < 0
\]

(3.20)

The solution for the inequality (3.20) is the same as the solution for the inequality (3.11). Thus, \(\operatorname{mse}(\hat{\gamma}_r(c)) < \operatorname{mse}(\hat{\gamma}_r)\) for \(c > 1\) if \(\sigma^2 > \gamma_j' \lambda_j\) for all \(j \in \{1, 2, \cdots, r\}\).

By contrast, if \(\sigma^2 < \gamma_j' \lambda_j\) for some \(j \in \{1, 2, \cdots, r\}\), \(\operatorname{mse}(\hat{\gamma}_r(c)) < \operatorname{mse}(\hat{\gamma}_r)\) for \(1 < c < \min \{\{a_j\}_{j=1}\} \) and \(\operatorname{mse}(\hat{\gamma}_r(c)) > \operatorname{mse}(\hat{\gamma}_r)\) for \(c > \max \{\{a_j\}_{j=1}\}\) where

\[
(a_j) = \frac{\lambda_j \gamma_j' + 2\lambda_j \sigma^2 + \sigma^2}{\lambda_j \gamma_j' - \sigma^2} \quad \text{for} \quad j \in \{1, 2, \cdots, r\}.
\]

The proof for Theorem 3.4 is completed.

Numerical Example

A numerical example illustrates Theorems 3.2, 3.3 and 3.4. The data set is from Ryan (1997, pp. 402-403). The data consists of one dependent variable and six independent variables. The regression model with standardized variables is

\[
Y = Z\gamma + \varepsilon,
\]

where \(Y\) is a 50×1 standardized observed random vector, \(Z\) is a 50×6 standardized known matrix with six independent variables, \(\gamma\) is a 6×1 vector of parameters and \(\varepsilon\) is a 50×1 vector of errors.

Multicollinearity diagnostic indicates the presence of multicollinearity in the data. The least squares estimates are given by \(\hat{\gamma}_1 = -5.218, \hat{\gamma}_2 = -0.376, \hat{\gamma}_3 = 8.869, \hat{\gamma}_4 = -1.755, \hat{\gamma}_5 = -0.320\) and \(\hat{\gamma}_6 = -1.178\). The estimated variance of the error term is given by \(\hat{\sigma}^2 = 0.000655\) while the eigenvalues are given by \(\hat{\lambda}_1 = 5.80831,\)
\[ \lambda_1 = 0.11749, \quad \lambda_2 = 0.04812, \quad \lambda_3 = 0.02501, \quad \lambda_4 = 0.00081, \quad \lambda_5 = 0.00025. \]

For practical purpose, \( \gamma \) and \( \sigma^2 \) in Theorems 3.2, 3.3 and 3.4 are substituted by the estimated \( \hat{\gamma} \) and \( \hat{\sigma}^2 \) in this numerical example. In this numerical example, the Principal Component Regression Estimator has the smallest mean squared error when \( r = 4 \). Thus, \( p = 6 \) and \( r = 4 \) are used throughout this example.

It was found that the condition for Theorem 3.2 is satisfied, that is, \( \sigma^2 > \gamma^2_j \lambda_j \) for \( j \in \{5, 6\} \), and \( \{a_j\} = \{4.43639, 1.42839\} \) for \( j \in \{5, 6\} \) was obtained. Taking the value of \( c = 1.3 \) where \( 1 < c < \min \{\{a_j\}\} \), it was found that \( \text{mse}(\hat{\gamma}, (c)) = 5.62615 \), that is, less than \( \text{mse}(\hat{\gamma}) = 6.21735 \). This agrees with Theorem 3.2(a). On the other hand, taking \( c = 4.5 \) where \( c > \max \{\{a_j\}\} \), it was found that \( \text{mse}(\hat{\gamma}) = 52.7144 \) and \( \text{mse}(\hat{\gamma}, (c)) = 53.1363 \), thus, \( \text{mse}(\hat{\gamma}, (c)) < \text{mse}(\hat{\gamma}) \). This is also in line with Theorem 3.2(b).

For Theorem 3.3, the value of \( \sum_{j=5}^{6} \gamma^2_j \lambda_j - \sigma^2 = -1.8931 < 0 \). It was also found that the values for \( \gamma^2_j \lambda_j = \{158.13, 0.01662, 3.78485, 0.07705\} \) for \( j = \{1, 2, 3, 4\} \). This shows that the condition for Theorem 3.3(a) is not satisfied since \( \sigma^2 < \gamma^2_j \lambda_j \) for \( j = \{1, 2, 3, 4\} \). Thus, Theorem 3.3(b) will be used to illustrate Theorem 3.3.

The values of \( \{a_j\} = \{1.00006, 1.09168, 1.00036, 1.01758\} \) for \( j = \{1, 2, 3, 4\} \). Choosing \( c = 1.00003 \) where \( 1 < c < \min \{\{a_j\}\} \), \( \text{mse}(\hat{\gamma}, (c)) = 1.536355 \), was obtained that is, smaller than \( \text{mse}(\hat{\gamma}) = 3.42946 \). By contrast, taking \( c = 1.3 \) where \( c > \max \{\{a_j\}\} \), it was found that \( \text{mse}(\hat{\gamma}, (c)) = 5.62615 > \text{mse}(\hat{\gamma}) = 3.42946 \). This agrees with Theorem 3.3.

For Theorem 3.4, it was found that \( \sigma^2 < \gamma^2_j \lambda_j \) for \( j = \{1, 2, 3, 4\} \). Since the condition for Theorem 3.4(b) is satisfied, Theorem 3.4(b) was used to illustrate Theorem 3.4. Choosing \( c = 1.00003 \) where \( 1 < c < \min \{\{a_j\}\} \), \( \text{mse}(\hat{\gamma}, (c)) = 1.536356 \) and \( \text{mse}(\hat{\gamma}) = 1.536359 \) were obtained. This shows that new estimator has a smaller mean squared error for \( 1 < c < \min \{\{a_j\}\} \). By contrast, taking \( c = 1.3 \) where \( c > \max \{\{a_j\}\} \), it was found that \( \text{mse}(\hat{\gamma}, (c)) = 5.62615 > \text{mse}(\hat{\gamma}) = 1.536358 \). This is in line with the Theorem 3.4. This numerical comparison was shown to be in line with the theoretical comparison.

Conclusion
The new biased estimator was obtained by combining the Principal Component Regression Estimator and the special case of Liu-type estimator. When certain conditions are satisfied, the new estimator has been shown to have smaller mean squared error compared to the special case of Liu-type estimator, the Ordinary Least Squares Estimator and the Principal Component Regression Estimator. The numerical comparison was also shown to be in line with the theoretical comparison.

In conclusion, the proposed new estimator was shown be an improvement in terms of reduction in mean squared error. Thus, the new estimator could be considered as an alternative to the unbiased Ordinary Least Squares Estimator when multicollinearity is detected in a linear regression model.

Acknowledgement
This research is supported by the Fundamental Research Grant Scheme.

References


Estimations on the Generalized Exponential Distribution Using Grouped Data

Hassan Pazira Parviz Nasiri
Payame Noor University of Tehran
Tehran, Iran

Classical and Bayesian estimators are obtained for the shape parameter of the Generalized-Exponential distribution under grouped data. In Bayesian estimation, three types of loss functions are considered: the Squared Error loss function which is classified as a symmetric function, the LINEX and Precautionary loss functions which are asymmetric. These estimators are compared with the corresponding estimators derived from un-grouped data empirically using Monte-Carlo simulation.

Key words: Generalized-Exponential; Bayesian estimation; Grouped data; LINEX loss; Precautionary loss; Newton-Raphson method; Fisher's information number; Monte-Carlo simulation.

Introduction

In various fields of science such as biology, engineering, and medicine it is not possible to obtain the measurements of a statistical experiment exactly, but is possible to classify them into intervals, rectangles or disjoint subsets (Alodat & Al-Saleh, 2002; Heitjan, 1989; Surles & Padgett, 2001; Wu & Perloff, 2005; Pipper & Ritz, 2006). For example, in life testing experiments, the failure time of a component may be observed to the nearest hour, day or month. Data for which true values are known only up to subsets of the sample space are called grouped data. In general, grouped data can be formulated as follows: Let \( X_1, X_2, ..., X_n \) be a random sample from the density \( f(x; \theta), \ x \in \chi, \ \theta \in \Theta \) and \( X_1, X_2, ..., X_{k+1} \) be a partition of the sample space \( \chi \) and \( N_j \) = the number of \( X_i \)'s that fall in \( \chi_j \) for \( j=1,2, ..., k+1 \). The set of pairs \( \{(X_1, N_1), ..., (X_{k+1}, N_{k+1})\} \) is called grouped data, and these data are used to draw inferences about the parameter \( \theta \). Due to a lack of complete information about the sample, there is a loss in information due to the grouping. Schervish (1995, p. 114), shows the following

\[
I_{\theta}(\theta) = I_{\hat{\theta}}(\theta) + E_\theta[I_{\hat{\theta}}(\theta|Y)]
\]

where \( I_{\theta}(\theta) \) and \( I_{\hat{\theta}}(\theta) \) are Fisher's information numbers obtained from \( X \) and \( Y \), respectively, and \( E_\theta[I_{\hat{\theta}}(\theta|Y)] \) is the conditional score function. If \( Y \) is replaced by the grouped sample \( \underline{n} = (N_1, N_2, ..., N_{k+1}) \), then \( I_{\hat{\theta}}(\theta) \geq I_{\hat{\theta}}(\theta) \) for all \( \theta \), meaning that the information in the sample \( X \) about \( \theta \) is reduced to \( I_{\hat{\theta}}(\theta) \) due to grouping.

Kuldorff (1961) considered non-Bayesian estimation from grouped data when the data were from normal and exponential distributions. Alodat and Al-Saleh (2000) considered the Bayesian estimation from grouped data when the underlying distribution is exponential. Alodat, et al. (2007) obtained Bayesian prediction intervals from grouped data when the underlying distribution is exponential. Aludaat, et al. (2008) obtained the Bayesian and non-Bayesian estimation from grouped data when the underlying distribution is Burr type X.
Recently a new distribution, called the Generalized-Exponential (GE) distribution, has been introduced. This distribution can be used quite effectively in situations where a skewed distribution is needed. It has been studied extensively (Gupta & Kundu, 1999, 2001a, 2001b, 2002, 2003a, 2003b, 2004; Raqab, 2002; Raqab & Ahsanullah, 2001; Zheng, 2002; Kundu, Gupta & Manglick, 2004). Singh, et al. (2008) estimated the parameters of this distribution under some symmetric and asymmetric loss functions using Lindley’s approximation technique. Note that the generalized exponential distribution is a sub-model of the exponentiated Weibull distribution introduced by Mudholkar and Srivastava (1993) and later studied by Mudholkar, Srivastava and Freimer (1995) and Mudholkar and Hutson (1996). Also recently, Nasiri and Pazira (2010) conducted Bayesian and non-Bayesian estimations on the Generalized Exponential distribution in the presence of outliers.

This article considers the group data problem when the density \( f(x; \theta) \) is Generalized-Exponential, that is,

\[
f(x; \theta) = \theta e^{-x} (1 - e^{-x})^{\theta - 1}; \quad x > 0, \quad \theta > 0, \quad (1)
\]

and the corresponding distribution function is

\[
F(x; \theta) = (1 - e^{-x})^\theta; \quad x > 0, \quad \theta > 0. \quad (2)
\]

The density functions of the generalized exponential distribution can take different shapes. For \( \theta \leq 1 \), it is a decreasing function and for \( \theta > 1 \), it is a unimodal, skewed, right tailed similar to the Weibull, or gamma density, function. It is observed that even for very large shape parameter \( \theta \), it is not symmetric. For this density function (1), the mode is at \( \log \theta \) for both \( \theta > 1 \) and \( \theta \leq 1 \), the mode is at 0 and the median is at \( -\ln \left(1 - \frac{\sqrt{0.5}}{\theta}\right) \). The mean, median and mode are non-linear functions of the shape parameter \( \theta \) and, as this parameter goes to infinity, all of them tend to infinity. For large values of \( \theta \), the mean, median and mode are approximately equal to \( \theta \) but they converge at different rates. Figure 1 shows the shape of \( f(x; \theta) \) for different values of \( \theta \).

Likelihood Function and M.L.E from Grouped Data

The likelihood density based on the grouped data is derived as follows. Let \( X_1, X_2, \ldots, X_n \) be a random sample from \( \text{GE}(\theta) \). Assume that the sample space of \( f(x; \theta) \) is partitioned into \( k + 1 \) equally-spaced intervals as follows. Let \( I_j = [(j - 1)\delta, j\delta) \), \( j = 1, \ldots, k \) and \( I_{k+1} = [k\delta, \infty) \), \( \delta > 0 \). If \( N_j \) denotes the number of
Given $X_j$'s that fall in $I_j$, $j=1,2,...,k+1$, then \( n = N_1 + ... + N_{k+1} \). Let

\[
P_j = P_j(\theta) = P(X \in I_j) = P((j-1)\delta < X < j\delta) = (1 - e^{-j\delta})^{\theta} - (1 - e^{-(j-1)\delta})^{\theta},
\]

for $j = 1,\ldots,k$ and

\[
P_{k+1} = P_{k+1}(\theta) = P(X > k\delta) = 1 - (1 - e^{-k\delta})^{\theta}.
\]

If $A_j = \log(1 - e^{-(j-1)\delta})$, then

\[
P_j = e^{\theta A_j} - e^{\theta A_j},
\]

for $j = 1,\ldots,k$ and

\[
P_{k+1} = 1 - e^{\theta A_{k+1}}.
\]

Thus, the density of $n = (N_1, N_2, ..., N_{k+1})$ is given by the multinomial distribution as follows:

\[
f(n; \theta) = \frac{n!}{n_1! \cdots n_{k+1}!} P_1^{n_1} \cdots P_{k+1}^{n_{k+1}}
\]

\[
= C \left(1 - e^{\theta A_{k+1}}\right)^{n_{k+1}} \prod_{j=1}^{k} \left(e^{\theta A_{j\mid i}} - e^{\theta A_j}\right)^{n_j}.
\]

where $C$ is a normalizing constant.

Next, find the MLE of $\theta$ based on the density (3) by maximizing the log-likelihood function

\[
\log f(n; \theta) = \text{constant} + \sum_{j=1}^{k} n_j \log(e^{\theta A_{j\mid i}} - e^{\theta A_j}) + n_{k+1} \log(1 - e^{\theta A_{k+1}}).
\]

The first derivative of the log-likelihood is

\[
\frac{\partial \log f(n; \theta)}{\partial \theta} = \frac{k}{\sum_{j=1}^{k} n_j A_{j\mid i} e^{\theta A_{j\mid i}} - A_j e^{\theta A_j}} - n_{k+1} A_{k+1} e^{\theta A_{k+1}}
\]

(4)

The MLE for $\theta$ is the solution of $\partial \log f(n; \theta)/\partial \theta = 0$, so the MLE is $\hat{\theta}$ such that

\[
\sum_{j=1}^{k} n_j A_{j\mid i} e^{\hat{\theta} A_{j\mid i}} - A_j e^{\hat{\theta} A_j} = n_{k+1} A_{k+1} e^{\hat{\theta} A_{k+1}}.
\]

(5)

The notation $\hat{\theta}_{MG}$ is used to denote the MLE of $\theta$ obtained from the grouped data. The Newton-Raphson method can be used to solve (5), thus, the solution of the equation is

\[
\theta_{i+1} = \theta_i - \frac{h(\theta_i)}{h'(\theta_i)}; \quad i = 1,2,3,\ldots
\]

(6)

where

\[
h(\theta) = \sum_{j=1}^{k} n_j A_{j\mid i} e^{\theta A_{j\mid i}} - A_j e^{\theta A_j} - n_{k+1} A_{k+1} e^{\theta A_{k+1}},
\]

and

\[
h'(\theta) = -\sum_{j=1}^{k} n_j \frac{(A_{j\mid i} - A_j)^2 e^{(\theta A_{j\mid i} + A_{j\mid i})}}{\left( e^{\theta A_{j\mid i}} - e^{\theta A_j} \right)^2}
\]

\[-n_{k+1} \frac{A_{k+1}^2 e^{\theta A_{k+1}}}{\left(1 - e^{\theta A_{k+1}}\right)^2}.
\]

Here, the initial solution $\theta_0$ should be selected from the MLE of $\theta$ based on the un-grouped data. Gupta and Kundu (1999 & 2001a) showed
that the M.L.E of $\theta$ based on the un-grouped sample $X_1, X_2, ..., X_n$ is

$$\hat{\theta}_n = \frac{-n}{\sum_{i=1}^{n} \log(1 - e^{-X_i})}.$$  \hspace{1cm} (7)

Fisher’s Information Number

To find the Fisher’s information number contained in the grouped sample about $\theta$, find the expectation of the second derivative of the log-likelihood

$$\partial^2 \log f(n; \theta) = -\sum_{j=1}^{k} n_j \Psi_{1j}(\theta) - n_{k+1} \Psi_{2}(\theta)$$

where

$$\Psi_{1j}(\theta) = \frac{(A_{j+1} - A_j)^2 \cdot e^{\theta(A_{j+1} + A_{j+1})}}{(e^{\theta A_{j+1}} - e^{\theta A_j})^2}$$

and

$$\Psi_{2}(\theta) = \frac{A_{k+1}^2 \cdot e^{\theta A_{k+1}}}{(1 - e^{\theta A_{k+1}})^2}.$$  \hspace{1cm} (8)

If $I_G(\theta)$ denotes the Fisher’s information number from grouped data, then

$$I_G(\theta) = -E \left[ \frac{\partial^2 \log f(n; \theta)}{\partial \theta^2} \right],$$

and, because $E[N_j] = nP_j$,

$$I_G(\theta) = E \left[ \sum_{j=1}^{k} n_j \Psi_{1j}(\theta) + n_{k+1} \Psi_{2}(\theta) \right]$$

$$= n \sum_{j=1}^{k} (A_{j+1} - A_j)^2 \frac{e^{\theta(A_{j+1} + A_{j+1})}}{e^{\theta A_{j+1}} - e^{\theta A_j}} + n A_{k+1}^2 \frac{e^{\theta A_{k+1}}}{1 - e^{\theta A_{k+1}}}.$$  \hspace{1cm} (9)

Using $I_G(\theta)$, a large sample $(1 - \alpha)$100% confidence interval for $\theta$ can be found as follows:

$$\hat{\theta}_{MG} \pm Z_{1-\alpha/2} \sqrt{I_G(\hat{\theta}_{MG})^{-1}}.$$  \hspace{1cm} (10)

Simple calculations can show that the Fisher’s information number about $\theta$ in a random sample $X_1, X_2, ..., X_n$ from (1), that is, from un-grouped data, is $I(\theta) = n / \theta^2$.

Bayesian Estimation

In classical statistics, the likelihood of the data, $L(\theta) = f(n; \theta)$, is used to estimate the parameter $\theta$. In Bayesian statistics, it is assumed that $\theta$ has a prior distribution, for example, $\pi(\theta)$. The likelihood data is then combined with the prior distribution to obtain the posterior distribution $\pi(\theta | n)$. Herein, three loss functions based on the grouped data are used.

First the squared error loss function, that is a symmetric loss function, and with respect to it, the posterior mean is used to estimate $\theta$, for example $\hat{\theta}_{BSG}$, because it minimizes the posterior expected loss, $E[(\hat{\theta} - \theta)^2 | n]$. The second is the asymmetric LINEX loss function, which was introduced by Varian (1975). These loss functions were widely used by several authors; among of them Basu and Ebrahimi (1991), Pandey (1997), Soliman (2000) and Nassar and Eissa (2004). The LINEX loss function may be expressed as

$$L(\Delta) = b(e^e - c\Delta - 1), \quad c \neq 0, \quad b > 0,$$  \hspace{1cm} (11)

where $b$ and $c$ are the scale and shape parameters, respectively, and $\Delta = \hat{\theta} - \theta$. The sign and magnitude of the shape parameter $c$ reflects the direction and degree of asymmetry, respectively. (If $c > 0$, the overestimation is more serious than underestimation, and vice-versa). For $c$ close to zero, the LINEX loss is approximately squared error loss and therefore almost symmetric. The posterior expectation of the LINEX loss function equation (11) is

$$E[L(\hat{\theta} - \theta) | n] \approx e^e E[e^{-\theta} | n] - c(\hat{\theta} - E[\theta | n]) - 1.$$  \hspace{1cm} (12)

where $E[\cdot | n]$ denotes posterior expectation with respect to the posterior density of $\theta$. Based
on results from Zellner (1986), the (unique) Bayes estimator of \( \theta \), denoted by \( \hat{\theta}_{BLG} \) under the LINEX loss is the value \( \hat{\theta} \) which minimizes (12), is given by

\[
\hat{\theta}_{BLG} = -\frac{1}{c} \ln(\mathbb{E}[e^{-c\theta} | n]),
\]

provided that the expectation \( \mathbb{E}[e^{-c\theta} | n] \) exists and is finite (see Calabria & Pulcini, 1996), then the precautionary loss function introduced by Norstrom (1996) is used. Norstrom introduced an alternative asymmetric precautionary loss function and also presented a general class of precautionary loss function with the quadratic loss function as a special case. These loss functions approach infinitely near the origin to prevent underestimation, thus giving conservative estimators - especially when low failure rates are being estimated. These estimators are very useful when underestimation may lead to serious consequences. A useful and simple asymmetric precautionary loss function is

\[
\theta = \begin{cases} 
\theta & \text{if } \theta > 0 \\
0 & \text{if } \theta = 0 \\
\infty & \text{if } \theta < 0
\end{cases}
\]

Using the Binomial theorem, the likelihood function of the grouped data is re-written as follows:

\[
f(n; \theta) = C \prod_{j=1}^{k} \sum_{r_j=0}^{n_j} \binom{n_j}{r_j} (-1)^{r_j} \left(e^{\theta A_j}\right)^{r_j} \cdot \left(\frac{\theta}{\alpha} - \beta\right)^{\alpha-1} e^{-\theta - \beta} \left(\frac{1}{\beta} + H(r_j)\right)^{-\alpha}
\]

where

\[
H(r_j, \ldots, r_{k+1}) = r_{k+1} A_{k+1} + \sum_{j=1}^{k} (n_j - r_j) A_j + \sum_{j=1}^{k} r_j A_j.
\]

Combining the likelihood information with the prior information yields the posterior distribution of \( \theta \) given \( n \),

\[
\pi(\theta | n) = \frac{\int f(n; \theta) \cdot \pi(\theta) \, d\theta}{\int f(n; \theta) \cdot \pi(\theta) \, d\theta}
\]

resulting in

\[
\pi(\theta | n) = \frac{\sum_{r_j=0}^{n_j} \sum_{r_{k+1}=0}^{n_{k+1}} \binom{n_j}{r_j} \binom{n_{k+1}}{r_{k+1}} (-1)^{r_j + \ldots + r_{k+1}} \theta^{\alpha-1} e^{-\theta - \beta} \left(\frac{1}{\beta} + H(r_j, \ldots, r_{k+1})\right)^{-\alpha}}{\Gamma(\alpha) \sum_{r_j=0}^{n_j} \sum_{r_{k+1}=0}^{n_{k+1}} \binom{n_j}{r_j} \binom{n_{k+1}}{r_{k+1}} (-1)^{r_j + \ldots + r_{k+1}} \left(\frac{1}{\beta} + H(r_j, \ldots, r_{k+1})\right)^{-\alpha}}
\]

where \( H = H(r_1, \ldots, r_{k+1}) \).
The Bayesian estimate of \( \theta \) with respect to the squared error loss function, based on the grouped data, is the posterior mean,

\[
\hat{\theta}_{\text{BSG}} = E(\theta | \mathbf{n}) = \alpha \sum_{r_j=0}^{n_1} \cdots \sum_{r_k=0}^{n_{k+1}} \left( \frac{n_1}{r_1} \cdots \frac{n_{k+1}}{r_{k+1}} \right) (-1)^{r_1+\cdots+r_{k+1}} \left( \frac{1}{\beta} - H \right)^{-\alpha - 1}
\]

(18)

Kundu and Gupta (2008) obtained the Bayes estimator of \( \theta \), based on the un-grouped data, only under square error loss function as follows

\[
\hat{\theta}_{\text{BS}} = \frac{\alpha + n}{\beta^{-1} - \sum_{i=1}^{n} \log(1-e^{-x_i})}, \quad (19)
\]

Note that, for the non-informative prior - when \( \alpha = \beta^{-1} = 0 \) - the above Bayes estimator (\( \hat{\theta}_{\text{BS}} \)) and MLE of \( \theta \) from the un-grouped data (\( \hat{\theta}_d \)), are identical.

The Bayesian estimate of \( \theta \) with respect to the LINEX loss function, based on the grouped data, is obtained as follows:

\[
\hat{\theta}_{\text{BLG}} = E\left( \frac{1}{\theta} | \mathbf{n} \right)
\]

\[
= -\frac{1}{c} \ln \left( E\left[ e^{-\theta} | \mathbf{n} \right] \right)
\]

\[
= -\frac{1}{c} \ln \left( \frac{\sum_{r_j=0}^{n_1} \cdots \sum_{r_k=0}^{n_{k+1}} \left( \frac{n_1}{r_1} \cdots \frac{n_{k+1}}{r_{k+1}} \right) (-1)^{r_1+\cdots+r_{k+1}} \left( \frac{1}{\beta} - H \right)^{-\alpha}}{\sum_{r_j=0}^{n_1} \cdots \sum_{r_k=0}^{n_{k+1}} \left( \frac{n_1}{r_1} \cdots \frac{n_{k+1}}{r_{k+1}} \right) (-1)^{r_1+\cdots+r_{k+1}} \left( \frac{1}{\beta} - H \right)^{-\alpha}} \right)
\]

(20)

Here, the Bayes estimator of \( \theta \) is obtained under the LINEX loss function based on the un-grouped data. It may be shown that the posterior density of \( \theta \) based on the un-grouped data is gamma with the shape and scale parameters as \( \alpha + n \) and \( \beta^{-1} - \sum_{i=1}^{n} \log(1-e^{-x_i}) \), respectively.

\[
E[e^{-\theta} | \mathbf{x}] = \left( \frac{1}{\beta - \sum_{i=1}^{n} \log(1-e^{-x_i})} \right)^{-\alpha - n}
\]

thus, the Bayes estimator of \( \theta \) under the LINEX loss function based on the un-grouped data is as follows

\[
\hat{\theta}_{\text{BL}} = \frac{\alpha + n}{c} \ln \left( \frac{1 + \frac{c}{\beta^{-1} - \sum_{i=1}^{n} \log(1-e^{-x_i})}}{1} \right).
\]

(21)

The Bayesian estimate of \( \theta \) with respect to the precautionary loss function, based on the grouped data, is obtained as follows:

\[
\hat{\theta}_{\text{BPG}} = \frac{E(\theta | \mathbf{n})}{\sqrt{E(\theta^{-1} | \mathbf{n})}}
\]

\[
= \left( \frac{\alpha(\alpha - 1) \sum_{r_j=0}^{n_1} \cdots \sum_{r_k=0}^{n_{k+1}} \left( \frac{n_1}{r_1} \cdots \frac{n_{k+1}}{r_{k+1}} \right) (-1)^{r_1+\cdots+r_{k+1}} \left( \frac{1}{\beta} - H \right)^{-\alpha - 1}}{\sum_{r_j=0}^{n_1} \cdots \sum_{r_k=0}^{n_{k+1}} \left( \frac{n_1}{r_1} \cdots \frac{n_{k+1}}{r_{k+1}} \right) (-1)^{r_1+\cdots+r_{k+1}} \left( \frac{1}{\beta} - H \right)^{-\alpha + 1}} \right)^{1/2}
\]

(22)

Also, the Bayes estimator of \( \theta \) is obtained under the precautionary loss function based on the un-grouped data. Whereas

\[
\theta | \mathbf{x} \sim \Gamma(\alpha + n, \beta^{-1} - \sum_{i=1}^{n} \log(1-e^{-x_i}))
\]

it may be shown that

\[
E(\frac{1}{\theta} | \mathbf{x}) = \frac{\beta^{-1} - \sum_{i=1}^{n} \log(1-e^{-x_i})}{(\alpha + n - 1)}
\]

thus, the Bayes estimator of \( \theta \) under the precautionary loss function based on the un-grouped data is as follows
\[ \hat{\theta}_{bp} = \frac{\sqrt{(\alpha+n)(\alpha+n-1)}}{\beta} - \sum_{i=1}^{n} \log(1-e^{-\beta_1}) \]  

(23)

Methodology

Simulation Study

The estimator \( \hat{\theta}_M \) is the Maximum Likelihood Estimator (MLE) of the shape parameter of the generalized exponential distribution obtained from the un-grouped data; whereas, \( \hat{\theta}_{MG} \) is the MLE of \( \theta \) obtained from the grouped data. \( \hat{\theta}_{BS}, \hat{\theta}_{BP} \) and \( \hat{\theta}_{BL} \) are Bayes estimators under squared-error, precautionary and LINEX loss functions, respectively, based on un-grouped data. Also, \( \hat{\theta}_{BSG}, \hat{\theta}_{BPG} \) and \( \hat{\theta}_{BLG} \) are Bayes estimators under squared-error, precautionary and LINEX loss functions, respectively, based on grouped data. The notation CLG is used to denote the confidence length for \( \theta \) based on the grouped data. Because the large sample properties of these estimators are unknown, the bootstrap method can be used to assess the precision of estimates, but construction of bootstrap confidence interval is computationally more demanding than asymptotic confidence interval. Therefore, the 95% confidence interval is computed based on the MLE’s. The main goal is to compare the estimators in terms of biases and MSE’s.

As noted, \( \hat{\theta}_{MG} \) and hence its MSE cannot be put in a convenient closed form. Therefore, MSE’s of the estimators are empirically evaluated based on a Monte-Carlo simulation study of 1,000 samples by MATLAB mainly for small sample sizes. The simulation study was carried out for \( \theta = 1 \) with sample sizes \( n = 6, 9, 12, 15, 18 \) and 20. These samples were placed into five intervals \( (k = 4) \) with \( \delta = 1 \). The loss and prior parameters are arbitrarily taken as \( c = -1.5, -1, -0.5, 0.5, 1 \) and 1.5, \( \alpha = 2 \) and \( \beta = 0.5 \). Results are summarized in Table 1-2 and Figures 1-5.

| Table 1: Bias and MSE of the MLE’s and Three Bayes Estimates from Un-grouped Data, when \( k = 4, \delta = 1, \theta = 1, \alpha = 2 \) and \( \beta = 0.5 \) (MSE in parenthesis) |
|---|---|---|---|---|---|---|---|
| \( n \) | \( \hat{\theta}_M \) | \( \hat{\theta}_{BS} \) | \( \hat{\theta}_{BP} \) | \( \hat{\theta}_{BL} \) |
| 6 | 0.2569 (0.4697) | 0.1297 (0.1419) | 0.0567 (0.1126) | 0.2896 (0.3063) | 0.2288 (0.2314) | 0.1761 (0.1792) | 0.1671 (0.1570) | 0.0882 (0.1150) | 0.0508 (0.0957) | 0.0269 (0.0821) |
| 9 | 0.1551 (0.1893) | 0.1025 (0.0963) | 0.0512 (0.0806) | 0.2031 (0.1655) | 0.1666 (0.1366) | 0.1333 (0.1139) | 0.0740 (0.0825) | 0.0475 (0.0719) | 0.0227 (0.0638) |
| 12 | 0.1096 (0.1205) | 0.0850 (0.0751) | 0.0462 (0.0665) | 0.1357 (0.1118) | 0.1206 (0.0970) | 0.1071 (0.0849) | 0.0580 (0.0672) | 0.0412 (0.0609) | 0.0199 (0.0559) |
| 15 | 0.0929 (0.0918) | 0.0730 (0.0642) | 0.0409 (0.0571) | 0.1304 (0.0902) | 0.1103 (0.0800) | 0.0912 (0.0714) | 0.0506 (0.0581) | 0.0389 (0.0531) | 0.0187 (0.0490) |
| 18 | 0.0704 (0.0728) | 0.0570 (0.0547) | 0.0303 (0.0497) | 0.1036 (0.0722) | 0.0874 (0.0654) | 0.0719 (0.0596) | 0.0427 (0.0504) | 0.0289 (0.0469) | 0.0156 (0.0439) |
| 20 | 0.0629 (0.0668) | 0.0519 (0.0510) | 0.0277 (0.0469) | 0.0934 (0.0658) | 0.0791 (0.0601) | 0.0652 (0.0553) | 0.0390 (0.0474) | 0.0265 (0.0443) | 0.0144 (0.0417) |
Table 2: Bias and MSE of the MLE’s and Three Bayes Estimates from the Grouped Data, when $k=4$, $\delta=1$, $\theta=1$, $\alpha=2$ and $\beta=0.5$ (MSE in parenthesis)

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\hat{\theta}_{MG}$</th>
<th>$\hat{\theta}_{BSG}$</th>
<th>$\hat{\theta}_{BPG}$</th>
<th>$\hat{\theta}_{BLG}$</th>
<th>CLG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.3274 (0.6874)</td>
<td>0.1386 (0.1280)</td>
<td>0.0079 (0.1038)</td>
<td>0.4326 (0.3930)</td>
<td>0.3127 (0.2575)</td>
</tr>
<tr>
<td>9</td>
<td>0.1817 (0.3099)</td>
<td>0.0983 (0.1025)</td>
<td>-0.0033 (0.0925)</td>
<td>0.3032 (0.2379)</td>
<td>0.2237 (0.1731)</td>
</tr>
<tr>
<td>12</td>
<td>0.1205 (0.2380)</td>
<td>0.0726 (0.1010)</td>
<td>-0.0104 (0.0880)</td>
<td>0.2299 (0.1899)</td>
<td>0.1707 (0.1492)</td>
</tr>
<tr>
<td>15</td>
<td>0.0896 (0.1732)</td>
<td>0.0708 (0.0868)</td>
<td>0.0338 (0.0819)</td>
<td>0.1917 (0.1446)</td>
<td>0.1588 (0.1183)</td>
</tr>
<tr>
<td>18</td>
<td>0.0885 (0.1560)</td>
<td>0.0635 (0.0837)</td>
<td>0.0016 (0.0800)</td>
<td>0.1725 (0.1363)</td>
<td>0.1327 (0.1152)</td>
</tr>
<tr>
<td>20</td>
<td>0.0619 (0.1312)</td>
<td>0.0443 (0.0783)</td>
<td>-0.0133 (0.0729)</td>
<td>0.1418 (0.1156)</td>
<td>0.1066 (0.0996)</td>
</tr>
</tbody>
</table>

Figure 1: MSE’s of the Estimators $\hat{\theta}_M$, $\hat{\theta}_{BS}$, $\hat{\theta}_{BP}$ and $\hat{\theta}_{BL}$ from the Un-grouped Data, Based on Table 1

MSE's of the estimators from the un-grouped data when $c = -1 \& 1$
Figure 2: MSE’s of the Estimators $\hat{\theta}_{MG}$, $\hat{\theta}_{BSG}$, $\hat{\theta}_{BPG}$ and $\hat{\theta}_{BLG}$ from the Grouped Data, Based on Table 2

![Graph showing MSE’s of the estimators from the Grouped data when $c = -1$ & 1]

Figure 3: MSE’s of the Bayes Estimators under Three Loss Functions Based on the Un-grouped and Grouped Data (when the overestimation is more serious than underestimation, $c = 1$)

![Graph showing MSE’s of the Bayes estimators based on both Grouped and un-grouped data when $c = 1$]
Figure 4: MSE’s of the Estimators as Function of Loss Parameter Based on the Un-grouped Data when \( n = 20 \)

MSE’s of the estimators from un-grouped data as function of loss parameter, when \( n = 20 \)

Figure 5: MSE’s of the Estimators as Function of Loss Parameter Based on the Grouped Data when \( n = 20 \)

MSE’s of the estimators from Grouped data as function of loss parameter, when \( n = 20 \)
Conclusion
In this study, the Bayesian and non-Bayesian estimators for the shape parameter of the generalized exponential distribution based on grouped and un-grouped data were obtained. In addition, the interval estimator for $\theta$ was considered when the data are given in groups. The Bayes estimators under symmetric squared error loss function and asymmetric precautionary and LINEX loss functions were derived. Observations regarding the results are stated in the following points.

Results Based on Un-Grouped Data
Table 1 shows that all of the estimators are overestimations. Also, it is noted that derived estimators are consistent, because both bias and MSE decrease as the sample size increases. The Bayes estimates have the smallest estimated MSE’s compared with the classical estimate. The Bayes estimates under the precautionary loss function ($\hat{\theta}_{BP}$) always have the smallest estimated MSE’s as compared with the Bayes estimates under Squared error loss function, $\hat{\theta}_{BS}$ (see Figures 1 and 4). When the underestimation is more serious than overestimation (for $c \leq 0.5$), the performance of the Bayes estimates under precautionary loss function ($\hat{\theta}_{BP}$) are better than the rest: however, when the overestimation is more serious than underestimation (for $c > 0.5$), the performance of the Bayes estimates under LINEX loss function, $\hat{\theta}_{BL}$, are better than the rest (see Figures 2 and 5). Otherwise, the confidence intervals work quite well.

General Conclusions
In general, when the data are given in groups, the proposed Bayes estimators ($\hat{\theta}_{BG}$) are more efficient than the corresponding Bayes estimators based on un-grouped data ($\hat{\theta}_{B}$) for very small sample sizes, thus these estimators work very well (see Figure 3). Also, whereas the proposed Bayes estimators are better than the proposed estimators by Gupta and Kundu (1999, 2001a, 2008), it is suggested that the Bayes estimators be used for estimating the shape parameter of GE distribution when the data are given in groups, for example, in life testing experiments.

References
GENERALIZED EXPONENTIAL ESTIMATIONS USING GROUPED DATA


Ranked Set Sampling Using Auxiliary Variables of a Randomized Response Procedure for Estimating the Mean of a Sensitive Quantitative Character

Carlos N. Bouza
Universidad de la Habana
Havana, Cuba

The analysis of the behavior of estimators of the mean of a sensitive variable is considered when a randomized response procedure is used. The results deal with the inference based on simple random sampling with replacement study design. A study of the behavior of the procedures for a ranked set sampling design is developed. A gain in accuracy is generally associated with the proposed alternative model.

Key words: Order statistics, expected model variance, unbiasedness.

Introduction
Let Y be a sensitive variable evaluated in a finite population U = \{u_1, \ldots, u_N\}. The individual u_i with a value of Y that carries a stigma will tend to give incorrect information or to refuse to answer. It is well known that, when dealing with sensitive questions, researchers face the need to reduce response refusals as well as response bias. A possibility is to replace a direct response to a sensitive question by using a random response query.

The seminal work on this issue was conducted by Warner (1965), who dealt with a qualitative question with possible responses of yes or no when one question is identified to have the stigma. The goal of the surveyor is to estimate the probability of having the stigma. It is expected that a large percent of the persons bearing it will either lie or refuse to answer. Warner’s method consists of placing the question associated with the stigma together with some insensitive questions. The respondent randomly chooses a question and answers it without revealing which was selected. When dealing with a quantitative character, similar reasoning can be used. Chaudhuri-Stenger (1992) proposed the use of a randomized responses (RR) model and they developed an approach for designing a general RR procedure when the stigma is related with a quantitative character. See, for example, Singh-Singh (1993) and Zou (1997), who analyzed the behavior of various estimators of the mean when simple random sampling is used for selecting the sample.

Ranked set sampling (RSS) is an alternative sample design that provides gains in accuracy with respect to simple random sampling with replacement (SRSWR). It was proposed by McIntyre (1952) to estimate mean pasture yield, and it was found to be more efficient than selecting the sample using a simple random sampling (SRS) design.

The units can be ranked by means of a cheap procedure and selecting an order statistic from each of the independent samples selected using SRS with replacement (SRSWR). Results indicated that the use of ranked set sampling is highly beneficial and leads to estimators that are more precise than the usual sample mean per unit. The method is now referred to as the ranked set sampling (RSS) method. (See Patil (2002) and Patil, et al. (1994, 1999) for detailed discussions.) This research developed a study of the use of alternative RR procedures when RSS was used instead of SRS with replacement (SRSWR) under a model.
A Model Based Randomized Responses Procedure Under SRSSWRR

The RR procedure proposed in Chaudhuri-Stenger (1992) begins with an individual \(u_i \in U\) and the sets of known variables \(A = \{A_1, \ldots, A_T\}\) and \(B = \{B_1, \ldots, B_S\}\). After these are fixed the following are calculated:

\[
\begin{align*}
\mu_A &= \frac{\sum_{t=1}^{T} A_t}{T} \\
\sigma_A^2 &= \frac{\sum_{t=1}^{T} (A_t - \mu_A)^2}{T}, \\
\mu_B &= \frac{\sum_{s=1}^{S} B_s}{S}, \\
\sigma_B^2 &= \frac{\sum_{s=1}^{S} (B_s - \mu_B)^2}{S}.
\end{align*}
\]

After \(u_i\) is selected from the population \(U = \{1, \ldots, N\}\) the respondent will not report the value of \(Y_i\) directly. Instead a random experiment is performed and its result is the independent selection of an \(a_i \in A\) and a \(b_i \in B\), \((a_i, b_i)\). The report made by the interviewee is: \(Z_i = a_iY_i + b_i\).

The first procedure for deriving information on \(Y_i\) is to use the report to compute

\[
R_i = \frac{Z_i - \mu_B}{\mu_A}.
\]

The model expectation is

\[
E(R_i) = Y_i
\]

and the corresponding variance is

\[
V(R_i) = \frac{\sigma_A^2 + \sigma_B^2}{\mu_A^2} = V_i.
\]

The independence of the selection, provided by the use of SRSWRR, supports that the model variance of (2.1) is:

\[
\sum_{i=1}^{n} E_d V(R_i) = \frac{\sum_{i=1}^{n} V_i^2}{n^2}.
\]

Ranked Set Sampling for the Randomized Responses Procedures

To implement RSS, \(m\) independent samples of size \(n\) are selected using SRSWRR. The units in each sample are ranked without knowing the value of \(Y_i\). Either personal judgment or the evaluation of a covariate \(X\) that is correlated with \(Y\) may be used to rank \(Y\) and the individual with rank \(i\) in the ranked sample \(s(i)\) is interviewed. The procedure is repeated independently \(r\) times (cycles) and \(n = mr\). David-Levine (1972) studied the effect of judgmental errors and established that it does not affect the properties of RSS.

Suppose that the ranking is made on \(Y\). The sampler asks interviewee \(u_i\) to randomly select \(A_i\) and \(B_i\). The report of the \(i^{th}\) ranked sample in the \(t^{th}\) cycle is: \(Z_{i(0)t} = A_iY_{i(0)t} + B_i\). Using this report, the following can be computed for each \(u_i\):
ESTIMATING THE MEAN USING RANKED SET SAMPLING RANDOMIZED RESPONSE

\[ R_{(i)t} = \frac{Z_{(i)t} - \mu_B}{\mu_A} \]

and, if its expectation is \( E_R(R_{(i)t}) = Y_{(i)t} \), then

\[ \bar{R}_t = \sum_{i=1}^{m} R_{(i)t} \]

and its unbiasedness may be derived using:

\[ E_d E_R(\bar{R}_t) = \sum_{i=1}^{m} \frac{\mu_{Y(i)}}{m} = \mu_Y. \]

From these results an estimator was derived that uses the information provided in the \( r \) cycles and is unbiased. Here it is proposed that

\[ R_{(rss)} = \frac{\sum_{i=1}^{r} \sum_{i=1}^{m} R_{(i)t}}{rm} \] (3.1)

be used. The fact that

\[ V_R(R_{(i)t}) = V_R\left( \frac{Z_{(i)t} - \mu_B}{\mu_A} \right) = Y_{(i)t}^2 \frac{\sigma_A^2 + \sigma_B^2}{\mu_A^2} \]

along with the independence of the involved variables are used for deriving the expected error of the RSS-estimator, which is:

\[ E_d V_R(\bar{R}_{(rss)}) = \frac{\sum_{i=1}^{m} (\sigma_{Y(i)}^2 + \mu_{Y(i)}^2) \sigma_A^2 + \sigma_B^2}{rm^2 \mu_A^2} \] (3.2)

given that \( E_d(Y_{(i)t}^2) = \sigma_{Y(i)}^2 + \mu_{Y(i)}^2 \).

The relation between \( \sigma_{Y(i)}^2 \), the variance of the \( i^{th} \) os, and the variance of the distribution \( \sigma_Y^2 \) permits (see Dell-Clutter, 1972): \( \sum_{i=1}^{m} \sigma_{Y(i)}^2 \)

Thus, (3.1) can be rewritten as:

\[ E_d V_R(\bar{R}_{(rss)}) = \frac{\sigma_A^2}{m} \left[ \frac{\sum_{i=1}^{m} \Delta_{Y(i)}^2}{n} + \frac{\sum_{i=1}^{m} \mu_{Y(i)}^2}{n \mu_A^2} \right] + \frac{\sigma_B^2}{n \mu_A^2} \]

and leads to the proposition 3.1.

Proposition 3.1

Consider the use of RRS for selecting a sample of \( n = rm \) individuals and the ranking with respect to \( Y \) of the reports \( Z_{(i)t}, i = 1, m, \) and \( t = 1, \ldots, r \). \( G(0, 1) \) represents a gain in accuracy measured by

\[ G(0,1) = \frac{\sigma_A^2}{n \mu_A^2} \left[ \sum_{i=1}^{m} \frac{\mu_{Y(i)}^2 - \mu_{Y(i)}^2}{m} + \sum_{i=1}^{m} \frac{\Delta_{Y(i)}^2}{m} \right], \]

\( \Delta_{Y(i)}^2 = (\mu_{Y(i)} - \mu_Y)^2 \), for estimating the mean using (3.1) instead of the SRSWR and the sample mean. This result is a natural extension of the classic RSS procedure (note that RSS is not necessarily more accurate than SRSWR).

The ranking may be implemented using the information provided by the selection of the auxiliary variables. The persons included in each sample randomly select the corresponding insensitive variables \( A \) and \( B \); they communicate their values to one another for establishing their ranks. The person in position \( j \) in the sample \( j \) gives the report. The procedure and the \( m \) independent samples in each cycle are evaluated. The report of an individual \( u_i \), is:

\[ Z_A(i) = A_{A(i)} Y_i + B_i \]

\[ Z_B(i) = A_{B(i)} Y_i + B_{B(i)} \]

If the ranks are made on \( A \), then the \( t^{th} \) cycle report is \( Z_{A(i)} = A_{A(i)} Y_i + B, \) and the response variable computed is

\[ Z(i) = \begin{cases} A_{A(i)} Y_i + B_i & \text{if } A \text{ denotes the order} \\ A_{B(i)} Y_i + B_{(i)} & \text{if } B \text{ denotes the order} \end{cases} \]
The model expectation is the value of the sensitive variable, \( E_R(R_{A(i)}) = Y \), and averaging the reports generates an unbiased estimation of the mean of \( Y \). The corresponding results are fixed in the Proposition 3.2.

**Proposition 3.2**

The use of RRS for selecting a sample of \( n = rm \) individuals and the ranking, with respect to \( A \), of the reports \( Z_{i(t)}, \ i = 1, \ldots, m \) and \( t = 1, \ldots, r \), represents a gain in accuracy when

\[
\sum_{i=1}^{m} \frac{1}{m \mu_{A(i)} \mu_A} + \sum_{i=1}^{m} \frac{\Delta_{A(i)}}{nm \mu_{A(i)} \mu_A} > \sum_{i=1}^{m} \frac{\mu_{A(i)}^2}{m \mu_{A(i)} \mu_A}
\]

with the estimator given by:

\[
\overline{R}_{A(rss)} = \frac{\sum_{i=1}^{m} \sum_{i=1}^{r} R_{A(i)t}}{rm}
\]

which is unbiased. The proof for this is as follows.

The unbiasedness of (3.3) follows from the fact that the reports are model unbiased for the corresponding \( Y_i \), and the arithmetic mean is design unbiased. Its model variance for the \( i^{th} \) os in the cycle \( t \) is:

\[
V_R(R_{A(i)}) = \frac{Y_i^2 \sigma_{A(i)}^2 + \sigma_B^2}{\mu_{A(i)}^2}
\]

where \( \sigma_{A(i)}^2 \) and \( \mu_{A(i)} \) are the variance and mean of \( A_{(i)} \). The design expectation of the model error for the \( i^{th} \) os is:

\[
E_d V_R(R_{A(i)}) = \frac{(\sigma_Y^2 + \mu_Y^2) \sigma_{A(i)}^2 + \sigma_B^2}{\mu_{A(i)}^2}
\]

and

\[
E_d V_R(R_{A(i)}) = \left( \sigma_Y^2 + \mu_Y^2 \right) \sigma_{A(i)}^2 + \sigma_B^2
\]

is the expected error of the estimator.

The relation between the variance of an os and the population variance is used again for rewriting the design expectation of the model variance of the estimator. The result derived is:

\[
E_d V_R(R_{A(rss)}) = \left( \sigma_Y^2 + \mu_Y^2 \right) [\frac{1}{n} \sum_{i=1}^{m} \frac{1}{m \mu_{A(i)}^2} - \frac{1}{nm} \sum_{i=1}^{m} \frac{\Delta_{A(i)}}{nm \mu_{A(i)}^2}] + \frac{1}{nm} \sum_{i=1}^{m} \sigma_B^2
\]

(3.4)

where \( \Delta_{A(i)} = \mu_{A(i)} - \mu_A \).

Comparing (3.4) with (2.2) the gain is

\[
G(0,2) = \frac{(\sigma_Y^2 + \mu_Y^2) \sigma_A^2 + \sigma_B^2}{n} \left[ \sum_{i=1}^{m} \frac{\mu_A^2 - \mu_A(i)^2 + \sum_{i=1}^{m} \frac{\Delta_{A(i)}}{nm \mu_{A(i)}^2}}{m \mu_{A(i)}^2 \mu_A^2} \right]
\]

which is positive only if the relation stated holds. Note that the conditions stated for granting use of the strategy characterized in the propositions are better than the SRSW strategy when

\[
\sum_{i=1}^{m} \frac{\mu_{C(i)} \mu_C(i)}{m \mu_{C(i)}^2 \mu_C(i)} > 0, \ C = Y, A
\]

is satisfied. The designer of the inquiry is able to fix the possible values of \( A \) and calculate the expectation of the different order statistics. Thus, it is possible to have a previous evaluation of the gain in accuracy when the model is based on ranking the auxiliary variable \( A \).

When the ranking is based on the results for \( B \) the report is \( Z_{B(i)t} = A_i Y_i + B_{(i)} \) and the following is computed


\[ R_{B(i)t} = \frac{Z_{B(i)t} - \mu_{B(i)}}{\mu_A}. \]

Once again the model expectation provides the true value of \( Y \), that is \( E\left( R_{B(i)t} \right) = Y_i \) and an unbiased estimation of \( \mu_Y \) is

\[ \bar{R}_{B(rss)} = \frac{\sum_{i=1}^{m} \sum_{t=1}^{r} R_{B(i)t}}{rm} \quad (3.5) \]

because \( E(Z_{B(i)}) = \mu_A Y_i + \mu_{B(i)} \). Denoting by \( \sigma_{B(i)}^2 \) and \( \mu_{B(i)} \) the variance and mean of the \( i^{th} \) os of \( B \) gives the design expectation of the error of the proposed estimator as:

\[
\begin{align*}
E_d(V_R) (\bar{R}_{B(rss)}) &= \left( \sigma_Y^2 + \mu_Y^2 \right) \left( \frac{\sigma_A^2}{n \mu_A^2} \right) + \sum_{i=1}^{m} \frac{\sigma_{B(i)}^2}{nm \mu_A^2} \\
&= \left( \sigma_Y^2 + \mu_Y^2 \right) \left( \frac{\sigma_A^2}{n \mu_A^2} \right) + \frac{\sigma_A^2}{n \mu_A^2} - \sum_{i=1}^{m} \frac{\Delta_{B(i)}^2}{nm \mu_A^2}
\end{align*}
\]

Therefore, to rank in \( B \) provides a gain in accuracy that does not depend on the differences between the mean of the variable used for ranking and expectation of the involved os's. Proposition 3.3 states this result.

Proposition 3.3

The use of RRS for selecting a sample of \( n = rm \) individuals and the ranking, with respect to \( B \), of the reports \( Z_{(i)t} \), \( i = 1, \ldots, m \) and \( t = 1,\ldots,r \), and of (3.5) for estimating the mean of \( Y \) is more accurate than the SRSWR strategy and the gain in accuracy is measured by \( \sum_{i=1}^{m} \Delta_{B(i)}^2 / m \), where \( \Delta_{B(i)} = \mu_{B(i)} - \mu_A \). Therefore the surveyor is able to increase the gain in accuracy by taking into account that the smaller the mean of \( A \), the larger the gain.

Comparison of the Different Alternatives

Different strategies for estimating the mean of a sensitive character are available because respondents may be ranked using \( Y \), \( A \) or \( B \). It is known that when \( Y \) is obtained by a direct response the RSS estimator of the mean is generally more accurate than SRSWR (see Patil, et al., 1999 for a clarifying discussion of this fact). The comparison of the accuracy of the estimators will be made analyzing the design expected model variances.

Comparing the SRSWR’s estimator and its RSS counterpart when the rank is made using the sampler’s judgment on \( Y \). Taking

\[ D_{rss,srs} = E_d(V_R(\bar{R})) - E_d(V_R(\bar{R}_{rss})) \]

results in

\[
D_{rss,srs} = \frac{\sum_{i=1}^{m} \Delta_{Y(i)}^2}{mn \mu_A^2} + \frac{\sigma_A^2}{n \mu_A^2} - \frac{\mu_Y^2}{m} \]

This difference depends heavily on the parameters of \( A \). Note that a large variance will provide a large value of the second term in \( D_{rss,srs} \). It may be argued that this fact allows the design of a RR procedure that diminishes the sample error by determining an adequate set \{\( A_1, \ldots, A_T \)\}.

In practice, ranking \( Y \) when it is a sensitive variable should be difficult. A more practical approach is that the sampled persons rank the selected \( A_i \)'s and provides the pair \( (A_i, \text{rank}(A_i)) \) to the sampler. Comparing the expected errors of (2.1) and (3.2) the following are obtained

\[
D_{A(rss),srs} = A^* + A^{**} - A^{***},
\]

where

\[
A^* = \frac{\sigma_Y^2 + \mu_Y^2}{nm \mu_A^2} \left[ \sum_{i=1}^{m} \left( \frac{\mu_{A(i)}^2 - \mu_A^2}{\mu_{A(i)}^2} \right) \right]
\]

\[
A^{**} = \frac{\sigma_B^2}{nm \mu_A^2} \left[ \sum_{i=1}^{m} \left( \frac{\mu_{B(i)}^2 - \mu_A^2}{\mu_{B(i)}^2} \right) \right]
\]

\[
A^{***} = \frac{\sigma_Y^2 + \mu_Y^2}{nm} \left[ \sum_{i=1}^{m} \left( \frac{\Delta_{A(i)}^2}{\mu_{A(i)}^2} \right) \right]
\]

252
The gain in accuracy of RSS thus depends on the magnitude of differences between the expectation of the m os and the population mean of the ranked variable $A$.

The ranking of $B$ produces the expression

$$\Delta_{(i)} = \mu_{(i)} - \mu_{B}$$

with

$$D_{rss,b(rss)} = -\sum_{i=1}^{m} \left( \frac{\Delta_{(i)}}{mn\mu_{A}^2} \right)$$

A comparison between two RSS alternatives may be developed by comparing their gains in accuracy. The ratio is proposed

$$D_{p,q} = \frac{D_{p/rss}}{D_{q/rss}} \quad (4.1)$$

Methodology

To evaluate the performance of each estimator Monte Carlo experiments were developed. The first set of experiments consisted in considering \{Y1 , ..., YN\}, \{A1, ..., AT\} and \{B1, ..., BS\} as independently distributed Uniform random variables in [0, 1]. After these sets were determined (4.1) was computed, $H = 1,000$ experiments were conducted and $G_{p,q} = \sum_{h=1}^{1,000} \Delta_{p,q,h}/1,000$ was calculated for the different combinations of $p$, $q = RSS$, A(RSS), B(RSS) and $m = 2,..,5$, $r = 1,..,5$; results are shown in Tables 4.1 - 4.3.

Results

Table 4.1 suggests that ranking on Y provides considerable gains in accuracy with respect to ranking on A. This fact should be generated by the role of A in the reports and the absence of errors in the ranking of Y. The gains seem to be increased when m is small with respect to r for a fixed sample size n=mr.

Table 4.2: Values of $G_{rss, b(rss)}$ for the U[0, 1] Distribution

<table>
<thead>
<tr>
<th>$r$</th>
<th>$m=2$</th>
<th>$m=3$</th>
<th>$m=4$</th>
<th>$m=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.530</td>
<td>0.260</td>
<td>0.179</td>
<td>0.135</td>
</tr>
<tr>
<td>2</td>
<td>0.537</td>
<td>0.274</td>
<td>0.167</td>
<td>0.141</td>
</tr>
<tr>
<td>3</td>
<td>0.500</td>
<td>0.266</td>
<td>0.275</td>
<td>0.186</td>
</tr>
<tr>
<td>4</td>
<td>0.524</td>
<td>0.258</td>
<td>0.175</td>
<td>0.139</td>
</tr>
<tr>
<td>5</td>
<td>0.588</td>
<td>0.282</td>
<td>0.200</td>
<td>0.118</td>
</tr>
</tbody>
</table>

Table 4.3: Values of $G_{b(rss), A(rss)}$ for the U[0, 1] Distribution

<table>
<thead>
<tr>
<th>$r$</th>
<th>$m=2$</th>
<th>$m=3$</th>
<th>$m=4$</th>
<th>$m=5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.377</td>
<td>0.310</td>
<td>0.264</td>
<td>0.254</td>
</tr>
<tr>
<td>2</td>
<td>0.934</td>
<td>0.310</td>
<td>0.264</td>
<td>0.253</td>
</tr>
<tr>
<td>3</td>
<td>0.971</td>
<td>0.310</td>
<td>0.258</td>
<td>0.257</td>
</tr>
<tr>
<td>4</td>
<td>1.111</td>
<td>0.310</td>
<td>0.203</td>
<td>0.251</td>
</tr>
<tr>
<td>5</td>
<td>0.746</td>
<td>0.310</td>
<td>0.252</td>
<td>0.246</td>
</tr>
</tbody>
</table>

Another set of experiments was developed considering an exponential with $\lambda = 1$ for generating the involved variables; results are shown in Tables 4.4 - 4.6. Table 4.4 suggests a
preference for ranking on Y. The gains in accuracy are larger than when the distribution is uniform.

The results shown in Table 4.5 establish that the ranking using B may be better than to rank using Y for m > 3; Table 4.6 indicates the preference for using B instead of A.

### Table 4.4: Values of G_{rss,A(rss)} for the Exp[1] Distribution

<table>
<thead>
<tr>
<th>.r</th>
<th>.m=2</th>
<th>.m=3</th>
<th>.m=4</th>
<th>.m=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.072</td>
<td>0.012</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>2</td>
<td>0.062</td>
<td>0.021</td>
<td>0.023</td>
<td>0.019</td>
</tr>
<tr>
<td>3</td>
<td>0.075</td>
<td>0.014</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td>4</td>
<td>0.077</td>
<td>0.016</td>
<td>0.016</td>
<td>0.012</td>
</tr>
<tr>
<td>5</td>
<td>0.093</td>
<td>0.012</td>
<td>0.015</td>
<td>0.016</td>
</tr>
</tbody>
</table>

### Table 4.5: Values of G_{rss,B(rss)} for the Exp[1] Distribution

<table>
<thead>
<tr>
<th>.r</th>
<th>.m=2</th>
<th>.m=3</th>
<th>.m=4</th>
<th>.m=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.688</td>
<td>0.342</td>
<td>1.234</td>
<td>1.413</td>
</tr>
<tr>
<td>2</td>
<td>0.672</td>
<td>0.543</td>
<td>1.340</td>
<td>1.238</td>
</tr>
<tr>
<td>3</td>
<td>0.700</td>
<td>0.320</td>
<td>1.467</td>
<td>1.350</td>
</tr>
<tr>
<td>4</td>
<td>0.683</td>
<td>0.345</td>
<td>1.650</td>
<td>1.379</td>
</tr>
<tr>
<td>5</td>
<td>0.780</td>
<td>0.300</td>
<td>1.350</td>
<td>1.626</td>
</tr>
</tbody>
</table>

### Table 4.6: Values of G_{B(rss),A(rss)} for the Exp[1] Distribution

<table>
<thead>
<tr>
<th>.r</th>
<th>.m=2</th>
<th>.m=3</th>
<th>.m=4</th>
<th>.m=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.105</td>
<td>0.035</td>
<td>0.013</td>
<td>0.013</td>
</tr>
<tr>
<td>2</td>
<td>0.092</td>
<td>0.039</td>
<td>0.017</td>
<td>0.015</td>
</tr>
<tr>
<td>3</td>
<td>0.107</td>
<td>0.043</td>
<td>0.012</td>
<td>0.011</td>
</tr>
<tr>
<td>4</td>
<td>0.113</td>
<td>0.043</td>
<td>0.010</td>
<td>0.015</td>
</tr>
<tr>
<td>5</td>
<td>0.119</td>
<td>0.034</td>
<td>0.011</td>
<td>0.009</td>
</tr>
</tbody>
</table>

### Conclusions

The results of these experiments indicate that the best strategy is to have the sampled persons communicate only the value of B and to rank its values accordingly. This indication is sustained by the usual difficulty for obtaining a perfect ranking of Y.

### References


Symmetry Plus Quasi Uniform Association Model and Its Orthogonal Decomposition for Square Contingency Tables

Kouji Yamamoto  
Osaka University Hospital,  
Suita City, Japan

Sadao Tomizawa  
Tokyo University of Science,  
Noda City, Japan

A model is proposed having the structure of both symmetry and quasi-uniform association (SQU model) and provides a decomposition of the SQU model. It is also shown with examples that the test statistic for goodness-of-fit of the SQU model is asymptotically equivalent to the sum of those for the decomposed models.

Key words: Likelihood ratio statistic, marginal homogeneity, quasi-symmetry, quasi-uniform association, separability.

Introduction
For the \( r \times r \) square contingency table, let \( p_{ij} \) denote the probability that an observation will fall in the \( i \)th row and \( j \)th column of the table (\( i = 1, \ldots, r; j = 1, \ldots, r \)). For the analysis of two-way contingency tables with ordered categories, Goodman (1979) considered some association models, for example, the uniform association model, which is a generalization of the independence model. Goodman (1979) also observed that regular multiplicative models for ordinal variables fit square contingency tables well when the cells on the main diagonal are ignored, thus, he proposed the quasi-uniform association (QU) model, defined by

\[
p_{ij} = \begin{cases} 
\mu \alpha_i \beta_j \psi_i (i \neq j), \\
\psi_i (i = j).
\end{cases}
\]

when \( \theta = 1 \), this model is the quasi-independence (QI) model (Bishop, Fienberg & Holland, 1975, p. 178).

The symmetry (S) model considered by Bowker (1948) is defined by

\[
p_{ij} = \psi_{ij} (i = 1, \ldots, r; j = 1, \ldots, r),
\]

where \( \psi_{ij} = \psi_{ji} \) (Bishop, et al., 1975, p. 282). This model describes a structure of symmetry of the cell probabilities \( \{p_{ij}\} \) with respect to the main diagonal of the table. The quasi-symmetry (QS) model considered by Caussinus (1965) is defined by

\[
p_{ij} = \mu \alpha_i \beta_j \psi_{ij} (i = 1, \ldots, r; j = 1, \ldots, r),
\]

where \( \psi_{ij} = \psi_{ji} \). The odds ratio for rows \( i \) and \( j \) (\( i > j \)), and columns \( s \) and \( t \) (\( s > t \)) are denoted by \( \theta_{(ij;st)} \); thus,

\[
\theta_{(ij;st)} = \frac{(p_{ij}p_{st})}{(p_{is}p_{jt})}.
\]

Using odds ratios, the QS model may be expressed as

\[
\theta_{(ij;st)} = \theta_{(st;ij)} (i < j; s < t).
\]
ORTHOGONAL DECOMPOSITION FOR QUASI UNIFORM ASSOCIATION MODEL

Therefore this model indicates a structure of symmetry with respect to the odds ratios. A special case of this model obtained by putting \( \{\alpha_i = \beta_i\} \) is the S model. Also, each of the QI and QU models is a special case of the QS model. The marginal homogeneity (MH) model is defined by

\[
p_i = p_i \quad (i = 1, \ldots, r),
\]

where \( p_i = \sum_{a=1}^{r} p_{ia} \) and \( p_i = \sum_{a=1}^{r} p_{ai} \) (Stuart, 1955).

Decomposition of the S Model (Caussinus, 1965)

Theorem 1

The S model holds if and only if both the QS and MH models hold.

The symmetry plus quasi-independence (SQI) model introduced by Goodman (1985) is defined by

\[
p_{ij} = \begin{cases} 
\mu \alpha_i \alpha_j & (i \neq j), \\
\psi_{ii} & (i = j).
\end{cases}
\]

This model is a special case of the S model obtained by substituting \( \{\psi_{ij} = \alpha_i \alpha_j\} \) for \( i \neq j \).

The purpose of this study is to: (1) propose a model that can be used to simultaneously study both symmetry and quasi-uniform association (the SQU model), (2) provide decomposition of the proposed model, and (3) show the orthogonality of decomposition with respect to the goodness-of-fit test statistic.

Proposed Model

Consider a model defined by:

\[
p_{ij} = \begin{cases} 
\mu \alpha_i \alpha_j \theta^i & (i \neq j), \\
\psi_{ii} & (i = j).
\end{cases}
\]

This model indicates that both the S and QU models hold simultaneously. Thus, this model shall be referred to as the symmetry plus quasi-uniform association (SQU) model. The SQU model is an extension of the SQI model. Under the SQU model, the row marginal distribution is identical with the column marginal distribution. Using odds ratios, the SQU model may be expressed as

\[
\theta_{(i;j:t)} = \theta^{(i-t)(i-t)} \quad (i \neq s, i \neq t, j \neq s, j \neq t).
\]

This model has uniform local association for cells off the main diagonal of the table. Figure 1 shows the relationships among the models.

Decompositions of the Models

Theorem 2

The SQU model holds if and only if both the QU and MH models hold.

Theorem 2 Proof

If the SQU model holds, then the QU and MH models hold. Conversely, if both the QU and MH models hold, then the QS model holds. Therefore, from Theorem 1 it may be stated that the S model holds. Thus, SQU model holds. The proof is completed and the following corollary is obtained because the SQI model is a special case of the SQU model with \( \theta = 1 \).

Corollary 1

The SQI model holds if and only if both the QI and MH models hold.

Orthogonality of Decomposition of Test Statistic for the Models

Let \( n_{ij} \) denote the observed frequency in the \((i,j)\)th cell of the table \((i = 1, \ldots, r; j = 1, \ldots, r)\). Assume that a multinomial distribution applies to the \(r \times r\) table. The maximum likelihood estimates of
expected frequencies under the models described in this paper could be obtained using an iterative procedure, for example, the general iterative procedure for log-linear models of Darroch and Ratcliff (1972) or using the Newton-Raphson method to the log-likelihood equations.

Each model can be tested for goodness-of-fit by, for example, the likelihood ratio Chi-squared statistic with the corresponding degrees of freedom (df). The numbers of df for the SQU, QU, and MH models are \( r^2 - 2r - 1 \), \( r(r - 3) \), and \( r - 1 \), respectively. Let \( G^2(\Omega) \) denote the likelihood ratio statistic for testing the goodness-of-fit of model \( \Omega \). Thus

\[
G^2(\Omega) = 2 \sum_{i=1}^{r} \sum_{j=1}^{r} n_{ij} \log \left( \frac{n_{ij}}{\hat{m}_{ij}} \right),
\]

where \( \hat{m}_{ij} \) is the maximum likelihood estimate of expected frequency \( m_{ij} \) under model \( \Omega \).

For the analysis of contingency tables, Lang and Agresti (1994) and Lang (1996) considered the simultaneous modeling of the joint distribution and of the marginal distribution. Aitchison (1962) discussed the asymptotic separability, which is equivalent to the orthogonality in Read (1977) and the independence in Darroch and Silvey (1963) of the test statistic for goodness-of-fit of two models (also see Lang & Agresti, 1994; Lang, 1996; Tomizawa & Tahata, 2007; Tahata & Tomizawa, 2008).

**Theorem 3**

The following asymptotic equivalence holds:

\[
G^2(SQU) = G^2(QU) + G^2(MH). \quad (1)
\]

The number of df for the SQU model equals the sum of the numbers of df for the QU and MH models.

**Theorem 3 Proof**

The QU model may be expressed in a log-linear form

\[
\log p_{ij} = \mu + \alpha_i + \beta_j + ij\theta + \psi_{ij} \quad (i = j) \quad (i = 1, \ldots, r; j = 1, \ldots, r),
\]

where \( \mu = \log \mu_i \), \( \alpha_i = \log \alpha_i \) (and so on) with \( \alpha_i = \beta_r = 0 \) without loss of generality, and where \( I(i = j) = 1 \) if \( i = j \) and 0 otherwise. Let

\[
p = (p_{11}, \ldots, p_{1r}, p_{21}, \ldots, p_{2r}, \ldots, p_{rr})',
\]

and

\[
\beta = (\mu^*, \beta_1, \beta_2, \beta_{12})',
\]

where "'" denotes the transposed,

\[
\beta_1 = (\alpha_1^*, \ldots, \alpha_{r-1}^*), \quad \beta_2 = (\beta_1^*, \ldots, \beta_{r-1}^*), \quad \text{and}
\]

\[
\beta_{12} = (\theta^*, \psi_{11}^*, \ldots, \psi_{rr}^*).
\]

The QU model is then expressed as

\[
\log p = X \beta = (1_r, X_1, X_2, X_{12}) \beta,
\]

where \( X \) is the \( r^2 \times 3r \) matrix and \( 1_s \) is the \( s \times 1 \) vector of 1 elements,

\[
X_1 = [I_{r-1} \otimes 1_r] ; \text{the } r^2 \times (r-1) \text{ matrix},
\]

\[
X_2 = 1_r \otimes [I_{r-1} \otimes 0_{r-1}] ; \text{the } r^2 \times (r-1) \text{ matrix},
\]

and \( X_{12} \) is the \( r^2 \times (r+1) \) matrix, determined from (2), \( I_{r-1} \) is the \( (r-1) \times (r-1) \) identity matrix, \( O_{st} \) is the \( s \times t \) zero matrix, \( 0_s \) is the \( s \times 1 \) zero vector, and \( \otimes \) denotes the Kronecker product. Note that the model matrix \( X \) is full column rank, which is \( 3r \). In a manner similar to Haber (1985) and Lang and Agresti (1994), the linear space spanned by the columns of the
matrix $X$ is denoted by $S(X)$ with the dimension $3r$. Let $U$ be an $r^2 \times d_i$ full column rank matrix, where $d_i = r^2 - 3r = r(r-3)$, such that the linear space spanned by the columns of $U$, that is $S(U)$, is the orthogonal complement of the space $S(X)$. Thus, $U^TX = O_{d_i,3r}$. Therefore the QU model is expressed as

$$h_1(p) = 0_{d_1},$$

where

$$h_i(p) = U'^i \log p.$$  

The MH model may be expressed as

$$h_2(p) = 0_{d_2},$$

where

$$d_2 = r-1, \quad h_2(p) = Wp,$$

and $W$ is the $d_2 \times r^2$ matrix with

$$W = [I_{r-1} \otimes I'_r, O_{r-1,r}] - I'_r \otimes [I_{r-1}, O_{r-1}].$$

Therefore, $W' = X_1 - X_2$ and thus the column vectors of $W'$ belong to the space $S(X)$, that is, $S(W') \subset S(X)$, hence, $WU = O_{d_i,d_i}$. From Theorem 2, the SQU model may be expressed as

$$h_3(p) = 0_{d_3},$$

where

$$d_3 = d_1 + d_2 = r^2 - 2r - 1,$$

and

$$h_3(p) = (h'_1, h'_2)' .$$

Note that $h_s(p), \ s = 1, 2, 3$ are the vectors of order $d_s \times 1$, and $d_s, \ s = 1, 2, 3$, are the numbers of df for testing goodness-of-fit of the QU, MH and SQU models, respectively.

Let $H_s(p), \ s = 1, 2, 3$ denote the $d_s \times r^2$ matrix of partial derivatives of $h_s(p)$ with respect to $p$, that is,

$$H_s(p) = \partial h_s(p)/\partial p'.$$

Let $\Sigma(p) = \text{diag}(p) - pp'$, where $\text{diag}(p)$ denotes a diagonal matrix with $i$th component of $p$ as $i$th diagonal component. It is observed that

$$H_1(p)p = U_1, \quad h_1(p) = 0_{d_1},$$

$$H_1(p)\text{diag}(p) = U', \quad H_2(p) = W.$$ 

Therefore,

$$H_1(p)\Sigma(p)H_2(p)' = U'W' = O_{d_i,d_2},$$

and $\Delta_3 = \Delta_1 + \Delta_2$, is obtained where

$$\Delta_s = h_s(p)' [H_s(p)\Sigma(p)H_s(p)']^{-1}h_s(p).$$ 

(3)

From the asymptotic equivalence of the Wald statistic and the likelihood ratio statistic (Rao, 1973, Sec. 6.3; Darroch & Silvey, 1963; Aitchison, 1962), and from (3), (1) is obtained, thus the proof is completed.

Corollary 2

The following asymptotic equivalence holds:

$$G^2(SQI) = G^2(QI) + G^2(MH).$$

The number of df for the SQI model equals the sum of numbers of df for the QI and MH models.

**Results**

Example 1

Table 1 contains data from a case-control study investigating a possible relationship between cataracts and the use of head coverings during the summer. Each case reporting to a clinic for cataract care was
matched with a control of the same gender and similar age not having a cataract. The row and column categories refer to the frequency with which the subject used head coverings. The SQU model applied to these data has $G^2(SQU) = 10.95$ with 7 df. Thus, the SQU model fits these data well. Under this model, the maximum likelihood estimate of $\theta$ is $\hat{\theta} = 0.808$. The SQU model indicates the structure of both the S and QU models. Therefore, under the SQU model, the probability that using a head covering for one case in a pair is always or almost always, and for the control in the pair is never, is estimated to equal the probability of using a head covering for a case in the pair is never, and for control in the pair is always or almost always.

For local $2 \times 2$ tables that do not contain a cell on the main diagonal, the odds that using head covering for a case in a pair is $s + 1$ instead of $s$ is estimated to be $\hat{\theta} = 0.808$ times when that for the control in the pair is $t + 1$ than when it is $t$. For $i < j$ and $s < t$ with $i \neq s$, $i \neq t$, $j \neq s$, $j \neq t$, the odds that the using head covering case in a pair is $j$ instead of $i$ is estimated to be $(0.808)^{(j-i)(t-s)}$ times higher when that for the control in the pair is $t$ than when it is $s$. For example, the odds that the using a head covering for a case in a pair is never instead of frequency is estimated to be $0.426 = (0.808)^4$ times higher when that for control in the pair is occasionally than when it is always or almost always.

### Table 1: Case-Control Study Investigating a Possible Relationship between Cataracts and the Use of Head Coverings during the Summer

<table>
<thead>
<tr>
<th>Cataract Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Always or Almost Always (1)</td>
</tr>
<tr>
<td>Always or Almost Always (1)</td>
<td>29 (29.00)*</td>
</tr>
<tr>
<td>Frequently (2)</td>
<td>5 (4.22)</td>
</tr>
<tr>
<td>Occasionally (3)</td>
<td>9 (5.12)</td>
</tr>
<tr>
<td>Never (4)</td>
<td>7 (6.16)</td>
</tr>
</tbody>
</table>

*Note: The parenthesized values are the maximum likelihood estimates of expected frequencies under the SQU model.
Example 2

Table 2 contains data from the Los Angeles study of endometrial cancer. These data were obtained from 59 matched pairs using four dose levels of conjugated oestrogen: (1) none, (2) 0.1-0.299 mg, (3) 0.3-0.625 mg, and (4) 0.625+ mg.

Table 3 shows that the SQU and MH models fit the data poorly while the QU model fits the data well. From Theorem 2, the poor fit of the SQU model may be said to be caused by the influence of the lack of structure of the MH model rather than the QU model. Because the QU model fits the data in Table 2 well, under this model, the cell probabilities $\{p_{ij}\}$ have a uniform local association for cells off the main diagonal of the table.

<table>
<thead>
<tr>
<th>Average Dose for Case (mg/day)</th>
<th>Average Dose for Control (mg/day)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 (1)</td>
<td>0 (1)</td>
</tr>
<tr>
<td></td>
<td>0.1-0.299 (2)</td>
<td>0.1-0.299 (2)</td>
</tr>
<tr>
<td></td>
<td>0.3-0.625 (3)</td>
<td>0.3-0.625 (3)</td>
</tr>
<tr>
<td></td>
<td>0.625+ (4)</td>
<td>0.625+ (4)</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Likelihood Ratio Chi-squared Values $G^2$ for Models Applied to Tables 1 and 2

<table>
<thead>
<tr>
<th>Applied Models</th>
<th>Table1</th>
<th>Table2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Degrees of Freedom</td>
<td>$G^2$</td>
</tr>
<tr>
<td>QI</td>
<td>5</td>
<td>6.99</td>
</tr>
<tr>
<td>QU</td>
<td>4</td>
<td>6.52</td>
</tr>
<tr>
<td>SQI</td>
<td>8</td>
<td>11.56</td>
</tr>
<tr>
<td>SQU</td>
<td>7</td>
<td>10.95</td>
</tr>
<tr>
<td>S</td>
<td>6</td>
<td>8.29</td>
</tr>
<tr>
<td>QS</td>
<td>3</td>
<td>3.85</td>
</tr>
<tr>
<td>MH</td>
<td>3</td>
<td>4.38</td>
</tr>
</tbody>
</table>

* means are significant at the 0.05 level
Conclusion

This article gives the decomposition of the SQU model and shows the orthogonality of test statistics. As observed in the examples, Theorem 2 would be useful for explaining the reason for the poor fit when the SQU model fits the data poorly.

From Theorem 3 it may be noted that the likelihood ratio statistic for testing goodness-of-fit of the SQU model - assuming that the QU model holds true - is $G^2(SQU) - G^2(QU)$ and this is asymptotically equivalent to the likelihood ratio statistic for testing goodness-of-fit of the MH model, that is, $G^2(MH)$. Namely, $G^2(SQU) - G^2(QU)$ would be used for testing goodness-of-fit of the MH model.

Suppose that model $\Omega_3$ holds if and only if both models $\Omega_1$ and $\Omega_2$ hold, where the number of df for $\Omega_3$ equals the sum of numbers of df for $\Omega_1$ and $\Omega_2$. Darroch and Silvey (1963) described that (i) when the asymptotic equivalence, then

$$G^2(\Omega_3) = G^2(\Omega_1) + G^2(\Omega_2) \tag{4}$$

holds, if both $\Omega_1$ and $\Omega_2$ are accepted (at the $\alpha$ significance level) with high probability, then $\Omega_3$ would be accepted; however, (ii) when equation (4) does not hold, it is possible for an incompatible situation to arise where both $\Omega_1$ and $\Omega_2$ are accepted with high probability but $\Omega_3$ is rejected with high probability (Darroch and Silvey (1963) show an interesting example). For the orthogonal decomposition of the SQU model into the QU and MH models, such an incompatible situation would not arise in terms of Theorem 3. Therefore, the orthogonal decomposition of the SQU model obtained herein is useful for analyzing data.

References


A Comparative Study for Bandwidth Selection in Kernel Density Estimation

Omar M. Eidous    Mohammad Abd Alrahem Shafeq Marie    Mohammed H. Baker Al-Haj Ebrahem
Yarmouk University, Irbid, Jordan

Nonparametric kernel density estimation method does not make any assumptions regarding the functional form of curves of interest; hence it allows flexible modeling of data. A crucial problem in kernel density estimation method is how to determine the bandwidth (smoothing) parameter. This article examines the most important bandwidth selection methods, in particular, least squares cross-validation, biased cross-validation, direct plug-in, solve-the-equation rules and contrast methods. Methods are described and expressions are presented. The main practical contribution is a comparative simulation study that aims to isolate the most promising methods. The performance of each method is evaluated on the basis of the mean integrated squared error for small-to-moderate sample size. Simulation results show that the contrast method is the most promising methods based on the simulated families considered.

Key words: Probability Density Function, Bandwidth, Least Squares Cross-Validation, Biased Cross-Validation, Contrast Method, Direct Plug-In, Solve-The-Equation Rules.

Introduction

The Kernel method is widely used in nonparametric density estimation. It produces a kernel estimator for the unknown probability density function \( f(x) \). Many researchers have observed that the choice of the bandwidth (smoothing) parameter, \( h \), is crucial for the effective performance of the kernel estimator (for example, see Scott, 1992). A method that uses the data \( X_1, X_2, ..., X_n \) to produce a value for the bandwidth \( h \) is termed a bandwidth selector or data-driven selector.

Various data-driven methods for selecting the bandwidth have been proposed and studied. Most of these methods are based on minimizing the MISE or AMISE. Unfortunately, none of these are available in practice because all of them depend on the unknown probability density function. (See Bowman, 1984; Stone, 1984; Hall & Marron, 1985; Scott & Terrell, 1987; Sheather & Jones, 1991.)


Let \( X_1, X_2, ..., X_n \) be a random sample of size \( n \) from a continuous univariate distribution with an unknown pdf \( f(x) \), then the kernel density estimator of \( f(x) \), \( x \in \mathbb{R} \) is defined by Silverman (1986) as

\[
\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i). \quad (1.1)
\]

Omar M. Eidous is Associate Professor on the Faculty of Science in the Department of Statistics. Email: omarm@yu.edu.jo. Mohammad Abd Alrahem Shafeq Marie is A graduate Student on the Faculty of Science in the Department of Statistics. Email: mohammadmarie@yahoo.com. Mohammed H. Baker Al-Haj Ebrahem is Associate Professor on the Faculty of Science in the Department of Statistics. Email: malhaj@yu.edu.jo.
where \( K_h(u) = h^{-1}K(uh^{-1}) \). \( K \) is the kernel function and is assumed to be symmetric (Silverman, 1986), and \( h \) is the bandwidth (or the smoothing parameter) that controls the degree of smoothing applied to the data. Both \( K \) and \( h \) are under the control of the user, therefore, their determination is necessary in order to analyze results about the kernel estimator.

The bandwidth can be chosen to minimize the asymptotic mean integrated square error, or AMISE (Silverman, 1986). This case, \( h \) can be obtained by solving the following equation

\[
\min_h \text{AMISE} = \left( \frac{1}{2} h^2 f''(x) k_2 \right)^2 + \frac{f(x)}{nh} \int K^2(t) dt
\]

Taking the derivative of \( \text{AMISE} \) with respect to \( h \) and equating to zero yields,

\[
h = k_2^{-2/5} \left( \int K^2(t) dt \right)^{2/5} \left( \int f''^2(t) dt \right)^{-1/5} n^{-1/5}
\]

Formula (1.3) is disappointing because the optimal bandwidth is a function of the second derivative of the density function being estimated. Therefore, unless the true density is known, it is impossible to know the optimal bandwidth. Moreover, when the true density is known, no estimation problem exists.

Nonetheless, the quantity \( \int (f''(x))^2 dx \) in (1.3) can be estimated by using a kernel estimator.

Methodology

Selecting the Bandwidth

The practical implementation of the kernel density estimator requires specification of the bandwidth \( h \). A widely used criterion is to choose an \( h \) that minimizes the AMISE: the bandwidth controls the smoothness of the fitted density curve. Note that a larger \( h \) provides a smoother estimate with smaller variance and larger bias, while a smaller \( h \) produces a rougher estimate with larger variance and smaller bias.

Most methods for choosing the bandwidth presented in the literature are proposed when the underlying probability density function, \( f(x) \) has support \((-\infty, \infty)\). In addition, by surveying the literature, it was found that the methods presented here are commonly used to estimate the smoothing parameter \( h \) in practice.

Least squares cross-validation (LSCV)

Least squares cross-validation (LSCV), proposed by Rudemo (1982) and Bowman (1984), is a completely automatic method for choosing the bandwidth \( h \). Following Rudemo’s (1982) derivations, the optimal bandwidth estimator can be obtained by minimizing:

\[
\text{LSCV}(h) = \int \hat{f}^2(x; h) dx - 2n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j \neq i} K_h(X_i - X_j)
\]

According to Rudemo (1982), formula (1.4) is derived based on the exact MISE. If the kernel function is Gaussian density, then
\[
\int_{-\infty}^{\infty} \hat{f}^2(x;h)dx = \frac{1}{n^2h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left(\frac{x-X_i}{h}\right)K\left(\frac{x-X_j}{h}\right)dx
\]

\[
= \frac{1}{2n^2h\sqrt{\pi}} \sum_{i=1}^{n} \sum_{j=1}^{n} e^{-\frac{(X_i-X_j)^2}{2h^2}}
\]

and

\[
\sum_{i=1}^{n} \sum_{j \neq i} K_h(X_i-X_j) = \frac{1}{h} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-X_j)^2}{2h^2}}.
\]

Therefore,

\[
\text{LSCV}(h) = \frac{1}{2n^2h\sqrt{\pi}} \sum_{i=1}^{n} \sum_{j=1}^{n} e^{-\frac{(X_i-X_j)^2}{2h^2}}
\]

\[
-2n^{-1}(n-1)^{-1}h^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-X_j)^2}{2h^2}}
\]

(1.5)

The optimal bandwidth \( h \) is obtained by minimizing \( \text{BCV}(h) \) over \( h \).

### Biased Cross-Validation (BCV)

While LSCV method used exact MISE, the biased cross-validation (BCV) is based on the AMISE (Scott & Terrell, 1987). The BCV method suggests the use of the second derivative of the traditional kernel estimator as opposed to the unknown second derivative of \( f(x) \). The BCV objective function is thus given by:

\[
\text{BCV}(h) = \frac{h^4}{4} k_2^2 \mu(\hat{f}''(x;h)) + (nh)^{-1} \mu(K)
\]

(1.6)

where \( \mu(f) = \int_{-\infty}^{\infty} f^2 dt \) and \( \hat{f}''(x;h) \) is the second derivative of the kernel estimator and \( K \) is the Gaussian kernel. Because \( k_2 = 1 \) and \( \mu(K) = \frac{1}{2\sqrt{\pi}} \), \( \text{BCV}(h) \) is given by

\[
\text{BCV}(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{3}{32n^2h^5\sqrt{\pi}} \left( h^4 \sum_{i=1}^{n} \sum_{j=1}^{n} e^{\frac{(X_i-X_j)^2}{4h^2}} - h^2 \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i-X_j)^2 e^{\frac{(X_i-X_j)^2}{4h^2}} + \frac{1}{12} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i-X_j)^4 e^{\frac{(X_i-X_j)^2}{4h^2}} \right)
\]

The optimal value of \( h \) is obtained by minimizing \( \text{BCV}(h) \) over \( h \).

### Direct Plug-In (DPI)

The DPI method is based on the idea of plugging in an estimate of unknown quantity \( \mu(f^{(r)}) \) in equation (1.6):

\[
\mu(f^{(r)}) = \int_{-\infty}^{\infty} (f^{(r)}(x))^2 dx, \ r = 2, 4, 6, 8, \ldots
\]

Sheather and Jones (1991) developed an estimator for \( \mu(f^{(r)}) \) based on the kernel estimator with bandwidth \( g \), which is given by:

\[
\mu(f^{(r)}) = n^{-1} \sum_{i=1}^{n} \hat{f}^{(r)}(X_i; g)
\]

\[
= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_g^{(r)}(X_i-X_j).
\]

(1.7)

According to Wand and Jones (1995), the bias term of the estimator (1.7) can be made to vanish by choosing \( g \) to be equal

\[
g = \left[ \frac{-2K^{(r)}(0)}{\mu(f^{(r+2)})k_2 n} \right]^{1/(r+1)}
\]

(1.8)
The problem is persistent because it is apparent from (1.8) that the optimal bandwidth \( g \) for estimating \( \mu(f^{(r)}) \) depends on \( \mu(f^{(r+2)}) \). To overcome this problem Sheather and Jones (1991) suggested estimating \( \mu(f^{(r)}) \) at some stage and using a simple estimate of bandwidth \( g \) chosen with reference to a parametric family, usually a normal density.

Thus, a family of DPI bandwidth selectors exist which depends on the number of stages of functional estimation before a normal scale (NS) is used. Such a rule will be called an \( l \)-stage DPI bandwidth selector and is denoted by \( \hat{h}_{DPI,l} \). The NS may be considered to be a zero-stage DPI bandwidth selector. Wand and Jones (1995) pointed out that no method exists for objective choice of the number of iterations that should be used. If \( f \) is a normal density with mean 0 and variance \( \sigma^2 \), then according to Wand and Jones (1995), \( r \) will be

\[
\mu(f^{(r)}) = \frac{(-1)^{r/2} \Gamma(r/2)}{(2\sigma)^{r+1}(r/2)!}\sqrt{\pi}.
\]

Note that simulation results presented for the DPI method in the simulation are based on the use of a two-stage DPI bandwidth selector to find the bandwidth. An algorithm for the two-stage DPI method is given by Sheather and Jones (1991).

Solve-the-Equation (STE)

The solve-the-equation (STE) rule is based on the formula for the AMISE-optimal bandwidth. Many authors (Scott, et al., 1977; Sheather, 1986; Park & Marron, 1990; Sheather & Jones, 1991) have required that \( h \) be chosen to satisfy the relationship:

\[
h = \left[ \frac{\mu(K)}{k_j^2\psi_4(\gamma(h))n} \right]^{1/5},
\]

where the pilot bandwidth for the estimation of \( \psi_4 \) is a function \( \gamma \) of \( h \). The choice of \( \gamma \) may be denoted by:

\[
\gamma(h) = \left[ \frac{2K^{(4)}(0)k_j}{R(K)} \right]^{1/7} (-\hat{\psi}_4(g_1)/\hat{\psi}_6(g_2))^{1/7} h^{5/7}
\]

where \( \hat{\psi}_4(g_1) \) and \( \hat{\psi}_6(g_2) \) are kernel estimates of \( \psi_4 \) and \( \psi_6 \), respectively (Sheather & Jones, 1991). The choice of \( g_1 \) and \( g_2 \) may be determined by using:

\[
g_1 = \left[ \frac{-2K^{(4)}(0)}{\psi_6 k_2 n} \right]^{1/7}
\]

and

\[
g_2 = \left[ \frac{-2K^{(6)}(0)}{\psi_6 k_2 n} \right]^{1/9}
\]

where:

\[
\psi_6 = \frac{105}{32\sqrt{\pi} \hat{\sigma}^9},
\]

\[
\psi_4 = \frac{-15}{16\sqrt{\pi} \hat{\sigma}^7},
\]

\[
K^{(4)}(0) = 3/\sqrt{2\pi},
\]

and

\[
K^{(6)}(0) = -15/\sqrt{2\pi}.
\]

Note that this two-stage STE bandwidth selector was used to find the bandwidth in the simulation and the algorithm used to find the \( \hat{h}_{STE,2} \) was based on Sheather and Jones (1991).

Contrast Method (CONT)

Ahmad and Ran (1998) introduced the concept of kernel contrast to select the bandwidth \( h \) by studying its finite sample and asymptotic properties. The first step in the CONT method is to define the kernel density estimations \( \hat{f}_j(x; h) \) based on \( q \) kernels, \( K_1, K_2, \ldots, K_q, \ q \geq 2 \). After selecting the contrast coefficients \( p_1, p_2, \ldots, p_q \), where

\[
\sum_{j=1}^q p_j = 0,
\]

the bandwidth that minimizes the \( \text{MISE}(h)_{\text{CONT}} \) is selected. However, a reasonable choice for estimating \( h \) is to minimize
EIDOUS, SHAFEQ MARIE & AL-HAJ EBRAHEM

ISE(h)_{CONT}, which does not depend on the unknown density function f(t). This method was proposed by Ahmad and Ran (2004), where

\[ MISE(h)_{CONT} = E \left[ \int \left( \sum_{j=1}^{q} p_j \hat{f}_j(x; h) \right)^2 dx \right] \]

and

\[ ISE(h)_{CONT} = \int \left( \sum_{j=1}^{q} p_j \hat{f}_j(x; h) \right)^2 dx. \]

Ahmad and Mugdadi (2003) showed that the estimator based on the ISE(h)_{CONT} for f(x) is consistent. The density estimation using a kernel contrast is denoted by

\[ \hat{f}(x; h) = \sum_{j=1}^{q} c_j \hat{f}_j(x; h). \]

The kernels may have an equal weight if q is chosen as an even integer, where

\[ \sum_{j=1}^{q} c_j = 1; \quad c_j = 1/q \quad \text{for } j=1, ..., q \]

and

\[ \sum_{j=1}^{q} p_j = 0; \quad p_j = -p_{2j} \quad \text{for } j=1, ..., q/2. \]

The simulation results in this article were found by taking, \( p_1 = -p_2, \quad p_2 = -1, \quad c_1 = c_2 = 1/2, \)

where \( K_1, K_2 \) are the two kernels N(0,1) and N(0,4), respectively. Therefore,

\[ ISE(h)_{CONT} = \frac{1}{2} \int_{-\infty}^{\infty} \left( \sum_{i=1}^{n} \frac{1}{nh} K_1 \left( \frac{x-X_i}{h} \right) - \sum_{j=1}^{n} \frac{1}{nh} K_2 \left( \frac{x-X_j}{h} \right) \right)^2 dx \]

Thus,

\[ ISE(h)_{CONT} = \left\{ \frac{1}{2n^2 h^2} \int_{-\infty}^{\infty} \left( \sum_{i=1}^{n} K_1 \left( \frac{x-X_i}{h} \right) \right)^2 dx - \frac{1}{n^2 h^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} K_1 \left( \frac{x-X_i}{h} \right) K_2 \left( \frac{x-X_j}{h} \right) dx \right\} \]

\[ + \left\{ \frac{1}{2n^2 h^2} \int_{-\infty}^{\infty} \sum_{j=1}^{n} K_2 \left( \frac{x-X_j}{h} \right)^2 dx \right\} \]

where:

\[ K_1 \left( \frac{x-X_i}{h} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-X_i)^2}{2h^2}} \]

and

\[ K_2 \left( \frac{x-X_j}{h} \right) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-X_j)^2}{8h^2}}. \]

Therefore,

\[ ISE(h)_{CONT} = \frac{3}{8h\sqrt{\pi}} - \frac{1}{n^2 h^2} \sum_{i=1}^{n} \sum_{j=1}^{n} e^{-\frac{(X_i-X_j)^2}{10h^2}} \]

Simulation Study

A simulation study was conducted to compare the several methods discussed for selecting the bandwidth of a kernel density estimator. The methods compared to estimate the bandwidth h - and consequently f(x) - are: least squares cross-validation (LSCV), biased cross-validation (BCV), direct plug-in (DPI), solve-the-equation (STE) rules and contrast (CONT). It is important to understand the effects
of the different methods for the estimator of \( f(x) \) for different values of the sample size, \( n \). In this study, four different normal mixture densities were simulated; these densities are (Marron & Wand, 1992):

a. Gaussian:
\[
f_1(x) = \phi(x).
\]

b. Kurtotic Unimodal:
\[
f_2(x) = \frac{2}{3} \phi(x) + \frac{1}{3} \phi_{1/10}(x)
\]
c. Bimodal:
\[
f_3(x) = \frac{1}{2} \phi_{2/3}(x + 1) + \frac{1}{2} \phi_{2/3}(x - 1)
\]
d. Strongly Skewed:
\[
f_4(x) = \sum_{l=1}^{8} \frac{1}{8^{(2/3)^{l}}} \{ x - 3[(2/3)^{l} - 1] \}
\]

(1.9)

where \( \phi_{\lambda}(u) = A^{-1} \phi(u/A) \) and \( \phi \) denotes the probability density function (pdf) of a standard normal variable, that is,
\[
\phi_A(u) = \frac{1}{\sqrt{2\pi A}} e^{-\frac{1}{2}(u/A)^2}.
\]

These densities represent Symmetric, Kurtotic Unimodal, Bimodal and Strongly Skewed distributions respectively. Figure 1 displays the shapes of these densities, which are a small subset of fifteen normal mixtures used by Marron and Wand (1992).

The general normal mixtures density is given by (Marron & Wand, 1992):
\[
f(x) = \sum_{l=1}^{k} w_l \phi_{\sigma_l}(x - \mu_l)
\]

where \(-\infty < \mu_l < \infty, \sigma_l > 0\) and \( w_l \) is a vector with positive entries summing to unity (weight), for \( l = 1, 2, ..., k \). It is assumed that \( f \) has a normal \( k \)-mixture density with parameters \( \{(w_l, \mu_l, \sigma_l^2) : l = 1, 2, ..., k \} \).

Fryer (1976) and Deheuvels (1977) first showed that the MISE could be calculated exactly when both the underlying density and the kernel function are Gaussian. The integrated squared error (ISE) of the estimator - if the true underlying density is known to be \( f(x) \) as in equation (1.37) - is given by Marron and Wand (1992) as

\[
\text{ISE}(\hat{f}) = \int_{-\infty}^{\infty} \left[ \hat{f}(x; h) - f(x) \right]^2 dx
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i=1}^{n} \phi_{\lambda}(X_{ij} - X_{il})
\]

\[
- \frac{2}{n^2} \sum_{i=1}^{n} \sum_{i=1}^{n} w_l \phi_{\lambda}(h^2 + \sigma^2_l)^{1/2} (X_i - \mu_l)
\]

\[+ U(h, 0) \]

where

\[
U(h, q) = \sum_{i=1}^{k} \sum_{i=1}^{k} w_l w_{l'} \phi_{\lambda}(q h^2 + \sigma^2_l + \sigma^2_{l'})^{1/2} (\mu_l - \mu_{l'})
\]

and the kernel function \( K \) is the standard normal. Thus, it is more appropriate to analyze the expected value of the ISE, called the MISE.

For each normal mixture density in (1.9) and each sample size \( n = 50, 100, 200, 500 \) that were simulated from \( f(x) \), 1,000 samples were artificially repeated from each \( f(x) \). For each sample, the bandwidth \( h \) based on LSCV, BCV, DPI, CONT and STE methods were obtained. Subsequently, for each sample the ISE values were obtained by using (1.9) according to the simulated density \( f(x) \). Subsequently, the MISE values were empirically determined as the mean of the ISE values obtained in each sample. Table 1 displays the simulation results and the MISEs against the sample sizes for the different underlying normal mixture densities. Moreover, the relative efficiencies of the contrast (CONT) method against LSCV, BCV, DPI and STE methods are given in Table 2. The rule of relative efficiency is given by

\[
RE(\hat{h}) = \frac{\text{MISE}(\hat{h})}{\text{MISE}(\hat{h}_{\text{CONT}})}.
\]
where $\hat{h}_*$ is the bandwidth which computed from the other methods (see Table 2).

Conclusion
Tables 1 and 2 show the main results of the simulation study. To provide insight into the effect of the sample size and different normal mixture densities on the performance of the various bandwidth selection methods, the following conclusions can be drawn:

1. The MISE for the kernel estimator $\hat{f}(x; h)$ decreases as the sample size increases for all simulated functions and for all different methods, which coincides with the theoretical properties of the kernel estimator.

2. In terms of the MISE of $\hat{f}(x; h)$, the performance of the BCV method is acceptable when the data are simulated from a very skewed density ($f_4(x)$), while its performance is inefficient for the other densities.

3. The MISE values of $\hat{f}(x; h)$ when $h$ is estimated based on the LSCV or BCV method are large compared with the MISE values produced by the other methods for all simulated densities and for all sample sizes.

Note that conclusions 2 and 3 suggest that these two methods should be disregarded as global method to select the bandwidth $h$. 
Table 1: The MISE ($\hat{f}$) for Different Methods to Choose the Value of Bandwidth

<table>
<thead>
<tr>
<th>Method</th>
<th>MISE ($\hat{f}$)</th>
<th>$f_1(x)$</th>
<th>$f_2(x)$</th>
<th>$f_3(x)$</th>
<th>$f_4(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sample Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPI(2-stage)</td>
<td>50</td>
<td>0.12846</td>
<td>0.23448</td>
<td>0.15910</td>
<td>0.76199</td>
</tr>
<tr>
<td>CONT</td>
<td></td>
<td>0.12481</td>
<td>0.22572</td>
<td>0.10643</td>
<td>0.58228</td>
</tr>
<tr>
<td>LSCV</td>
<td></td>
<td>0.19144</td>
<td>0.28647</td>
<td>0.25236</td>
<td>0.94230</td>
</tr>
<tr>
<td>BCV</td>
<td></td>
<td>0.40578</td>
<td>0.43591</td>
<td>0.39941</td>
<td>0.76748</td>
</tr>
<tr>
<td>STE(2-stage)</td>
<td></td>
<td>0.13070</td>
<td>0.23831</td>
<td>0.17873</td>
<td>0.78334</td>
</tr>
<tr>
<td>DPI(2-stage)</td>
<td>100</td>
<td>0.12730</td>
<td>0.21665</td>
<td>0.15063</td>
<td>0.75133</td>
</tr>
<tr>
<td>CONT</td>
<td></td>
<td>0.12373</td>
<td>0.22057</td>
<td>0.09582</td>
<td>0.56732</td>
</tr>
<tr>
<td>LSCV</td>
<td></td>
<td>0.16841</td>
<td>0.26467</td>
<td>0.23532</td>
<td>0.88068</td>
</tr>
<tr>
<td>BCV</td>
<td></td>
<td>0.31693</td>
<td>0.38360</td>
<td>0.30008</td>
<td>0.71768</td>
</tr>
<tr>
<td>STE(2-stage)</td>
<td></td>
<td>0.12530</td>
<td>0.23352</td>
<td>0.12739</td>
<td>0.76518</td>
</tr>
<tr>
<td>DPI(2-stage)</td>
<td>200</td>
<td>0.12215</td>
<td>0.20491</td>
<td>0.14057</td>
<td>0.74043</td>
</tr>
<tr>
<td>CONT</td>
<td></td>
<td>0.11160</td>
<td>0.21296</td>
<td>0.09271</td>
<td>0.55792</td>
</tr>
<tr>
<td>LSCV</td>
<td></td>
<td>0.15314</td>
<td>0.25512</td>
<td>0.21145</td>
<td>0.83185</td>
</tr>
<tr>
<td>BCV</td>
<td></td>
<td>0.25271</td>
<td>0.30184</td>
<td>0.23965</td>
<td>0.60122</td>
</tr>
<tr>
<td>STE(2-stage)</td>
<td></td>
<td>0.11947</td>
<td>0.19695</td>
<td>0.12676</td>
<td>0.73857</td>
</tr>
<tr>
<td>DPI(2-stage)</td>
<td>500</td>
<td>0.11903</td>
<td>0.19237</td>
<td>0.13929</td>
<td>0.73197</td>
</tr>
<tr>
<td>CONT</td>
<td></td>
<td>0.11208</td>
<td>0.20337</td>
<td>0.09019</td>
<td>0.55088</td>
</tr>
<tr>
<td>LSCV</td>
<td></td>
<td>0.13948</td>
<td>0.24905</td>
<td>0.19998</td>
<td>0.78332</td>
</tr>
<tr>
<td>BCV</td>
<td></td>
<td>0.20559</td>
<td>0.28995</td>
<td>0.16698</td>
<td>0.56810</td>
</tr>
<tr>
<td>STE(2-stage)</td>
<td></td>
<td>0.10785</td>
<td>0.18467</td>
<td>0.12599</td>
<td>0.71857</td>
</tr>
</tbody>
</table>
4. The DPI and STE methods produce similar results in term of their MISE values for all densities and for all sample sizes. The DPI method performs better than the STE method for small sample sizes and as the sample size increases the STE is better than the DPI method. This indicates that the convergence rate of the STE method is faster than that of the DPI method.

5. The performance of the CONT method generally is better than the performance of the other methods. A significant improvement for the CONT method over the other methods is clearly demonstrated in the bimodal \( f_3(x) \) and the strongly skewed \( f_4(x) \) models.

6. The relative efficiency values in Table 2 show that, for most of the densities and sample sizes, a considerable gain in the relative efficiency for the CONT method is achieved. The relative efficiency values are less than one in some cases, which indicates that the performance of the corresponding method is better than CONT method, but the relative efficiency remains acceptable in these cases.

7. Comparing the MISE values for different methods when the data are simulated from \( f_4(x) \) to the MISE values when the data are simulated from the other densities, it may be concluded that \( f_4(x) \) is difficult to estimate by any of the methods considered. That is, the strongly skewed density contains features that cannot be recovered from the sample sizes considered.

8. On the basis of the simulation results, the CONT method may be recommended as a global method to select the bandwidth \( h \) in kernel density estimation.
This study has shown that the CONT method is a useful technique for choosing the bandwidth of the kernel estimator. The CONT method produces reasonable estimates for \( f(x) \) in almost all cases considered (see Table 2). Although the conclusions are based on four different densities, many other candidate shapes exist for the densities from which it is assumed that the data was obtained (Marron & Wand, 1992). Therefore, it is not possible to claim that the CONT method performs better than the other methods for any set of data. However, based on the simulation study, the different methods can be ranked in ascending order (best to worst) according to their performances as follows:

1. CONT.
2. DPI (2-stage) and STE (2-stage)
3. LSCV
   and lastly,
4. BCV

References


Applying Multiple Imputation with Geostatistical Models to Account for Item Nonresponse in Environmental Data

Breda Munoz Virginia M. Lesser Ruben A. Smith
RTI International, RTP, NC
Oregon State University, Corvallis, OR

Methods proposed to solve the missing data problem in estimation procedures should consider the type of missing data, the missing data mechanism, the sampling design and the availability of auxiliary variables correlated with the process of interest. This article explores the use of geostatistical models with multiple imputation to deal with missing data in environmental surveys. The method is applied to the analysis of data generated from a probability survey to estimate Coho salmon abundance in streams located in western Oregon watersheds.

Key words: Environmental surveys; missing data; nonresponse.

Introduction
Environmental surveys are often subject to missing data. An entire observational unit, such as a sampling site, may be missing; conversely, one or a few variables for an observational unit may be missing. These types of missing data are referred to in the survey literature as either unit or item nonresponse, respectively (Lessler & Kalsbeek, 1992). Causes for missing data in environmental studies include failure of the measuring instruments (resulting in unit and/or item nonresponse), inaccessibility of the site (unit nonresponse), and data lost or damaged (unit and/or item nonresponse). A multiple imputation approach is proposed for handling missing item nonresponse data that occurs at one sample point in time data in environmental surveys.

Further study of the magnitude and factors resulting in missing data is necessary to interpret the data that has been collected. The impact of missing data in the estimation stage depends on the missing data mechanism or random process leading to it and also on whether the observed missingness is related to any variables in the dataset (Little & Rubin, 2002). Specifically, the impact of nonresponse on survey error depends on how the missing data occurred, the percent of nonresponse, and the parameters to be estimated (Lessler & Kalsbeek, 1992; Little & Rubin, 2002).

Let $Y = \begin{bmatrix} Y_{\text{obs}} \\ Y_{\text{miss}} \end{bmatrix}$ denote the matrix of complete data corresponding to observations of a random process, where $Y_{\text{miss}}$ and $Y_{\text{obs}}$ denote the missing and observed components of $Y$, respectively. Missing data can be classified as missing completely at random (MCAR), missing at random (MAR), and nonignorable or informative nonresponse (Little & Rubin, 2002). Data is called MCAR if the observed data ($Y_{\text{obs}}$) can be considered a representative sample of the population, that is, the missingness does not depend on the response ($Y$) or other variables.

Breda Munoz is a Senior Research Statistician at RTI International. Email her at: breda@rti.org. Virginia M. Lesser is a Professor of Statistics and Director of the Survey Research Center, Oregon State University. Email her at: lesser@science.oregonstate.edu. Ruben A. Smith currently serves as a Mathematical Statistician for the Applied Sciences Branch, Division of Reproductive Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention. Email him at: RASmith@cdc.gov.
measured at the site or regional level. Under this assumption, valid results are obtained when analysis techniques developed for complete data sets are performed on the observed data ($Y_{obs}$) (Little & Rubin, 2002; Lessler & Kalsbeek, 1992; Lohr 2001).

When the missingness does not depend on the unobserved response but depends only on observed values of auxiliary variables, then the missing data mechanism is known as MAR. This is also referred to as ignorable nonresponse. A model for this nonresponse mechanism can be formulated and incorporated into either design-based or model-based analysis techniques to explain and account for the nonresponse. For example, among the design based approaches, weighting methods - such as a weighting class adjustment - can be used to produce estimates to adjust for the nonresponse (Lohr, 2001).

Finally, if the probability of nonresponse depends on the response and cannot be completely explained by the values of the auxiliary variables, then the nonresponse is nonignorable (Little & Rubin, 2002). Models for the nonignorable missing mechanism are usually more complicated than models for ignorable nonresponse because they depend on the unobserved values.

Recognized approaches to handle missing data problems include deletion of the records, hot or cold deck imputation (Chen & Shao, 1999), substitution, parametric and semi-parametric modeling techniques (Rotnitzky, et al., 1998; Robins, 1995), and multiple imputation (Little & Rubin, 2002). More innovative techniques include neural networks (Gupta & Lam, 1996), Bayesian models (Sebastiani & Ramoni, 2000; Kleinman, et al., 1998), maximum likelihood estimation approaches (Little & Schluchter, 1985; Schneider, 2001; Little 1982), and linear and generalized linear model imputation assuming nonignorable missing data (Greenless, et al., 1982; Baker & Laird, 1988; Ibrahim, 1990).

Most of these approaches result in a single imputation of the missing data, generating one complete data set. Analyses are then applied to the complete data set. The results of data analysis on single imputation data neither reflect the missing-data uncertainty nor on the consequence of imputation. Furthermore, analyses based on a single imputation may result in under-estimated standard errors, incorrect p-values, and high Type I error rates. This problem increases as the rate of missing information and the number of model parameters increases (Schafer & Olsen, 1998).

Another method to deal with nonresponse is the well-known multiple imputation (MI) methodology. This method incorporates the uncertainty of the missing data into the inference (Rubin, 1987). MI replaces each missing item with $m$ values from a distribution of likely values. This process generates $m$ complete data sets on which the same analysis procedure is performed. The final inferences combine the individual estimates obtained from the $m$ complete data sets, thus allowing a researcher to account for the variability due to imputation and to analyze the data using standard techniques and software available for complete datasets (Schafer & Olsen, 1998; Schafer, 1997).

To account for the spatial variability inherent in environmental monitoring programs, a geostatistical model is considered as the imputation model. Kriging and other stochastic predictors for spatial data are referred to as geostatistical models in the spatial statistics literature (Diggle, et al., 1998). Kriging is a well-known technique for spatial interpolation that generates predictions for the unobserved values of the spatial random process at the unvisited sites. The kriging estimator is a minimum error weighted linear predictor that assumes a Gaussian distribution for the random process and a model for the variance-covariance matrix (see Cressie, 1993 for more details). Diggle, et al. (1998) extended the concept of geostatistical models to non-Gaussian situations within the framework of generalized linear models (see McCullagh & Nelder, 1989 for more details on generalized linear models).

In this study MI is explored using geostatistical models for handling missing data in environmental surveys for item nonresponse. An advantage of using geostatistical models in MI is the possibility of imputing missing values for both continuous and discrete environmental variables.
**Multiple Imputation**

Multiple imputation (MI) is a simulation-based approach analyzing missing data that incorporates the uncertainty of missing data into the inference (Rubin, 1987; Rubin, 2002, Harrel & Zhou, 2007). In MI, each missing datum is replaced by a set of $m > 1$ simulated plausible values from their predictive distribution creating $m$ complete data sets. Each complete data set is analyzed separately. The final estimator is the average of the estimators obtained in the individual analyses. The variability introduced by the $m$ analyses is combined with an estimate of the sample variance to provide a single variability measure for the parameters of interest (Schafer, 1997).

Following Rubin (1996) and Schafer (1997), $\hat{Q}_i$ is denoted as a point estimate (e.g., an estimate of salmon abundance in the State of Oregon) of the parameter of interest, $Q$ (e.g., salmon abundance in the State of Oregon), where $i = 1, \ldots, m$. Let $\hat{U}_i$ denote the estimated variance of $\hat{Q}_i$ obtained from the $i^{th}$ individual analysis, $i = 1, \ldots, m$. The overall point estimate is obtained as

$$\bar{Q}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$$

and the overall within imputation variance estimate is given by

$$\bar{U}_m = \frac{1}{m} \sum_{i=1}^{m} \hat{U}_i.$$ 

The between imputation variance estimate, defined as

$$B_m = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q}_m)^2,$$

reflects the extra inferential uncertainty due to the imputation of the missing data. The total variance of $\bar{Q}_m$, is calculated as

$$T_m = \bar{U}_m + (1 + m^{-1})B_m.$$ 

A confidence interval for the parameter of interest, $Q$, can be obtained as: $\bar{Q}_m \pm t_{df} \sqrt{T_m}$, where $t_{df}$ is the $df$-quantile of the $t$-Student distribution, and

$$df = (m-1) \left(1 + \frac{m\bar{U}_m}{(m+1)B_m}\right)^2$$

denotes the corresponding degrees of freedom (Barnard & Rubin, 1999).

To ensure valid inferences when using MI, researchers must assume a mechanism of missingness, a model for the complete data $f(Y_{\text{miss}}, Y_{\text{obs}})$, and a prior distribution for the parameters of the model. A MAR mechanism for the missing data was assumed and imputations for $Y_{\text{miss}}(s)$ from the posterior predictive distribution of the missing data $f(Y_{\text{miss}} \mid Y_{\text{obs}})$ were generated. The posterior predictive distribution of $Y_{\text{miss}}$ can be obtained by Bayes’s Theorem as

$$f(Y_{\text{miss}} \mid Y_{\text{obs}}) = \int f(Y_{\text{miss}} \mid \Theta, Y_{\text{obs}}) f(\Theta \mid Y_{\text{obs}}) d\Theta$$

(1)

where $\Theta$ represents the vector of parameters of the imputation model for the complete data (e.g., $f(Y_{\text{miss}}, Y_{\text{obs}})$), $f(Y_{\text{miss}} \mid Y_{\text{obs}})$ is the posterior predictive distribution of $Y_{\text{miss}}$ given $\Theta$ and the observed data (e.g., $Y_{\text{obs}}$), $f(\Theta \mid Y_{\text{obs}})$ is the posterior distribution of $\Theta$ given the observed data (e.g., $Y_{\text{obs}}$), and $\Theta$ denotes the parameter space (Schafer, 1997; Little & Rubin, 2002). It can be shown that $f(\Theta \mid Y_{\text{obs}}) \propto L(\Theta \mid Y_{\text{obs}}) \pi(\Theta)$, where $L(\Theta \mid Y_{\text{obs}})$ is the observed data likelihood, and $\pi(\Theta)$ is an assumed prior for $\Theta$.

The resulting posterior predictive density of $Y_{\text{miss}}(s)$, $f(Y_{\text{miss}} \mid Y_{\text{obs}})$, may not be a recognizable distribution. Whether the
distribution is recognizable depends on the assumptions adopted for the conditional distributions and the priors. In some cases \( f(Y_{miss} \mid Y_{obs}) \) can be written as the product of conditional and marginal known densities.

In other cases, only an approximation can be obtained by means of computational analyses such as the Markov Chain Monte Carlo (MCMC) methods, which consist of a collection of techniques for drawing pseudo random values from approximate or exact predictive distributions (Schafer, 1997; Gelman, et al., 1995). These methods include the Gibbs sampling algorithm, data augmentation methods, the Metropolis-Hasting algorithm and a series of hybrid algorithms.

MCMC is one of the primary methods for generating MI’s in nontrivial problems. MCMC is discussed in the literature for parameter simulation by creating a dependent sequence of random draws of parameters from Bayesian posterior distributions under complicated parametric models (Gilks, et al., 1996). However, in MI-related applications MCMC is used to create a small number of independent draws of the missing data from a predictive distribution; these draws are then used for multiple-imputation inference (Schaffer, 1997; Rubin, 2003).

The MCMC methods generate sequential realizations of the posterior predictive density of \( Y_{miss}(s) \), \( \{ Y^{(t)}_{miss}(s) : t = 1, 2, \ldots \} \). Each term in the sequence (e.g., \( Y^{(t)}_{miss}(s) \)) depends on the preceding one, and the limiting distribution of the sequence converges to the posterior predictive density of \( Y_{miss}(s) \). These methods are attractive because the convergence of the MCMC algorithms does not require that the starting values for the distribution of \( Y_{miss}(s) \) to be actual realizations of the posterior predictive density of \( Y_{miss}(s) \). Close starting values are recommended, however, to assure faster convergence (Gelman & Rubin, 1992; Shafer, 1997). Finally, the posterior predictive mean is defined as the expected value of the posterior predictive distribution of \( Y_{miss} \), \( E(Y_{miss} \mid Y_{obs}, \theta) \). Diagnostic assessment of the convergence of the MCMC chains can be made using the convergence diagnostics of Geweke (1992) and Heidelberger and Welch (1983). Both convergence diagnostics assess the stationary convergence assumption of the chain.

Geostatistical Models

In environmental science, researchers use geostatistical techniques to model environmental processes that evolve in space and time. Geostatistical models are proposed (Handcock & Stein, 1993; Le & Zidek, 1992; Diggle, et al., 1998; Diggle & Ribeiro, 2002; Christensen & Waagepetersen, 2002) in conjunction with MI (Schafer, 1997; Rubin, 1996; Little & Rubin, 2002) to handle missing data in environmental surveys.

An environmental process of interest is generated by an unobserved spatial random field, \( Y \), defined over a continuous region of interest, \( D \subset \mathbb{R}^2 \). \( Y(s) \) denotes the outcome of the process of interest at location \( s \), and \( s \) be the coordinates of a site or point in \( D \), \( s \in D \). The observed data is collected from a finite number of sites, \( S = \{ s_1, s_2, \ldots, s_n \} \). The sites can be selected either from a probability or a non-probability sampling design. Missing data occurs in \( n_1 \) of the \( n \) sites, with \( n_1 < n \).

For each point \( s \) in \( D \), the random process of interest, \( Y \), has a distribution with mean \( \mu(s) \), \( E[Y(s)] = \mu(s) \). A continuous differentiable function \( g \) of \( \mu \) exists, such that \( g[\mu(s)] = X\beta + Z(s) + \varepsilon(s) \), where \( X \) is a vector of covariates, correlated with the random process \( Y \), that is available at the site level, and \( \beta \) is a vector of unknown parameters. \( Z \) denotes a spatial random effect with mean 0 and its variance-covariance matrix \( \sigma_Z^2 \mathbf{R}(\theta) \). \( \mathbf{R}(\theta) \) is a correlation matrix. This correlation matrix is a function of the distance between two sites and \( \theta \), where \( \theta \) is a vector of unknown correlation parameters and \( \sigma_Z^2 \) is the unknown structural parameter or constant variance. In addition, \( \varepsilon \) denotes an independent non-spatial random effect with mean 0 and variance-covariance matrix \( \sigma^2 \mathbf{I} \). In this case, \( \sigma^2 \) represents the classical nugget effect and captures
measurement error or a combined effect of measurement error and any small scale spatial variation (Diggle & Ribeiro, 2002).

The posterior predictive density $Y_{\text{miss}}(s)$ is obtained by integrating the following expression with respect to the parameters $\beta, \theta, \sigma^2_\varepsilon$ and $\sigma^2_Z$ (see Equation 1) is:

$$f(Y_{\text{miss}} | Y_{\text{obs}}, \beta, \theta, \sigma^2_\varepsilon, \sigma^2_Z, Z) \propto f(Y_{\text{miss}} | Y_{\text{obs}}, \beta, \theta, \sigma^2_\varepsilon, \sigma^2_Z, Z)f(Z | \theta, \sigma^2_s) \times f(\theta | Y_{\text{obs}})f(\sigma^2_\varepsilon | Y_{\text{obs}}) \pi(\beta)\pi(\theta)\pi(\sigma^2_\varepsilon)\pi(\sigma^2_Z)$$

An exact expression for the integral will depend on the distribution (such as normal, Poisson, gamma, Bernoulli, binomial) assumed for the complete data, $f(Y_{\text{miss}}, Y_{\text{obs}})$, the distributions assumed for the two random components of the model, $f(Z | \theta, \sigma^2_s)$ and $f(\varepsilon | \sigma^2_\varepsilon)$, and the priors assumed for the parameters, $\pi(\beta), \pi(\theta), \pi(\sigma^2_\varepsilon)$ and $\pi(\sigma^2_Z)$.

Diggle and Ribeiro (2002), Handcock and Stein (1993) and Omre and Halvorsen (1989) investigated the case assuming a Gaussian distribution for the data and a number of prior distributions for the parameters; their results are applied when selecting appropriate priors for the simulation and illustrative examples herein.

Methodology
The use of MI with a geostatistical model was assessed in a simulation. In addition, these procedures were applied to data collected from a 2002 probability survey of Coho salmon located in streams in western Oregon watersheds.

Simulation Example
One realization from a multivariate normal process with mean vector equal to $0$, and a variance covariance matrix equal to $\sigma^2_Z R(\theta) + \sigma^2_\varepsilon I$ over a 21 by 21 regular grid was generated and variances were chosen to be unequal and small. The variance, $\sigma^2_Z = 0.8$ is the variance of the latent spatial random process and $\sigma^2_\varepsilon = 0.2$ is the variance of the non-spatial random process. $R(\theta)$ denotes the one-parameter 21 by 21 correlation matrix generated assuming an exponential correlation function, $e^{-||s_i - s_j||/\theta}$, with $s_i$ and $s_j$ denoting two different sites, and $\theta = 2$ denoting the maximum distance where correlation between two sites is expected.

The parameter $\theta$ is known as the scale parameter and controls how fast the correlation decays with distance. Large values of $\theta$ correspond to a strong spatial correlation and small values to a weak spatial correlation. $I$ is the 21 by 21 identity matrix. This simulated process accounts for spatial variation and measurement error. The collection of 441 observations defines the population values.

To induce a missing at random (MAR) mechanism on the response, stratification was imposed to the region of interest by dividing it into seven equal area vertical regions and then assigning a different response rate to each stratum; each stratum consists of 63 sites. Specification of the response rate range was based on the observed response rates from seven environmental surveys ranging from 0.69 to 0.90, as reported by Herger and Hayslip (2000) and Flitcroft, et al (2002). A range of response rates from 0.70 to 0.90 was assumed and randomly assigned to the seven strata. Within each stratum, 63 values of a uniform random variable $P$ was assigned randomly to the 63 sites. A site, $s$, if selected, would be missing if $P(s) \leq 1 - \alpha$, where $P(s)$ denotes the value of the random variable $P$ assigned to the site $s$, and $\alpha$ denotes the stratum response rate.

Samples of size $n = 152$ were selected at random using equal allocation. Missing rates of 5%, 15%, 25%, 35% and 45% were assumed. For each missing rate, the number of missing sites in the sample was allocated proportional to the stratum response rates. Using the same sampling design, 2,000 samples of size $n = 152$ were generated. The Horvitz-Thompson (HT) mean and variance estimators for the continuous domain (Cordy, 1993) were calculated under the
following settings: (1) the observed data; (2) hot deck imputation; (3) a single imputation obtained from the geostatistical imputation model; (4) the predictive posterior mean imputation calculated as the mean of independent realizations from the predictive posterior distribution at each missing site; (5) hot deck multiple imputation using five and ten multiple imputations for the missing data and (6) multiple imputations for the predictive posterior mean imputation using five and ten multiple imputations for the missing data.

For the single and multiple imputation approaches, a multivariate mixed Gaussian model with constant mean $\beta$ and variance covariance matrix $\sigma^2 R(\theta) + \sigma^2 I$ was assumed. $R(\theta)$ is a correlation matrix that is a function of the distance between sites and an unknown parameter $\theta$. The parameters of the posterior distribution were estimated by implementing MCMC techniques using a MATLAB program (Smith, 2004). An exponential correlation function and a uniform prior for $\beta$, an exponential prior for the correlation parameter with mean 1, and an inverse gamma distribution with parameters $\alpha = 0.1$ and $\beta = 10$ for the variance parameters $\sigma^2_\varepsilon$ and $\sigma^2_Z$ were assumed.

As discussed by both Diggle and Ribeiro (2002) and Banerjee, et al. (2004), these prior selections lead to proper posterior distributions.

Imputation values for the missing data were obtained after verifying that the sample auto-correlations of the MCMC traces were less than 0.01 to ensure independence of the MCMC realizations. Values were randomly selected from the collection of independent realizations and used for the single and multiple imputations.

Salmon Example

This approach was illustrated with the 2002 winter Coho salmon spawning probability survey conducted by the Oregon Department of Fish and Wildlife (ODFW). This survey provides annual inventories of the Coho salmon abundance in streams located within western Oregon watersheds. These streams drain into the Pacific Ocean south of the Columbia River and are considered suitable habitat for salmon (Flitcroft, et al., 2002). The target population consists of all streams located in a United States Geographical Survey (USGS) hydrography data layer of Oregon, except those streams located upstream of large dams that blocked anadromous fish passage (Flitcroft, et al., 2002).

The ODFW uses a generalized random tessellation stratified (GRTS) probability design (Stevens & Olsen, 1999) to select the sample site locations within the population of stream segments. The objective of these surveys is to estimate spawning Coho salmon abundance in both the entire area as well as within five monitoring areas (MA): North Coast, Mid Coast, Mid South Coast, Umpqua and South Coast.

Approximately 120 sites are selected per year within each MA, except in the South Coast MA where the sample size is about 60 sites per year. A total of 495 sites were surveyed in 2002. An additional 61 sites were originally selected in the sample but not visited because of time constraints or inaccessibility of the site location, resulting in 11% missing rate. It was assumed that these missing values resulted from a MAR mechanism. Figure 1 shows the location of the surveyed and missing sites corresponding to the year 2002. Stars represent surveyed sites, and open dots denote the missing sites in the same year. Each sampling site is approximately one-mile in length. At each selected site, counts of spawning Coho are obtained by visual observation. The population abundance of returning adult Coho in individual sites is estimated using area-under-the curve (AUC) techniques (Jacobs, et al., 2002).

Let $Y_i$ denote the total number (abundance) of spawning Coho salmon observed at site $s_i$ in 2002 and $l_i$ be the length of the site $s_i$(in kilometers). Let $\lambda_i$ be the density of spawning Coho salmon (counts per kilometer) at site $s_i, i = 1, \ldots, n$, where $n$ is the total number of surveyed sites. The total number of spawning Coho salmon at each site, $Y_i$, was assumed a noisy version of an unobserved spatial random process $Z_i$, and that conditional on $Z_i$, $Y_i$ has a Poisson distribution with mean $l_i\lambda_i$. In other words, $Y_i | Z_i \sim \text{Poisson}(l_i\lambda_i)$, where $\log(\lambda_i) = \mu_i + Z_i + \varepsilon_i$, where $\mu_i$ denotes a
systematic component, $Z_i$ denotes the spatial random component and $\varepsilon_i$ the non-spatial random component, $i = 1, \ldots, n$.

The systematic component is assumed constant within each MA:

$$
\mu_i = \beta_0 + \sum_{j=1}^{4} \beta_j x_{ij}
$$

where $\beta_1, \beta_2, \beta_3$ and $\beta_4$ are the regression coefficients measuring the MA effects (North Coast, Mid-Coast, Mid-South and Umpqua, respectively, compared to the South Coast MA). The variable $x_{ij}$, is denoted by the value 1 if the $i$th site is located in MA $j$, and 0 otherwise, $i = 1, \ldots, n$, $j = 1, 2, 3, 4$.

Figure 1: Site Locations for ODFW 2002 Spawning Locations
The spatial random process $Z$ is assumed to have a multivariate normal distribution with $\mathbf{0}$ mean vector and variance-covariance matrix given by $\sigma_Z^2 \mathbf{R}(\theta)$, where $\theta$ is the spatial correlation parameter, and $R_{ij}(\theta) = e^{-|s_i - s_j|/\theta}$ denotes the exponential model. The non-spatial random effects, $\varepsilon_i$, are assumed to be independent and normally distributed with mean $0$ and variance $\sigma_{\varepsilon}^2$.

All parameters are assumed independent; vague prior distributions for the parameters were also assumed based on discussions from scientists experienced with these studies. An inverse-gamma ($\alpha = 0.1, \beta = 10$) prior for $\sigma_Z^2$ and $\sigma_{\varepsilon}^2$, which has a wide distribution due to a long tail, and a proper prior $\pi(\theta) = 1/\theta^2$ for $\theta$ on the interval $[0.01, 50]$ was assumed. Selection of the upper limit of 50 kilometers was based on the assumption that it is unlikely to observe spatial correlation beyond this value. For the components of $\beta$, independent improper uniform priors were used. Mathematical expressions for the marginal posterior distributions follow those presented in Christensen and Waagepetersen (2002).

A MATLAB program was used to obtain realizations from the posterior distributions of $\theta$, $\sigma_Z^2$ and $\sigma_{\varepsilon}^2$, and each of the elements of $Z$ and $\beta$ (Smith, 2004). The MCMC simulation was run for 250,000 iterations after a 250,000 burn-in period. In order to reduce serial correlation in the simulated values, particularly in the chain for the parameter $\theta$, each chain was re-sampled to obtain a final sample of 2,500 values of almost uncorrelated values (auto-correlation = 0.01) from the posterior for $\theta, \sigma_Z^2, \sigma_{\varepsilon}^2$ and each of the elements of $\beta, Z,$ and $\log(\lambda)$.

Results
Simulation Example
The Geweke’s statistics and two sided p-value for the model parameters $\beta, \theta, \sigma_Z^2$ and $\sigma_{\varepsilon}^2$ are 0.107 and 0.915; 0.875 and 0.382; 0.871 and 0.384; and 0.826 and 0.401, respectively, suggesting no evidence exists against convergence for each parameter. Similar results were achieved with the Heidelberger and Welch test for the model parameters, suggesting that chain convergence was achieved immediately after the 10,000 burn-in period for each model parameter (p-values for $\beta, \theta, \sigma_Z^2$ and $\sigma_{\varepsilon}^2$ are 0.552, 0.891, 0.926 and 0.784, respectively).

Table 1 shows the simulated root mean squared error (RSME), the average width of the 95% confidence interval, and the coverage rate of the simulated 95% confidence interval for each missing rate. A number of observations can be made from this simulation. As the percentage of missing data increases, the coverage rate decreases. As the missing rate increases, the imputation approaches all appear to be much closer to the 95% coverage as compared to the observed data. The multiple imputation approaches increase the RMSE slightly as compared to the simple and posterior mean imputation approach. In general, all multiple imputation methods ($M = 20$ not shown) performed similarly suggesting that there is no considerable gain in precision with more than 5 imputations.

Salmon Example
Sensitivity to selection of hyper-parameters was explored and no meaningful change was observed in the results. The convergence of the MCMC traces was assessed with the Geweke’s statistic and the Heidelberger and Welch test. The Geweke’s statistics and two sided p-values for the model parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \theta, \sigma_Z^2$ and $\sigma_{\varepsilon}^2$ are $-0.052$ and $0.959$, $-1.081$ and $0.230$, $0.222$ and $0.824$, $-0.154$ and $0.878$, $-0.240$ and $0.810$, $-0.588$ and $0.556$, $0.910$ and $0.363$, and $0.551$ and $0.5821$, respectively, suggesting that no evidence exists against convergence for each parameter. Similar results were achieved with the Cramer-von-Mises statistics for the model parameters, suggesting that chain convergence was achieved for each model parameter (p-values: 0.886, 0.753, 0.921, 0.989, 0.667, 0.410, 0.944, and 0.366). As a result, the iterations
Table 1 Simulated Root Mean Squared Error (RMSE) of the Mean Estimate, Average Width and Coverage Rate of the 95% Confidence Interval for 5%, 15%, 25%, 35% and 45% Missing Rates

<table>
<thead>
<tr>
<th>Missing Response Rate</th>
<th>Analysis Method</th>
<th>RMSE × 100</th>
<th>Width of Interval × 100</th>
<th>Coverage Rate(%) × 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% Missing</td>
<td>Observed Data</td>
<td>5.502</td>
<td>21.569</td>
<td>95.10</td>
</tr>
<tr>
<td></td>
<td>Single Posterior Imputation</td>
<td>5.425</td>
<td>21.266</td>
<td>95.85</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Imputation</td>
<td>5.677</td>
<td>21.319</td>
<td>96.11</td>
</tr>
<tr>
<td></td>
<td>Posterior Mean Imputation</td>
<td>5.423</td>
<td>21.259</td>
<td>96.00</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=5)</td>
<td>5.446</td>
<td>21.349</td>
<td>95.94</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=10)</td>
<td>5.446</td>
<td>21.351</td>
<td>96.00</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=5)</td>
<td>5.601</td>
<td>21.956</td>
<td>93.80</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=10)</td>
<td>5.553</td>
<td>21.768</td>
<td>94.10</td>
</tr>
<tr>
<td>15% Missing</td>
<td>Observed Data</td>
<td>5.480</td>
<td>21.482</td>
<td>92.05</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Imputation</td>
<td>5.509</td>
<td>20.693</td>
<td>93.81</td>
</tr>
<tr>
<td></td>
<td>Single Imputation Data</td>
<td>5.264</td>
<td>20.636</td>
<td>94.55</td>
</tr>
<tr>
<td></td>
<td>Predictive Posterior Mean Imputation</td>
<td>5.259</td>
<td>20.615</td>
<td>94.65</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=5)</td>
<td>5.280</td>
<td>20.700</td>
<td>94.83</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=10)</td>
<td>5.279</td>
<td>20.705</td>
<td>94.85</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=5)</td>
<td>5.432</td>
<td>21.293</td>
<td>93.20</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=10)</td>
<td>5.354</td>
<td>20.988</td>
<td>93.73</td>
</tr>
<tr>
<td>25% Missing</td>
<td>Observed Data</td>
<td>5.477</td>
<td>21.468</td>
<td>88.40</td>
</tr>
<tr>
<td></td>
<td>Single Imputation Data</td>
<td>5.103</td>
<td>20.001</td>
<td>93.10</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Imputation</td>
<td>5.174</td>
<td>20.056</td>
<td>93.36</td>
</tr>
<tr>
<td></td>
<td>Predictive Posterior Mean Imputation</td>
<td>5.093</td>
<td>19.964</td>
<td>92.90</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=5)</td>
<td>5.111</td>
<td>20.035</td>
<td>90.14</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=10)</td>
<td>5.110</td>
<td>20.051</td>
<td>93.35</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=5)</td>
<td>5.382</td>
<td>21.097</td>
<td>90.21</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=10)</td>
<td>5.313</td>
<td>20.827</td>
<td>93.23</td>
</tr>
<tr>
<td>35% Missing</td>
<td>Observed Data</td>
<td>5.490</td>
<td>21.519</td>
<td>82.20</td>
</tr>
<tr>
<td></td>
<td>Single Imputation Data</td>
<td>4.944</td>
<td>19.381</td>
<td>91.45</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Imputation</td>
<td>5.174</td>
<td>19.434</td>
<td>91.70</td>
</tr>
<tr>
<td></td>
<td>Predictive Posterior Mean Imputation</td>
<td>4.931</td>
<td>19.330</td>
<td>91.20</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=5)</td>
<td>4.952</td>
<td>19.414</td>
<td>92.00</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=10)</td>
<td>4.950</td>
<td>19.433</td>
<td>91.90</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=5)</td>
<td>5.264</td>
<td>20.634</td>
<td>89.23</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=10)</td>
<td>5.271</td>
<td>21.662</td>
<td>90.30</td>
</tr>
<tr>
<td></td>
<td>Observed Data</td>
<td>5.480</td>
<td>21.483</td>
<td>73.05</td>
</tr>
<tr>
<td>45% Missing</td>
<td>Single Imputation Data</td>
<td>4.810</td>
<td>18.854</td>
<td>91.55</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Imputation</td>
<td>5.033</td>
<td>18.837</td>
<td>91.80</td>
</tr>
<tr>
<td></td>
<td>Predictive Posterior Mean Imputation</td>
<td>4.792</td>
<td>18.785</td>
<td>90.85</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=5)</td>
<td>4.811</td>
<td>18.863</td>
<td>91.24</td>
</tr>
<tr>
<td></td>
<td>Multiple Imputation (M=10)</td>
<td>4.809</td>
<td>18.887</td>
<td>91.45</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=5)</td>
<td>5.124</td>
<td>20.086</td>
<td>88.70</td>
</tr>
<tr>
<td></td>
<td>Hot Deck Multiple Imputation (M=10)</td>
<td>5.212</td>
<td>20.431</td>
<td>89.23</td>
</tr>
</tbody>
</table>
\( \beta^{(t)}, \theta^{(t)}, \sigma_{Z}^{2}, \sigma_{\varepsilon}^{2}, Z^{(t)} \) and \( \log(\lambda^{(t)}) \) for \( t = 1, \ldots, 2,500 \) can be treated as a sample from the joint posterior distribution \( p(\log(\lambda), Z, \beta, \theta, \sigma_{Z}^{2}, \sigma_{\varepsilon}^{2} | Y) \).

The posterior mean, median and the 95% Bayesian credible interval for each of the parameters in the model are shown in Table 2. The regression coefficients for the region covariates indicate that the MAs Mid-Coast, North Coast, Mid-South Coast and Umpqua tend to have a higher abundance of spawning Coho salmon than the MA South Coast. In addition, the posterior 95% Bayesian credible intervals for all region parameters except the Mid-Coast include zero, suggesting that all MAs except the Mid-Coast have a similar abundance of spawning Coho salmon.

The quantiles for \( \sigma_{Z}^{2} \) (1.93; 4.73) (on the log scale) are above zero, indicating that after the inclusion of the five-level region covariate and the spatial random effect, some additional variability may be attributed to observation error and other small-scale variation not accounted for in the model.

Using the 2,500 iterations of the posterior predictive parameters, the geostatistical imputation model is compared with hot deck imputation. The single imputation method was obtained by selecting one independent draw from the posterior predictive distribution. Multiple imputation was used to assess the impact of the error for this method using five and ten draws. This method was compared to the hot deck imputation, also employing both five and ten imputations.

Finally, the mean of the 2,500 values from the predictive posterior distribution of each missing site was used to estimate the predictive posterior mean for the missing site. These imputation methods are compared with the complete observed data ignoring the missing values. The predicted values were back transformed and the Horvitz-Thompson (HT) estimator for the total estimate for the abundance of spawning Coho salmon, the standard error using the local-variance estimator (Stevens & Olsen, 2003), and the 95% confidence intervals for the total were calculated.

---

**Table 2: Mean, Median, and 95% Bayesian Credible Intervals for the Parameters of the Model**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Median</th>
<th>0.025 Quantile</th>
<th>0.975 Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_{0} ) (South Coast)</td>
<td>0.17</td>
<td>0.16</td>
<td>-1.06</td>
<td>1.41</td>
</tr>
<tr>
<td>( \beta_{1} ) (North Coast)</td>
<td>1.64</td>
<td>1.67</td>
<td>-0.19</td>
<td>3.39</td>
</tr>
<tr>
<td>( \beta_{2} ) (Mid-Coast)</td>
<td>2.48</td>
<td>2.50</td>
<td>0.87</td>
<td>4.07</td>
</tr>
<tr>
<td>( \beta_{3} ) (Mid-South)</td>
<td>1.52</td>
<td>1.51</td>
<td>-0.03</td>
<td>3.11</td>
</tr>
<tr>
<td>( \beta_{4} ) (Umpqua)</td>
<td>1.28</td>
<td>1.28</td>
<td>-0.16</td>
<td>2.68</td>
</tr>
<tr>
<td>( \theta )</td>
<td>17.49</td>
<td>16.10</td>
<td>8.50</td>
<td>34.66</td>
</tr>
<tr>
<td>( \sigma_{Z}^{2} )</td>
<td>3.07</td>
<td>2.98</td>
<td>1.93</td>
<td>4.73</td>
</tr>
<tr>
<td>( \sigma_{\varepsilon}^{2} )</td>
<td>1.39</td>
<td>1.39</td>
<td>0.82</td>
<td>1.95</td>
</tr>
</tbody>
</table>
Table 3 shows a summary of the results; the total estimate using only the observed data provides the lowest total counts estimate of all approaches. No adjustment for missing data was made for this estimate. Examination of the data reveals that the highest level of missing data was found in the Mid-Coast and the highest abundance values were located in this region. All imputation methods that made adjustments for this differential nonresponse across regions provided larger total estimates than the observed data.

The single posterior imputation obtains just one draw and may be more variable than an imputation based on multiple or the mean of multiple draws. The standard error for the MI method is larger than that obtained with the other methods: this was expected because MI accounts for uncertainty due to the imputations (Schafer, 1997). As a result, the 95% confidence intervals using only the observed data (ignoring the missing values), single imputation and mean imputation, are less conservative than that which uses multiple imputation.

Conclusion
Statistical techniques that incorporate the spatial structure of the data in the random and/or systematic part of a model are currently used for modeling environmental phenomena, either discrete or continuous. Therefore, it seems natural to explore the efficiency of a multiple imputation approach that incorporates the spatial structure of the latent process while accounting for missing data. The use of generalized mixed models to account for the missing data in environmental surveys was explored in this article. Generalized mixed models are recent techniques used for modeling environmental phenomena in an attempt to capture any spatial and/or temporal structure in the data. The possibility of implementing generalized linear models to different data distributions make them appealing for handling missing data in environmental surveys. Evaluations of the selection of the priors and the model specifications are performed before any imputation is conducted. This allows the researcher to explore different models for the covariance matrix and different priors that may better reflect the study data.

Simulation results from this study suggest that all imputation methods perform well at 5% and 15% missing rates. When the missing rate is 15% or higher, the performance of the statistics decays similarly for all techniques considered. However, the coverage rates for the 95% confidence intervals for all imputation methods are improved over no imputation. The performance of the statistics observed with 5 and 10 multiple imputations at all response rates, suggests that as in human populations (Schafer, 1997, Little & Rubin, 2002), little is gained when the number of imputations exceeds 5.

The method was illustrated by estimating the mean of an environmental

<table>
<thead>
<tr>
<th>Imputation Technique</th>
<th>Total</th>
<th>SE</th>
<th>0.025 Quantile</th>
<th>0.975 Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Data (No Imputation)</td>
<td>227,885</td>
<td>16,648</td>
<td>195,255</td>
<td>260,514</td>
</tr>
<tr>
<td>Hot Deck Imputation</td>
<td>249,271</td>
<td>16,966</td>
<td>216,018</td>
<td>282,524</td>
</tr>
<tr>
<td>Single Posterior Imputation</td>
<td>238,185</td>
<td>16,919</td>
<td>205,023</td>
<td>271,346</td>
</tr>
<tr>
<td>Posterior Mean Imputation</td>
<td>250,921</td>
<td>16,519</td>
<td>218,543</td>
<td>283,298</td>
</tr>
<tr>
<td>MI Hot Deck (m=5)</td>
<td>257,931</td>
<td>18,193</td>
<td>222,274</td>
<td>293,589</td>
</tr>
<tr>
<td>MI Posterior (m=5)</td>
<td>250,213</td>
<td>21,689</td>
<td>206,302</td>
<td>294,127</td>
</tr>
</tbody>
</table>
variable, the abundance of spawning Coho salmon in the Oregon coastal streams. It is expected that multiple imputation methods which incorporate auxiliary information into the systematic part may render better results than the observed data. By incorporating auxiliary variables correlated with the process of interest into an imputation geostatistical model, the variances of the spatial component and the measurement error may be reduced resulting in narrowed posterior prediction intervals for the missing data. This implies that imputations may be closer to the unobserved true value, which will improve the imputation results. However, given the variability expected in natural environments, it is important to account for the imputation error through a multiple imputation approach.

References


MI TO ACCOUNT FOR ITEM NONRESPONSE IN ENVIRONMENTAL DATA


On the Appropriate Transformation Technique and Model Selection in Forecasting Economic Time Series: An Application to Botswana GDP Data

University of Botswana, Botswana

Selected data transformation techniques in time series modeling are evaluated using real-life data on Botswana Gross Domestic Product (GDP). The transformation techniques considered were modified, although reasonable estimates of the original with no significant difference at \( \alpha = 0.05 \) level were obtained: minimizing square of first difference (MFD) and minimizing square of second difference (MSD) provided the best transformation for GDP, whereas the Goldstein and Khan (GKM) method had a deficiency of losing data points. The Box-Jenkins procedure was adapted to fit suitable ARIMA \((p, d, q)\) models to both the original and transformed series, with AIC and SIC as model order criteria. ARIMA \((3, 1, 0)\) and ARIMA \((1, 0, 0)\) were identified, respectively, to the original and log of the transformed series. All estimates of the fitted stationary series were significant and provided a reliable forecast.

Key words: Data transformation technique, autoregressive integrated moving average, model order criteria, forecast, gross domestic product.

Introduction
The foremost difficulty with economic research in developing countries is the dearth of data. Much of the available economic time series data are constructed out of bits and pieces that must be shaped and arranged to yield a final series that is usable for model building. One way to circumvent this problem is to estimate some components for dates for which time series is not readily available from known values of that component for other dates. For example, the US real Gross Domestic Product (GDP) and German real GDP are produced and publicly released at quarterly intervals, although both US and German economic analysts and business-decision makers often need monthly GDP forecasts (Stum & Wollmershauser, 2005), quarterly figures may be required only when the series of annual data are available. This problem has led to several transformations of the data to the form required by researchers for particular research objectives. Economists use many transformations of time series data to help extract economically relevant information (Cohen, 2001).

A facet of the research conducted focuses on the interpolation of some values of a series at a given time period by a related series (Friedman, 1962). The problem with this technique is that it assumes that a related series - as well as some values of the series to be interpolated - are readily available: this may not be the case in developing countries. Various studies have been concerned with the derivation of quarterly figures from annual data, including Lismman and Sandee (1964), Boot, et al. (1967) and Goldstein & Khan (1976); in each of these examples the value of a quarterly figure for each year \( t \), is considered as a weighted average of the totals of the years. A system of equations is built from which weighted coefficients were calculated subject to some criteria.

D. K. Shangodoyin is Associate Professor at the Department of Statistics. Email: shangodoyink@mopipi.ub.bw. K. Setlhare is Senior Lecturer at the Department of Statistics. Email: setlharek@mopipi.ub.bw. K. K. Moseki is a Lecturer at the Department of Statistics. Email: mpsekikk@mopipi.ub.bw. K. Sediakgotla is a Lecturer at the Department of Statistics. Email: sediakgotla@mopipi.ub.bw.
The challenge, therefore, is to explore the efficiency of the transformation techniques and analyze their prediction potential. Some transformation techniques can be found in Boot, et al. (1967) which introduced two methods that involve minimizing the squared first differences (MSFD) and minimizing the squared second differences (MSSD). Goldstein and Khan (1976) proposed an interpolation technique based on the quadratic function: the transformed data could be modeled appropriately by checking the order of the fitted model using model order selection criteria as discussed by Shibata (1976).

In this article, the focus is to evaluate the efficacy of data transformation techniques with the aim of using two known models’ order determination criteria to produce the best model order-transformation technique for forecasting economic time series with application to Botswana GDP data. This is considered a challenge to analysts in view of the dearth of quarterly economic series data in some sectors of a nation’s economy where only annual data is available.

Methodology
The Technique and Model Determination

Boot, et al. (1967) considered two procedures for the interpolation of quarterly figures given only annual data; the basis of their research is the work of Lisman & Sandel (1964). The first approach is based on the criterion that minimizes the sum of square of the first difference (MFD) between the successive quarterly values, which are subject to the constraint that, each year, the sum of the quarterly total should equal the yearly totals.

Mathematically, consider \( n \) years for which it is necessary to minimize

\[
\sum_{i=2}^{4n} (x_i - x_{i-1})^2 \quad \text{subject to} \quad \sum_{i=4t-3}^{4k} x_i = y_i, \quad t = 1, 2, \ldots n
\]

where \( x_i \) is the estimated quarterly value in years \( k = 1, 2, 3 \) and quarters \( i = 1, 2, 3, 4 \), \( t = 1, 2, 3, 4 \) and \( y_{i-1}, y_i, y_{i+1} \) are the totals for the three successive years, and

\[
\begin{aligned}
X_{11} &= \begin{pmatrix} 569 & -135 & 25 \\ 525 & -87 & 15 \\ 437 & 27 & -5 \\ 305 & 189 & -35 \end{pmatrix} \\
X_{12} &= \begin{pmatrix} 129 & 405 & -75 \\ 7 & 513 & -61 \\ -61 & 513 & 7 \\ -75 & 405 & 129 \end{pmatrix} \\
X_{13} &= \begin{pmatrix} -35 & 189 & 305 \\ -5 & 27 & 437 \\ 15 & -187 & 525 \\ 25 & -135 & 569 \end{pmatrix}
\end{aligned}
\]

The MFD derived formula for calculating the estimated quarterly total within three successive years is given as:

\[
\begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \end{pmatrix} = \frac{1}{1836} \begin{pmatrix} X_{11} & \cdots & y_{t-1} \\ \cdots & \cdots & \cdots \\ X_{12} & \cdots & y_{t} \\ \cdots & \cdots & \cdots \\ X_{13} & \cdots & y_{t+1} \end{pmatrix} \quad (i)
\]

The second approach is the minimization of the sum of square of the second difference (MSD) in which

\[
\sum_{i=2}^{4n} (\Delta x_i - \Delta x_{i-1})^2 \quad ,
\]

is minimized, where \( \Delta x_i = x_{i+1} - x_i \), is subject to the constraint

\[
\sum_{i=4t-3}^{4t} x_i = y_i, \quad t = 1, 2, \ldots, n,
\]

and the \( x_i \)'s are as defined above. Similar to the MSFD, the problem is solved by considering the Lagrangean expression.
In the expressions, the first year will have \( y_{t-1} = y_t = 0 \) and the second year will have \( y_{t-1} = 0 \) in the computation of the quarterly total for the years, assuming \( y_{t-1}, y_t \) and \( y_{t+1} \) are independent aggregates. Lisman and Sandel (1964) assumed that the quarterly data, for example, \( Z_i \), was linearly dependent on three successive annual totals and proposed the computation of quarterly data from the following:

\[
\left( \begin{array}{c}
0.0729 \\
-0.0103 \\
-0.0415 \\
-0.0211
\end{array} \right) = \left( \begin{array}{ccc}
0.1982 & -0.0211 \\
0.3018 & -0.0415 \\
0.3018 & -0.0103 \\
0.1982 & 0.0729
\end{array} \right) \left( \begin{array}{c}
y_{t-1} \\
y_t \\
y_{t+1}
\end{array} \right)
\]

(iv)

All of these methods are known to have limitations (Boot, et al., 1967), thus other mathematical methods of interpolation have been developed by researchers such as Glejer (1966), Boots and Feibes (1967) and Vangrevelinghe (1966). The choice of method as described in (i)-(iv) is based on the similarity in their computation. It would be of tremendous assistance to analysts if the various methods are subjected to real-life data experimentation, while the transformed data are modeled with an appropriate check on the models order to ascertain their suitability in forecasting.

In this article it is assumed that the \( y \)'s are moving by 3 points, models are run up to \( n-2 \), and the identified (or fitted) model is used to compute \( n-1 \) and \( n \) so that no year is omitted and the model provides a reasonable degree appropriateness for the transformed data. The Box-Jenkins modeling was performed on both the original and transformed data with a view to forecast. However, the unknown value of the model order, \( P \), may constitute a casualty in modeling as attempts to under fit increases the residual variance, while over fitting results in too many parameters which eventually causes unreliability (Jones, 1975; Shibata, 1976). Various selection criteria have been advanced for model order selection (Box, Jenkins & Reinsel, 1994), in this article, three similar
criteria were employed vis-à-vis the Akaike information criteria (AIC) \( \{ N \ln \sigma_p^2 + 2p \} \), final predictor error (FPE) \( \{ \frac{N + p}{N - p} \sigma_p^2 \} \) and Schwarz’s criterion (SIC) \( \{ N \ln \sigma_p^2 + P \ln N \} \). The order in which two of these criteria agree shall be considered to be the best order for the data.

Results

Data Analysis: Transformation and Modeling of Botswana GDP Data

Data presented in Appendix I shows that no significant variation exists between the average values of data computed by the three techniques and the original data. The test of difference conducted between the original series and the transformed series indicates that there is no significant difference between the means of the GDP, MFD, MSD and GKM. It was observed (see Appendix II), that the MFD and MSD provided the best transformation for the Botswana GDP data while the GKM had a deficiency of losing data points. The proposed method of moving point incorporated into the selected techniques is shown to be worthwhile because neither the MFD nor the MSD lost any data.

Model Selection and Order Determination

The original GDP series is made stationary by taking the first difference (see Appendix II) - an autoregressive process of order 3 is identified as the most suitable model. Based on AIC and SIC criteria, the fitted values (Appendix II) are adequate as indicated in Figure 2 and the bounds placed on the fitted values appear to have accommodated the original values adequately.

The MFD, MSD and GKM series became stationary only when the log-transformation was taken, the AIC, SIC and model RESIDUALS were the criteria used in selecting the best order for the model and these identified the AR (1) models to MFD, MSD and GKM. The behavior of the fitted values (see Appendix III, Figures 1-4 and Tables 1--4) indicate the appropriateness of the model as confirmed by the Portmanteaux test for model adequacy.

Conclusion

The moving point method introduced into the transformation techniques utilized in this research has shown a tremendous improvement over the MFD and MSD. It was observed that both MFD and MSD give nearly the same fitted values as the original series; thus confirming the findings of Shangodoyn and Adubi (2000) who used Nigeria GDP data. The choice of the model order should not, however, be limited to the order determination criteria but also to the model residual variance.

References


Appendix I: One-Way Analysis of Variance

The results indicate that no significant difference exists between the means of the four series (original, FMD, MSD, and GKM), at $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor</td>
<td>3</td>
<td>497763</td>
<td>165921</td>
<td>0.06</td>
<td>0.981</td>
</tr>
<tr>
<td>Error</td>
<td>132</td>
<td>372383781</td>
<td>2821089</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>135</td>
<td>372881545</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Individual 95% CIs for Mean Based on Pooled StDev

| Level     | N   | Mean | StDev | ---+---------+---------+---------+--- |
|-----------|-----|------|-------|-----+---------+---------+---------+--- |
| ORIGINAL  | 36  | 5068 | 1765  | (---------------*---------------) |
| MFD       | 36  | 5068 | 1742  | (---------------*---------------) |
| MSD       | 36  | 5068 | 1745  | (---------------*---------------) |
| GKM       | 28  | 4918 | 1368  | (---------------*---------------) |
|           |     |      |       | +-----------------+-----------------+--- |

Pooled StDev = 1680 4400 4800 5200 5600
### Appendix II: Original and Transformed Series with Forecast Values

<table>
<thead>
<tr>
<th>Actual GDP</th>
<th>MFD</th>
<th>MSD</th>
<th>GKM</th>
<th>Fitted GDP Values</th>
<th>Fitted MFD Values</th>
<th>Fitted MSD Values</th>
<th>Fitted GKM Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>2918.5</td>
<td>2713.68</td>
<td>2669.984</td>
<td>2924.033</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>2633</td>
<td>2732.35</td>
<td>2728.983</td>
<td>3009.161</td>
<td>NA</td>
<td>2812.433</td>
<td>2766.022</td>
<td>3049.796</td>
</tr>
<tr>
<td>2667.8</td>
<td>2769.68</td>
<td>2789.227</td>
<td>2962.78</td>
<td>NA</td>
<td>2913.945</td>
<td>2864.974</td>
<td>3178.91</td>
</tr>
<tr>
<td>2822</td>
<td>2825.68</td>
<td>2853.205</td>
<td>3228.221</td>
<td>NA</td>
<td>3018.262</td>
<td>2966.907</td>
<td>3311.378</td>
</tr>
<tr>
<td>2964.8</td>
<td>2900.35</td>
<td>2924.651</td>
<td>3308.047</td>
<td>3310.34</td>
<td>3125.433</td>
<td>3071.894</td>
<td>3447.203</td>
</tr>
<tr>
<td>2959.3</td>
<td>2996.24</td>
<td>3008.543</td>
<td>3451.497</td>
<td>3004.924</td>
<td>3235.503</td>
<td>3180.005</td>
<td>3586.381</td>
</tr>
<tr>
<td>3074.2</td>
<td>3113.37</td>
<td>3107.371</td>
<td>3407.489</td>
<td>2939.061</td>
<td>3348.52</td>
<td>3291.311</td>
<td>3728.91</td>
</tr>
<tr>
<td>3263.4</td>
<td>3251.74</td>
<td>3221.135</td>
<td>3704.262</td>
<td>3311.223</td>
<td>3464.531</td>
<td>3405.888</td>
<td>3874.781</td>
</tr>
<tr>
<td>3469.9</td>
<td>3411.33</td>
<td>3347.346</td>
<td>4148.561</td>
<td>3598.809</td>
<td>3583.582</td>
<td>3523.809</td>
<td>4023.983</td>
</tr>
<tr>
<td>3330.3</td>
<td>3531.03</td>
<td>3481.024</td>
<td>4349.12</td>
<td>3559.041</td>
<td>3705.721</td>
<td>3645.149</td>
<td>4176.504</td>
</tr>
<tr>
<td>3325.3</td>
<td>3610.82</td>
<td>3618.437</td>
<td>4274.65</td>
<td>3369.796</td>
<td>3830.992</td>
<td>3769.985</td>
<td>4332.328</td>
</tr>
<tr>
<td>4078.4</td>
<td>3650.72</td>
<td>3757.094</td>
<td>4604.015</td>
<td>3554.994</td>
<td>3959.443</td>
<td>3898.395</td>
<td>4491.436</td>
</tr>
<tr>
<td>4540.1</td>
<td>4308.45</td>
<td>4172.611</td>
<td>4856.836</td>
<td>4346.209</td>
<td>4091.118</td>
<td>4030.458</td>
<td>4653.806</td>
</tr>
<tr>
<td>4280.4</td>
<td>4359.09</td>
<td>4349.4</td>
<td>4987.995</td>
<td>4649.145</td>
<td>4226.064</td>
<td>4166.252</td>
<td>4819.415</td>
</tr>
<tr>
<td>4385</td>
<td>4460.37</td>
<td>4524.359</td>
<td>4895.197</td>
<td>4367.403</td>
<td>4364.324</td>
<td>4305.858</td>
<td>4988.234</td>
</tr>
<tr>
<td>4534.7</td>
<td>4612.29</td>
<td>4693.829</td>
<td>5311.871</td>
<td>4600.11</td>
<td>4505.945</td>
<td>4449.359</td>
<td>5160.236</td>
</tr>
<tr>
<td>4894.5</td>
<td>4814.85</td>
<td>4852.319</td>
<td>5175.132</td>
<td>4762.759</td>
<td>4650.971</td>
<td>4596.837</td>
<td>5335.387</td>
</tr>
<tr>
<td>5107.8</td>
<td>4986.19</td>
<td>4992.512</td>
<td>5287.936</td>
<td>5018.153</td>
<td>4799.445</td>
<td>4748.376</td>
<td>5513.653</td>
</tr>
<tr>
<td>4861.9</td>
<td>5126.32</td>
<td>5110.746</td>
<td>5208.404</td>
<td>5225.658</td>
<td>4951.41</td>
<td>4904.06</td>
<td>5694.997</td>
</tr>
<tr>
<td>5298.3</td>
<td>5235.24</td>
<td>5207.023</td>
<td>5697.818</td>
<td>5064.246</td>
<td>5106.911</td>
<td>5063.975</td>
<td>5879.38</td>
</tr>
<tr>
<td>5614.1</td>
<td>5312.94</td>
<td>5285.001</td>
<td>5906.844</td>
<td>5501.775</td>
<td>5265.99</td>
<td>5228.209</td>
<td>6066.761</td>
</tr>
<tr>
<td>5937.3</td>
<td>5371.21</td>
<td>5352</td>
<td>6120.156</td>
<td>5749.231</td>
<td>5428.688</td>
<td>5396.848</td>
<td>6257.095</td>
</tr>
<tr>
<td>4578.2</td>
<td>5410.06</td>
<td>5413.51</td>
<td>6021.465</td>
<td>6074.464</td>
<td>5595.048</td>
<td>5569.981</td>
<td>6450.337</td>
</tr>
<tr>
<td>5394.1</td>
<td>5429.49</td>
<td>5473.19</td>
<td>6532.291</td>
<td>4770.482</td>
<td>5765.11</td>
<td>5747.699</td>
<td>6646.438</td>
</tr>
<tr>
<td>6144.7</td>
<td>6059.24</td>
<td>5875.95</td>
<td>6831.44</td>
<td>5582.387</td>
<td>5938.915</td>
<td>5930.091</td>
<td>6845.349</td>
</tr>
<tr>
<td>6444.7</td>
<td>6129.85</td>
<td>6116.536</td>
<td>7050.604</td>
<td>6288.624</td>
<td>6116.503</td>
<td>6117.25</td>
<td>7047.018</td>
</tr>
<tr>
<td>5856.1</td>
<td>6271.08</td>
<td>6356.419</td>
<td>6929.72</td>
<td>6593.752</td>
<td>6297.914</td>
<td>6309.268</td>
<td>7251.391</td>
</tr>
<tr>
<td>6497.6</td>
<td>6482.93</td>
<td>6594.195</td>
<td>7519.338</td>
<td>6040.774</td>
<td>6483.185</td>
<td>6506.238</td>
<td>7458.412</td>
</tr>
<tr>
<td>7144.8</td>
<td>6765.39</td>
<td>6827.755</td>
<td>NA</td>
<td>6676.607</td>
<td>6672.355</td>
<td>6708.256</td>
<td>7668.025</td>
</tr>
<tr>
<td>7009.6</td>
<td>7035.87</td>
<td>7054.289</td>
<td>NA</td>
<td>7295.269</td>
<td>6865.461</td>
<td>6915.416</td>
<td>NA</td>
</tr>
<tr>
<td>6906.9</td>
<td>7294.36</td>
<td>7272.392</td>
<td>NA</td>
<td>7165.869</td>
<td>7062.54</td>
<td>7127.815</td>
<td>NA</td>
</tr>
<tr>
<td>7575.2</td>
<td>7540.87</td>
<td>7482.064</td>
<td>NA</td>
<td>7085.977</td>
<td>7263.628</td>
<td>7345.549</td>
<td>NA</td>
</tr>
<tr>
<td>7794.8</td>
<td>7775.39</td>
<td>7684.71</td>
<td>NA</td>
<td>7748.615</td>
<td>7468.76</td>
<td>7568.718</td>
<td>NA</td>
</tr>
<tr>
<td>7269.2</td>
<td>7951.28</td>
<td>7883.141</td>
<td>NA</td>
<td>7950.03</td>
<td>7677.97</td>
<td>7797.419</td>
<td>NA</td>
</tr>
<tr>
<td>7924.1</td>
<td>8068.55</td>
<td>8079.464</td>
<td>NA</td>
<td>7429.783</td>
<td>7891.292</td>
<td>8031.753</td>
<td>NA</td>
</tr>
<tr>
<td>8934.4</td>
<td>8127.18</td>
<td>8275.084</td>
<td>NA</td>
<td>8099.146</td>
<td>8108.759</td>
<td>8271.82</td>
<td>NA</td>
</tr>
</tbody>
</table>
### Appendix III: Graphs and Tables Results

#### Table 1: GDP at Constant 1993/94 Prices in P’000 000

<table>
<thead>
<tr>
<th>Identified Model</th>
<th>ARIMA(p,1,0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order of Model</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>179.8878(68.48002)</td>
</tr>
<tr>
<td>a1</td>
<td>-0.1669(0.1812)</td>
</tr>
<tr>
<td>a2</td>
<td>na</td>
</tr>
<tr>
<td>a3</td>
<td>na</td>
</tr>
<tr>
<td>AIC</td>
<td>15.1775</td>
</tr>
<tr>
<td>SIC</td>
<td>15.2673</td>
</tr>
<tr>
<td>Residual Var</td>
<td>6909122</td>
</tr>
<tr>
<td>Best Model:</td>
<td>ARIMA(3,1,0)</td>
</tr>
<tr>
<td>Forecasting Model</td>
<td>X_{t+m} = 166.0609 + GDP(-1) – 0.6071GDP(-1) – 0.8321GDP(-2) – 0.45589GDP(-3)</td>
</tr>
</tbody>
</table>

#### Table 2: Results of Fitted Model on MFD Series

<table>
<thead>
<tr>
<th>Identified Model</th>
<th>ARIMA (p, 0, 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Order of Model</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>12.3638(7.2757)</td>
</tr>
<tr>
<td>a1</td>
<td>0.992(0.015)</td>
</tr>
<tr>
<td>a2</td>
<td>na</td>
</tr>
<tr>
<td>a3</td>
<td>na</td>
</tr>
<tr>
<td>AIC</td>
<td>-4.098</td>
</tr>
<tr>
<td>SIC</td>
<td>-4.009</td>
</tr>
<tr>
<td>Residual Var</td>
<td>0.03066</td>
</tr>
<tr>
<td>Best Model:</td>
<td>ARIMA(1,1,0)</td>
</tr>
<tr>
<td>Forecasting Model</td>
<td>X_{t+m} = exp(12.3636 + 0.992MFD(-1))</td>
</tr>
</tbody>
</table>
Appendix III: Graphs and Tables Results (continued)

Table 3: Results of Fitted Model on MSD Series

<table>
<thead>
<tr>
<th>MSD</th>
<th>Stationary</th>
<th>Logarithm Transformation</th>
<th>ARIMA(p,0,0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified Model</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Order of Model</td>
<td>c</td>
<td>a1</td>
<td>a2</td>
</tr>
<tr>
<td></td>
<td>14.502(9.4691)</td>
<td>11.844(4.2843)</td>
<td>10.7887(2.4542)</td>
</tr>
<tr>
<td></td>
<td>0.9947(0.0084)</td>
<td>1.2938(0.1696)</td>
<td>1.2326(0.1822)</td>
</tr>
<tr>
<td></td>
<td>na</td>
<td>-0.3006(0.1690)</td>
<td>-0.0995(0.2917)</td>
</tr>
<tr>
<td></td>
<td>na</td>
<td>na</td>
<td>-0.1419(0.1803)</td>
</tr>
<tr>
<td>AIC</td>
<td>-5.252</td>
<td>-5.2796</td>
<td>-5.222</td>
</tr>
<tr>
<td>SIC</td>
<td>-5.163</td>
<td>-5.1449</td>
<td>-5.0412</td>
</tr>
<tr>
<td>Residual Variance</td>
<td>0.009471</td>
<td>0.08501</td>
<td>0.008718</td>
</tr>
<tr>
<td>Best Model</td>
<td>Log(SMD) = 14.502 + 0.9947SMD(-1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forecasting Model</td>
<td>( \hat{X}_{t+m} = \exp(14.502 + 0.9947SMD(-1)) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Forecast and MSD Values

Figure 4: Forecast and Values Of GKM
### Table 4: Results of Fitted Model on GKM Series

<table>
<thead>
<tr>
<th>Stationary</th>
<th>Logarithm Transformation</th>
<th>GKM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified Model</td>
<td>ARIMA(p,0,0)</td>
<td></td>
</tr>
<tr>
<td>Order of Model</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>10.7201(4.4355)</td>
<td>10.7123(3.4997)</td>
</tr>
<tr>
<td>a1</td>
<td>0.9846(0.0298)</td>
<td>0.6679(0.2054)</td>
</tr>
<tr>
<td>a2</td>
<td>na</td>
<td>0.3118(0.2022)</td>
</tr>
<tr>
<td>a3</td>
<td>na</td>
<td>na</td>
</tr>
<tr>
<td>AIC</td>
<td>-3.4161</td>
<td>-3.399</td>
</tr>
<tr>
<td>SIC</td>
<td>-3.3201</td>
<td>-3.2537</td>
</tr>
<tr>
<td>Residual Variance</td>
<td>0.044768</td>
<td>0.040382</td>
</tr>
<tr>
<td>Best Model</td>
<td>( \text{Log}(GKM) = 10.7201 + 0.9846GKM(-1) )</td>
<td>( \hat{X}_{t+m} = \exp(10.7201 + 0.9846GKM(-1)) )</td>
</tr>
</tbody>
</table>
Optimal Meter Placement by Reconciliation Conventional Measurements and Phasor Measurement Units (PMUs)

Reza Kaihani  Ali Reza Seifi
Shiraz University, Iran

The success of state estimation depends on the number, type and location of the established meters and RTUs on the system. A new method by incorporating conventional measurements and New Technology of Phasor Measurement Units (PMU) is proposed. Conventional meters (power injection and power flow measurements) are allocated in order to reduce the number of meters, RTUs, critical measurements, critical sets and leverage points, and also to improve the numerical stability of equations; a genetic algorithm is used for optimization. A second step involves adding PMUs in areas in which it is expected that the accuracy of state estimation will be improved.

Key words: State estimation, meter placement, network observability, Phasor Measurement Units, PMU, leverage points.

Introduction

Current energy management systems (EMSs) must accurately monitor power system state variables, i.e. the voltage phasors (voltages in module and phase) of each bus in real time. The primary monitoring tool is the state estimation (SE), which constitutes the core of all control operations. Installing a new state estimator or upgrading an existing one requires - among other considerations - evaluation of the metering configuration. Determination of the best possible combination of meters for monitoring a given power system is referred to as the optimal meter placement problem. Fundamentally, the metering scheme must provide enough information to allow power system state estimation. Planning metering systems for power system monitoring is a complex task, not only due to the problem dimension itself (number of possible configurations), but also to the need of establishing a trade-off between SE performance and metering system costs. With an adequate metering system, the SE can successfully process the available information and obtain reliable estimates of system operating conditions, which can then be used for further analyzes and for control actions.

In the design of measurement point locations, first it should be considered that the measurement system must satisfy the basic condition of state estimation: observability of the network. In addition to this essential prerequisite, it is also necessary to consider other issues such as accuracy, reliability and economy. Network observability analysis determines whether the network is observable or not by the type and placement of the measurements; the topology of network observability is related to the type and placement of the measurements.

Several methods of network observability analysis, such as, numerical (Abur & Expósito, 2004; Monticelli & Wu, 1986) and topological (Abur & Expósito, 2004; Krumpholz, Clements & Davis, 1980) have been introduced to determine if the network is observable or island observable. Implementation of synchronized phasor measurements presents
an opportunity for improvements in power system state estimation.

As an addition to standard real and reactive power and voltage and current magnitude measurements, the Phasor Measurement Unit (PMU) provides voltage and current phasor measurements \((\theta, Ie, Ve)\). PMUs provide positive sequence value measurements, are available from 20 times per cycle and can be synchronized with measurements obtained from another PMU. The PMUs are accurate and can take measurements synchronously, thus improving the performance of state estimation.

Several research articles have been published regarding algorithms developed with the objective of attaining optimal measurement plans for power system state estimation, taking into account some of the previously described requirements. Cost of installing new meters and reduction of the number of critical p-set measurements are taken into consideration in the formulations of (Souza, et al., 2005; Mori & Iida, 1993; Riccieri & Falcão, 1999; Mori & Matsuzaki, 1999; Allemong, Radu & Sasson, 1982; Antonio, Torrão & Filho, 2001). Accuracy of the weighted least squares state estimation for the chosen measurement design is also used as one of the objectives in these studies (Mori & Iida, 1993). In Monticelli and Wu (1986) and Magnago and Abur (2000) a metering system was designed for a basic network and possible occurrence of topology changes and/or measurement losses. Pioneering work in PMU development and utilization has been accomplished by Phadke, et al. (1986). For details on PMU placement problems in power systems, see Zovaneloc & Cairns, 1996; Milosevic & Begovic, 2003; and Rice & Heydt, 2006.

In this article planning of measurement systems is implemented in two steps. The first step uses conventional meters, power injection measurement and power flow measurement in pairs unit (P, Q) to achieve a primary outline. The objective function in this step is observability, reducing the cost of meters and RTU, decreasing critical measurements and critical sets, minimizing the number of leverage points and improving numerical stability. The second step adds PMUs to the primary design from step 1 to improve the accuracy state estimation and to speed up convergence. That detection of bad data in bad leverage points due to errors or malfunctions in meters is not simple and the number of leverage points is reduced to avoid this situation; as is known, however, the existence of good leverage points (free of bad data) causes the accuracy of state estimation to increase (Abur & Expósito, 2004). To avoid losing the positive qualification by adding PMUs, in step 2 the accuracy will be increased to compensate for the absence of leverage points in the pre-designed measurement system developed in step 1.

Linear State Estimation

The conventional method for power system state estimation is the weighted least squares (WLS) state estimation (Abur & Expósito, 2004). The WLS state estimator equations relating to the measurements and the state vector are:

\[
\begin{align*}
  z &= Hx + e
\end{align*}
\]

where \(x\) and \(z\) are the \(n\times1\) state and \(m\times1\) measurement vectors; \(H\) is the \(m\times n\) Jacobian matrix, \(e\) is the \(m\times1\) measurement error vector, \(m\) is the number of measurements and \(n\) is the number of buses. The SE can be formulated as weighted least-squares (WLS) problem

\[
\min J(x) = [z - Hx]R^{-1}[z - Hx]
\]

The state estimate \(\hat{x}\) by minimization \(J(x)\) in \([\ldots]\) can be obtained through the WLS method by satisfying the following Optimality condition:

\[
\frac{\partial J(x)}{\partial x} = H^TR^{-1}[z - H\hat{x}] = 0
\]

\[
\hat{x} = G^{-1}H^TR^{-1}z
\]

where \(G = H^TR^{-1}H\) is known as the gain matrix.
Observability Analysis

Observability analysis is a search process for portions of a power network; given the network and measurement topology, state estimation can be performed. Usually, the linearized and decoupled state estimator is adopted to perform observability analysis. Hereafter, for simplicity, the Po (active power-angle) model will be used. A system is said to be observable if the gain matrix is nonsingular, which can be verified during its triangular factorization (no zero pivots, if the reference bus angle is not included) (see Abur & Expósito, 2004; Monticelli & Wu, 1986; Krumpholz, Clements & Davis, 1980).

Condition Number

The condition number of a nonsingular square matrix $A$ is defined as:

$$K_G = \|A\|\|A^{-1}\|$$

where $\|\cdot\|$ represents a matrix norm (Abur & Expósito, 2004; Rice & Heydt, 2007; Reza & Ross, 2001). If 2-norm is used, the condition number can be calculated using

$$K_G = \frac{\lambda_s}{\lambda_1}$$

where $\lambda$ denotes the eigenvalues of $A$ respectively, subscript $s$ refers to the largest eigenvalues and subscript 1 refers to the smallest values. The condition number is equal to unity for identity matrices and tends to infinity for matrices approaching singularity. A large condition number in value is indicative of an ill-conditioned matrix (Abur & Expósito, 2004).

In state estimation, the sensitivity of the estimate of $x$ to noise is improved (lessened) when $K_G$ (the condition number of gain matrix) is small, and the sensitivity is worsened (increased) when $K_G$ is large. Typical threshold values of $K_G$ in state estimation applications, beyond which designers of a state estimator become concerned, are approximately 105 (Rice & Heydt, 2007).

Leverage Points

Some measurements of a power system may have a much stronger influence on the state estimate than others due to their location, the local measurement redundancy, the network topology and parameters. These points are outliers in the space spanned by the row vectors of the Jacobian matrix, meaning that they do not follow the pattern of the point cloud in that space. Such measurements, referred to as leverage measurements, will distort the solution of the least absolute value estimation when they carry bad data.

Two cases are associated with leverage points. When a measurement is a leverage point and has a wrong metered value, it is a bad leverage point; identification of the bad measurement becomes very difficult by conventional methods. Residual covariance for these measurements will be numerically insignificant. If, however, the measurement is a leverage point and has a good metered value, it is a good leverage point and heavily reinforces the M-estimator’s performance.

Projection Statistics

A robust measure of levering the effect of a measurement was applied to the power system state estimation by Mili, et al., 1996; this measure is called the projection statistic (PSi) and is defined for a measurement $i$ as

$$PS_i = \max_k \left| H_i^T \cdot H_i \right| / \beta$$

for $k = 1, 2, ..., m$

where

$$\beta = 1.926 \text{lomed}_i \left\{ \text{lomed}_{j \neq i} \left| H_j^T \cdot H_k + H_j^T \cdot H_k \right| \right\}$$

$1 \leq i, j, k \leq m$

and $\text{lomed}_i \{x\}$ is the low median of the $m$ number in $x = \{x_1, x_2, ..., x_m\}$.

The projection statistic $PS_i$ can be shown to behave approximately like a Chi-square random variable. Measurement $i$ is the related to the sparsity structure of the row $H_i$. Hence, measurement $i$ is identified as a leverage point if $PS_i > \chi^2_{k, 0.975}$ where, $k$ is the number of
nonzero entries in the row $H_i$ of the measurement Jacobian $H$.

Classification of Measurements

Critical Measurement

Critical measurement is one whose elimination from the measurement will result in an unobservable system. The residual and standard deviation associated with a critical measurement always equals zero.

\[
\begin{align*}
    r(i) &= z(i) - \hat{z}(i) \\
    \sigma_E(i) &= \sqrt{E(i,i)} = 0
\end{align*}
\]

where

\[
E = R - HG^{-1}H^T
\]

and

\[
\hat{z} = H \hat{x}.
\]

Redundant Measurement

A redundant measurement is a measurement that is not critical. Only redundant measurements may have nonzero residuals.

Critical Set

A critical set (Cset) is defined as a group of measurements (non-critical) in which the removal of any of such measurements makes the remaining of the group critical. Normalized residuals of measurements pertaining to a critical set (Cset) are equal and their correlation coefficients present maximum values. Suppose that measurements $i$ and $j$ belong to the same critical set, it follows that:

\[
\begin{align*}
    \rho_{ij} &= \frac{r_N(i)}{r_N(j)} = 1 \\
    \gamma_{ij} &= \frac{E(i,j)}{\sqrt{E(i,i)E(j,j)}} = 1
\end{align*}
\]

This article employs a method detailed in Filho, et al., 2001 to detect critical measurements and sets using equations presented.

Genetic Algorithm

A genetic algorithm (GA) stems from an analogy of the Natural selection process. The GA has the following advantages:

(i) It is expected that the GA is capable of evaluating the global minimum; the GA is based on the multi-point search and does not get stuck with local minima; and

(ii) It is not necessary that the objective function is differentiable, that is, the objective function is arbitrary.

The GA evaluates the optimal solution to maximize the objective function called the fitness. Using the genetic operators such as crossover, mutation, and reproduction the optimal solution is searched to maximize the fitness. In this article, a GA is used to determine the optimal solution for redundant measurements for static state estimation. The specified values of the load flow calculation are taken as the basic measurements, the GA was archived in step 1 for designing the primary outline of the metering system and the measurement set is assumed to contain only the conventional measurements such as, power injections and power flows.

Methodology

Step 1: Metering System Design

To reduce the number (cost) of meters and RTUs, to abate critical measurement and critical set and leverage points and to decrease condition numbers, it is necessary that the SE equation converge rapidly and avoid ill-conditioned cases type of measurement. In step 1, power injection measurements and power flow measurements are used. The random measurement error standard deviation is:

\[
\sqrt{R_i} = \frac{(0.02m + 0.005f_s)}{3}
\]

where $m = \sqrt{P_i^2 + Q_i^2}$ is the true measurement value, and $f_s$ is full scale value.
Fitness Function
To comply with requirements discussed a fitness function is proposed as:

$$FF = N\text{meas} + k_1 \times N\text{rtu}$$

$$+ k_2$$

$$+ k_3 \times N\text{lepo} + k_4 \times N\text{scr} + k_5 \times N\text{pcr}$$

where FF is the fitness function, Nmeas is the number of measurements, Nrtu is the number of RTUs, Nlepo is the number of leverage points, Nscr is the number of critical measurements, Npcr is the number of critical 2-set and k1, k2, k3, k4, k5 are constants.

Step 2: Addition of Voltage Phase Angle Measurement
The addition of a voltage phase angle measurement to a conventional state estimator could greatly increase the accuracy of the state estimator if implemented correctly. In this step, adding PMUs to the pre-designed metering system developed in step 1 will increase accuracy.

PMU Placement
PMU placement can be accomplished via several different criteria including security concerns, observability and improvement in state estimation. Here the criterion used to determine the location of PMUs will be improvements in the state estimator performance. It is possible to examine the deviation of $\hat{x}$ from the exact value of $x$. Typically this comparison is not possible, however, due to the use of test beds with a known solution, it is possible to use normalized error, NE, to assess the accuracy of $\hat{x}$ with

$$NE = \frac{\|x_{\text{exact}} - \hat{x}\|_2}{\|x_{\text{exact}}\|_2}$$

To allocate the PMUs, first, the residual vector of states $(x_{\text{exact}} - \hat{x})$ is calculated, followed by the difference of the residual vector to mean. The greatest number in this vector will be the best candidate for installation as the first PMU. By this method, the addition of PMUs to the measurement set will be continued until changes in NE are not observed. (See Milosevic & Begovic, 2003 for details.)

Results
Step 1
The meter placement problem was modeled through GAs considering a binary encoding system in which each individual (chromosome) of a population corresponds to a proposed solution for the problem (metering system). A chromosome is represented by a vector whose elements are associated with meter types and locations. The chromosome dimension then corresponds to the maximum number of meters that can be installed in a given network (twice the number of branches plus the number of buses). The chromosome elements (genes) assume binary values and will be equal to 1 if a meter is placed and equal to 0 otherwise. It is assumed that all the power measurements are in active-reactive pairs; therefore, a single gene represents a pair of measurements. (The proposed method is applied to analyze the measurement placement plan of the IEEE-14 bus power system shown in Figure 1.)

During the search procedure, different values for GA parameters (crossover probability, mutation rate, and population size) were tested. The search process stopping criterion was based on a previously defined maximum number of generations. The genetic algorithm parameters used in the Step 1 to run the search for the optimal set of measurements are as follows:

- Maximum generation = 200
- Population size = 100
- Crossover probability = 0.7
- Mutation probability = 0.01
- Constants in FF are: $k_1 = 20$, $k_3 = 100$, $k_4 = 10^4$ and $k_5 = 100$.

$$k_2 = \begin{cases} 
0 & \text{if condition number} \leq 1,000 \\
10^3 & \text{if } 1,000 < \text{condition number} < 100,000 \\
10^6 & \text{if condition number} \geq 100,000 
\end{cases}$$

The IEEE-14 bus system example with its measurement configuration shown in Figure 1 illustrates the proposed method (Step 1). Five
injections measurements are located at buses 1, 3, 7, 10, 13 and 16, and line flow measurements on lines 1-2, 1-5, 3-2, 3-4, 7-4, 6-5, 6-11, 6-12, 13-6, 7-8, 7-9, 9-10, 9-14, 10-11, 13-12 and 13-14. The evolution of the fitness for the best individual in each generation is presented in Figure 2.

Step 2
The procedure in Step 1 will be achieved to determine the optimal placement and number of added measurements (PMUs) to increase the accuracy of state estimation and to decrease the condition number of the gain matrix. In addition, it is assumed that PMUs are
added to voltage phase angle measurements. The measurement standard deviation of PMUs is assumed to be 0.002 (radian).

Table 1 shows the results of the simulation for normalized errors when PMUs are added by the method discussed herein. An improvement in state estimation accuracy was distinguished as PMUs were added to System. Figure 3 shows the condition number variation along with the PMU amount with the increase of PMU amount, the scale of the estimator and the condition number of the equation became smaller. Thus, the model can improve the numerical stability of the SE equations.

Table 1: Change in Normalized Error due to Install PMUs

<table>
<thead>
<tr>
<th>Location of PMUs</th>
<th>Normalized Error</th>
<th>Change in Normalized Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(No PMU)</td>
<td>0.0015755</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.00031086</td>
<td>-80.2691</td>
</tr>
<tr>
<td>6,12</td>
<td>0.000055957</td>
<td>-81.9993</td>
</tr>
<tr>
<td>6,11,12</td>
<td>0.000046249</td>
<td>-17.3490</td>
</tr>
<tr>
<td>6,11,12,13</td>
<td>0.000040274</td>
<td>-12.9191</td>
</tr>
<tr>
<td>6,11,12,13,14</td>
<td>0.000040091</td>
<td>-0.4544</td>
</tr>
</tbody>
</table>

Figure 3: Change in Condition Number

Conclusion
An optimization problem has been formulated to determine the optimal design principle for arranging measurements where a number of metering systems should be minimized while some performance requirements should be observed. A genetic algorithm was applied to solve the optimal meter placement problem. Test results with the IEEE 14 bus system show that the proposed methodology is capable of obtaining optimal metering systems. Further the metering system was reinforced by adding PMUs to the system designed and the simulation shows that the new model can improve accuracy, the SE equations numerical stability and the convergence speed.
References


An equivalence test is proposed which is based on the P-value of a test for a difference and the sample size. This test may be especially appropriate for an exploratory re-analysis if only a non-significant test for a difference was reported. Thus, neither a confidence interval is available, nor is there access to the raw data. The test is illustrated using two examples; for both applications the smallest equivalence range for which equivalence could be demonstrated is calculated.

Key words: Equivalence; P-value; re-analysis; reverse test.

Introduction

Two or more groups are often compared in applied research, thus begging the question: What should be done in the case of a non-significant difference between the groups? Concluding that the null hypothesis of no difference is true without any further support is not correct. Here, it is shown that an equivalence test can be performed without access to raw data if the sample size and the P-value of a test for a difference are known, and if the test statistic is at least approximately normally distributed. This allows any reader to perform a re-analysis and it is possible to determine the smallest difference for which equivalence can be established.

A procedure sometimes performed in case of a non-significant difference is a retrospective power analysis, but such a retrospective power analysis has logical flaws and shortcomings (Hoening & Heisey, 2001; Nakagawa & Foster, 2004). When the aim is to demonstrate the absence of a relevant difference it is necessary to reverse the traditional hypotheses in an equivalence test (McBride, 1999; Hoening & Heisey, 2001). The null hypothesis will then state that there is a relevant difference, whereas there is essentially no difference – that is, a negligible difference only – under the alternative. Defining the effect size Cohen’s $d$, calculated as $d = \frac{\mu_1 - \mu_2}{\sigma}$ (Cohen, 1988) where $\mu_i$ denotes the population mean of group $i$ and $\sigma$ the population standard deviation, results in

$$H_{0,\text{equiv.}}: d \leq -\theta \text{ or } d \geq \theta$$

vs.

$$H_{1,\text{equiv.}}: -\theta < d < \theta \text{ (with } \theta > 0).$$

When the appropriate confidence interval for $d$ is completely included within the equivalence range $-\theta$ to $\theta$, the equivalence test’s null hypothesis $H_{0,\text{equiv.}}$ can be rejected (Steinijans et al., 2000). Hence, the alternative $H_{1,\text{equiv.}}$ cannot be $d = \theta$ only, the entire confidence interval has to be consistent with $H_{1,\text{equiv.}}$.

Parkhurst (2001) suggested performing such an equivalence test whenever a classical test with a no-effect hypothesis has failed to yield a significant difference and he introduced the term reverse test for an equivalence test applied in this context. Parkhurst’s suggestion has not become common practice. However, reporting a confidence interval would allow a
reader to check whether the interval lies within an assumed equivalence range and therefore to judge the biological importance of a result. Unfortunately, reporting confidence intervals is also not commonplace, although it is often recommended (see Nakagawa & Foster, 2004, and references therein). By contrast, two measures are almost always given when a null hypothesis of no difference is tested: the P-value p and the sample size n. It is the aim of this article to demonstrate how an equivalence test can be carried out based on n and p only.

The Proposed Equivalence Test

It is assumed that the test statistic is at least approximately normally distributed, which is true for a wide variety of commonly applied tests. Under the null hypothesis of no difference, the one-sided P-value has a uniform distribution over the interval [0, 1] regardless of the sample size n. Under the alternative hypothesis, that is, under the assumption that there is a difference, the probability for a small P-value increases. In this case, the P-value’s distribution depends on n and d (Hung, et al., 1997).

First, consider a one-sample test with H0: \( \mu = 0 \) vs. the one-sided alternative H1: \( \mu > 0 \). If the effect size is defined as \( d = \frac{\mu}{\sigma} \), then the distribution function of the P-value p is

\[
G_d(p) = 1 - \Phi(Z_p - \sqrt{nd}),
\]

where \( \Phi \) denotes the distribution function of the standard normal distribution and \( Z_p \) the \((1 - p)\)th percentile of that distribution, i.e. \( \Phi(Z_p) = 1 - p \) (Hung, et al., 1997).

The P-value, \( p \), of the test for a difference, can be used as the test statistic for the equivalence test. The critical region of the resultant equivalence test is \([G_\alpha^{-1}(1 - \alpha), 1]\), that is, whenever \( p \) lies within this interval equivalence can be concluded. The equivalence test’s P-value is \( p_{\text{equiv.}} = 1 - G_\theta(p) \).

When two samples with \( m_1 \) and \( m_2 \) observations, respectively, are compared H0: \( \mu_1 = \mu_2 \) may be tested vs. the one-sided alternative H1: \( \mu_1 > \mu_2 \). With the effect size \( d = \frac{\mu_1 - \mu_2}{\sigma} \), and \( n = \frac{m_1 m_2}{m_1 + m_2} \) the above-mentioned formulas for the one-sample scenario can be used (Hung, et al., 1997).

The formulas discussed apply to one-tailed tests. In the case that a two-tailed P-value is reported, a one-tailed P-value of the test for a difference can be calculated because the original test statistic is assumed to be at least approximately normally distributed (George & Mudholkar, 1990).

Applications

Scantlebury, et al. (2006) investigated the energy expenditure of the Damaraland mole-rat (Cryptomys damarensis). No significant change in body mass during the experimental period was found for any category of animal and condition. Consider frequent workers during dry conditions; in that case, \( n = 21 \) and Student’s one-sample t test gives a one-tailed P-value \( p = 0.18 \).

When assuming that a moderate effect, \( d = 0.5 \), corresponds to a negligible change in body mass, the equivalence range is any effect size between \(-0.5 \) and \(0.5 \). The critical region of the resulting equivalence test with \( \alpha = 0.05 \) is \([0.259, 1]\), hence equivalence cannot be concluded in this case because \(0.18 < 0.259\). The equivalence test’s P-value is \( p_{\text{equiv.}} = 0.084 \). The equivalence test with \( \alpha = 0.05 \) could demonstrate equivalence if an effect size with an absolute value of 0.559 or smaller would be regarded as a negligible difference. Thus, 0.559 is the smallest value of \( \theta \) for which equivalence can be demonstrated.

Richdale (1957) observed yellow-eyed penguins (Megadyptes antipodes) from different colonies on New Zealand’s South Island. He compared the number of days the birds were ashore as chicks between \( m_1 = 27 \) that were subsequently seen as juveniles or later, and \( m_2 = 58 \) chicks that were not seen again. Student’s t test gives a one-tailed P-value of 0.300. Again, equivalence cannot be concluded if the range is any absolute value of the effect size smaller than a moderate effect of \( d = 0.5 \). The critical region of the resultant equivalence test with \( \alpha = 0.05 \) is \([0.308, 1]\). Here, 0.505 is the smallest value of \( \theta \) for which equivalence can be demonstrated.
Is $\theta = 0.505$ a negligible effect in this example? Richdale (1957) reported means and standard deviations: 106.4 days (± 5.1) for the chicks not seen again and 105.8 days (± 4.4) for the other group, the estimated common standard deviation is 4.89. Hence, a mean difference of approximately 4 days would be a large effect of $d = 0.8$. A mean difference of approximately 2.5 days would give an effect of $d = 0.505$ for which equivalence can be demonstrated. Compared with the observed range of 97 to 118 days (Richdale, 1957) this difference appears to be negligible.

Conclusion
For any equivalence test the equivalence range has to be specified. Several proposals describe how to choose an equivalence range (see Ng, 2001 and references therein). Here, equivalence ranges based on the effect size $d$ are used. According to Cohen (1988) $d = 0.2$ is a small effect, $d = 0.5$ a medium effect, and $d = 0.8$ a large effect. These values may be used although they depend on the variance, in particular because the equivalence test is used here with an exploratory intention. Different researchers may favour different equivalence ranges; in this case, Parkhurst (2001) recommended calculating the minimum value for which equivalence can be concluded. A SAS program to compute this value, given in the applications described herein, is available by request.

A large difference between $\mu_1$ and $\mu_2$ is possible even when the test for a difference gives a large one-tailed $P$-value. This is the case when the observed difference is in the opposite direction than specified by the one-sided alternative hypothesis; in this situation it is not useful to decide for equivalence. Therefore a conservative approach is warranted: the smaller one of the two possible one-tailed $P$-values for the equivalence test should be used. Note that this was done in the examples analysed, because the $P$-values of the test for difference were both $\leq 0.5$.

When the equivalence test is performed as a reverse test after a non-significant test for difference, a multiple test problem occurs. It may be argued that the error rates of the entire procedure are not under control. However, the procedure is proposed here as a more exploratory means to allow a reader to gain additional information. When the aim of a study is to demonstrate equivalence of two treatments in a confirmatory manner an equivalence test must be performed as the first and main analysis. In this context it should be mentioned that Parkhurst (2001) recommended the reverse test particularly for basic science.

Finally, it should be noted that the idea of an original $P$-value-based equivalence test is not entirely new. Donahue (1999) mentioned that the temptation may exist to use the $P$-value in order to test for equivalence; however, he did not consider this idea any further because other equivalence tests exist. The situation considered herein is that, for a re-analysis, there is no access to raw data and no reporting of confidence intervals, hence, the equivalence test based on $p$ and $n$ may be the only choice. However, sometimes the $P$-value is not specified. If, instead of the $P$-value, a lower limit, such as $p > 0.45$ (e.g. Brown, et al., 2005), is specified the boundary can be used rather than the unknown $p$ for the then conservative equivalence test.

References


Derivation of Mass Independent Quantum Treatment of Phenomenon

David Parker
Oak Park, Michigan

The derivation and applications is presented of a spatial variable or spatial radius which is related to the inertia or mass-energy of any quantum body by a Lorentz invariant relation. Mass independent DeBroglie and Schroedinger equations are derived and applied to the resolution of the linguistic incompatibility between quantum theory and the geometrical weak equivalence principle. The equivalence principle is restated in terms of the spatial radius. The gravitational attraction between bodies and the relativistic energy are both presented in terms of the spatial radius follows. The ratio of the gravitational force to the Coulomb force at the Planck scale and is found to equal unity. The annihilation of particle and anti-particle is presented as an interference effect using the spatial radius.

Key words: Mass independent quantum mechanics, mass independent Schroedinger equation, mass independent DeBroglie equation, spatially dependent quantum mechanics.

Introduction
Quantum mechanics contains a collection of statistical phenomenon known as the wave-particle duality which refers to the unexpected and inexplicable appearance of wave like behavior for a large collection of particles. An example of the bizarre aspect of the wave-particle duality is the appearance of an interference-like pattern after the detection of thousands of electrons which apparently did not interact with each other. The connections between the wave and particle aspects of matter follow the Einstein-DeBroglie relations. However, there is a large conceptual gap between the wave and particle dynamical variables which are utilized in the Einstein-DeBroglie relations (Silverman, 1994).

For example, the wavelength of the wave phenomena associated with matter has a natural physical meaning relative to the wave phenomena but it does not have the same kind of physical relation relative to the particle momentum or inertia. Furthermore, the wavelength is a distance which is the measure of an interval and is a spatial variable which does not directly relate to the momentum or inertial variables. In order to bridge this conceptual gap the dynamical variables of mass and energy which characterize the inertial aspects of matter have been reformulated in terms of the spatial variable which also characterizes the wave aspects of matter. The spatial variable which is introduced in this article is called the spatial radius. The spatial variable also leads to a mass independent formulation of quantum theory and a physical model of the inner structure of the wavelength associated with matter which in turn might give some understanding of the wave-like transfer of inertia by particles. The phrases quantum body or elementary particles refer to particles or bodies in quantum theory, such as photons, bosons, fermions, and their theoretical constituents such as quarks.

The spatial radius is related to the inertia or the mass-energy of any quantum body by a Lorentz invariant relation derived from a generalization of the results of classical experiments. The extension of this concept to any bodies which are not a concern of quantum theory such as astronomical bodies is a work in progress. Physical significance of the spatial radius is found in scattering theory and is discussed below. The spatial radius is also applied to the gravitational attraction between bodies at the Planck scale and the treatment of relativistic energy.
Also discussed is an exclusion which follows from the Lorentz invariant form. The spatial radius concept enables the derivation of mass independent three dimensional forms of both the DeBroglie and Schroedinger equations which in turn implies a mass independent quantum treatment of phenomenon and the resolution in principle of the linguistic incompatibility between the existing mass dependent quantum interpretation of nature and the mass independent demands of the geometrical weak equivalence principle (Greenberger & Overhauser, 1980).

Using the spatial radius, the equivalence principle can be restated as the requirement that the fall of a point particle in the gravitational field of the earth is independent of the spatial radius of the point particle. Verification of the predictions and solutions which are dependent on the spatial radius of the DeBroglie and Schroedinger equations and are derived here can be acquired by comparison with the large body of solutions which have accumulated since the introductions of the original mass-dependent DeBroglie and Schroedinger equations in the previous century.

Derivation of the Spatial Radius

The spatial radius concept was derived from a generalization of the results of classical experiments which attempted to determine the motion or the self-force of electrons in various fields and implied that the motion of the assumed spherical electrons were dependent on the ratio of the electrostatic energy and the size or radius of the assumed spherical electron with a numerical factor dependent on charge distribution which can be written as e²/r, which is proportional to the proper energy of the electron where e is the electron charge (esu) and r the radius (Born, 1969). Experimentally the electron was known to be stable, however, the model of an assumed spherical electron was eventually demonstrated by Poincare to be unstable and so the spherical model was discarded (Heitler, 1953).

Without the spherical model the question arises as to the physical meaning of the electron radius (r) upon which the electron motion was found to be partially dependent. The dependency of the motion on the term e²/r was interpreted as implying that the inertial electron mass increased proportionally to e²/r and it was also assumed by some physicists that inertial mass of the electron was purely electromagnetic in origin, which seemed possible because the electron mass was so small. This assumption about the electromagnetic origin of the inertial mass implies that the electron radius assumes the value r = e²/(m × c²) and the radius with this particular value became known as the classical electron radius (Born, 1969).

An analysis of the role of the electron charge to the motion in these experiments shows that the role of the electron charge is involved with the inertial mass of the electron. This was interpreted herein as the possibility that the idea of the electron radius could be generalized to all bodies independently as to whether or not the body was charged. Thus, the classical electron radius can be viewed as the spatial radius associated with an electron. The specific form of the generalization follows the form in the Lorentz invariant equation for the relation between the classical electron radius and the electron mass.

The classical electron radius is sometimes presented as a length which is a combination of classical constants but is several orders of magnitude smaller than the average size atom. At present the value of the classical electron radius is probably historical. Contemporary physicists might be surprised however by Max Born’s claim that the classical electron radius has an actual physical significance and the simplest phenomena in which it occurs is the scattering of light in the calculation of an ‘effective cross scattering’ (Born, 1969).

The electron which has the smallest mass can be viewed as the unit of mass, then the mass of any body and the associated spatial radius can be written as a multiple of the electron mass. An indication of the interpretation and the actual physical significance or existence of the spatial radius as a generalization of the classical electron can be found in Max Born’s analysis and calculation of the coefficient of scattering of radiation for a free body. Born points out that in the formula for the total energy scattered by a free body I(f) when given an
initial energy $I(0)$ is $I(f) = \frac{8\pi}{3}(e^2/mc^3)\times I(0)$ which implies that a proton or nucleus scatters some millions of times less than an electron (Born, 1969). The point is that the generalized form of the spatial radius is found in the formula for the energy scattered and also has a physical interpretation. The minute amount of radiation scattered by a large mass means that the phenomena would not be apparent.

The exact expression for the relation between the spatial radius and the mass-energy of the body is $\Phi = R(M) \times M$ and when given the energy, $E$, of the body then $E \times R(E) = (\Phi \times c^2)$ with the important result that $R(E) = R(M)$ for any quantum body. The expressions for $R(M)$ and $R(E)$ are Lorentz invariant. From $e^2 = \alpha \times h \times c$ the constant $\Phi$ can also be written as $\Phi = (\alpha \times h) / c$ where $\alpha$ is the fine structure constant and $h$ is Planck’s reduced constant. The discussion of relativistic energy will result in the spatial radius having either a positive or negative value ($\pm R(M)$ or $\pm R(E)$) with the negative value referring to the anti-particle. Objection to the generalization of the properties of a charged electron to uncharged bodies should be reconsidered in terms of the quark theory.

The exclusion mentioned above refers to physical variables of any quantum body which originates within the Lorentz invariant definition of the spatial radius. The exclusion is that the spatial radius and the set of all physical variables which, during a finite time interval, describe the state of the body such as the mass, energy, etc. always have non-zero finite values. The sign of each variable is the sign of the spatial radius. The exclusion of zero-valued variables requires that any change of the sign of the spatial radius must be discontinuous which is consistent with the spirit of quantum theory. An example of a transition which is clearly discontinuous is mentioned by Heitler where he points out that a rapidly varying external field can cause a transition of an electron with positive energy to a negative energy state with positive charge (positron) (Heitler,1953). The assumption that all variables assume the sign of the spatial radius implies that the mass of anti-particles becomes negative does not necessarily refer to the inertial mass of the anti-particle. However, it is believed that negative inertial mass will behave or fall exactly like positive inertial mass in a weak gravitational field.

Algebra of the Spatial Radius

Given mass $M$ and mass $N$, then

$$M \times N = \frac{\Phi^2}{R(M) \times R(N)},$$

$$M + N = \frac{\Phi \times (R(N) + R(M))}{R(M) \times R(N)}$$

and

$$\frac{M}{N} = \frac{R(N)}{R(M)}.$$

Thus, the reduced mass ($\mu$) is:

$$\mu = \frac{\Phi}{R(M) + R(N)}.$$

Methodology

Derivation of the Mass Independent and Spatially Dependent DeBroglie Equation

The mass independent form of the DeBroglie equation for the wavelength of the wave phenomena associated with any quantum body is derived by first solving for the body mass in terms of the spatial radius and substituting for the mass in the momentum. Then the DeBroglie equation becomes:

$$\lambda = \frac{h}{M \times v}$$

$$= \frac{(h \times R(M))}{(\Phi \times v)}$$

$$= \frac{(h \times R(M) \times c)}{(\alpha \times h \times v)}$$

$$= \frac{(2\pi \times R(M) \times c)}{(\alpha \times v)}$$

$$= \frac{h}{p}$$

using $\Phi = \alpha \times h / c$, where $\alpha$ is the fine structure constant and $h$ is Planck’s constant in reduced form. For the spatial radius when given
as a function of the energy $R(E)$ the wavelength has the form

$$\lambda = \left( \frac{2\pi \times R(E)}{\alpha} \right).$$

(1a)

The arrangement of the wavelength in equations 1a and 1b is interpreted here as being the fine structure of the wavelength where the quantity $1/\alpha \approx 137$ is the number of spheres of size $2\pi \times R(M)$ which comprise the wavelength and are also arranged in a line in the direction of motion of the body. Each sphere has an inertial aspect in the direction of motion of the body and the sum of the spheres equals the total inertia in the direction of the motion. The point is that the total inertia of a body is not transmitted in one whole impulse but appears to be transferred in a linearly quantized manner like a wave in the direction of motion. From this assumption it appears that the inertial aspects are also related to the fine structure constant $\alpha$. The component of the wavelength in the direction of motion varies relativistically; this is consistent with the correlation of particle aspects at high energies.

Derivation of the Wavelength from the Uncertainty Principle

The spatially dependent wavelength can be easily derived from the calculation of the product of the uncertainties in momentum and position $\Delta p \times \Delta x \geq \hbar$ with the substitution of the spatial radius for the mass in the momentum $P = M \times v$ and by assuming that the wavelength $\lambda = \Delta x$.

Derivation of Non-Relativistic Mass Independent Schroedinger Equation

The simplest derivation is obtained by substituting the spatial radius for the mass into the co-efficient of the second derivative of the wave function with respect to space in the original non-relativistic mass-dependent Schroedinger equation. The resulting spatially dependent Schroedinger implies a spatial interpretation of quantum theory.

Gravitational Attraction between Bodies

The classical gravitational attraction between bodies in terms of the spatial radius is derived with the substitution of the spatial radius for the mass of each body. As the scale of events approaches the Planck scale it is generally believed that the ratio of the gravitational force to the Coulomb force approaches unity. The ratio of the gravitational and Coulomb forces in terms of the spatial radius is given in equation 3b, which is the gravitational force between two bodies each with the Planck mass, $m_\text{P}$, separated by the distance $S$. The Coulomb force is between two unit charges separated by the distance $S$. Note that $L \times a = R(m)$ where $L$ is the Planck length and $R(m)$ is the spatial radius of the Planck mass. It might be possible with the use of the spatial radius to smooth out the violent fluctuations at the Planck length. The gravitational force between two bodies with mass $M$ and $N$ is:

$$F = \frac{-[\Phi^2 \times G]}{[R(M) \times R(N) \times S]^2}$$

(3a)

$$\frac{F(\text{GRAV})}{F(\text{COUL})} = \frac{[G \times \Phi^2 \times \alpha \times \kappa]}{[R^2(m) \times S^2]} \kappa \times e^2 / S^2$$

(3b)

where $R(m)$ is the spatial radius of the Planck mass and $\kappa$ is the constant of proportionality in Coulomb’s law.

$$F(\text{GRAV}) \times [(\alpha \times \kappa)] = F(\text{COUL})$$

or

$$F(\text{GRAV}) = (\hbar \times c) / S^2$$

(3c)
Relativistic Energy and the Spatial Radius

If the relativistic energy in terms of the spatial radius is calculated for a body with a given rest mass $m(o)$ and velocity there will be a negative and a positive valued solution. The negative value is the energy of the anti-particle.

$$E(+) = \Phi \times c^2 / R (m(o))$$
$$= c^2 / R (m(o)) \quad (4a)$$

$$E(-) = -\Phi \times c^2 / R (m(o))$$
$$= -c^2 / R (m(o)) \quad (4b)$$

Interference Effects and Annihilation of Particles

When particle and anti-particle interact and mutually annihilate they generally form a pair or trio of photons. The assumption here is that the annihilation is caused by the destructive interference of the out of phase wave phenomenon associated with the wave aspects of particle and anti-particle which are related to the spatial radius.

The sign of the spatial radius refers to the in or out aspect of the particle's wave phenomena. Interference phenomenon requires a dimensional representation for the superposition of the waves, which is regarded as motivation for the assumption of three dimensionality. Interference effects are also assumed to have a role in the exclusion principle in the repositioning of the orbital electrons.

The Relation of the Planck Scale to All Quantum Bodies

The Planck scale is a set of physical variables formed from various combinations of the gravitational constant, Planck's constant (reduced) and light speed. Given any particle $P$ with finite non-zero mass $M$ and radius $R(M)$ it follows that $M \times R(M) = m(e) \times R(e) = \Phi$ where $m(e)$ is the electron mass and $R(e)$ is the electron radius. The product of the fine structure constant, the Planck mass and the Planck length is equal to $[m(e) \times L] \times \alpha = \Phi$. Then the spatial radius concept implies that the Planck mass and the Planck length are related to the mass and radius of any quantum body. It is also implied that the Planck energy has a relation to the energy all of the quantum bodies.

Conclusion and Brief Model of Quantum Bodies

All quantum bodies are assumed to exist as three dimensional structures during their lifetime. The axis of any quantum particle in the instantaneous direction of motion is called the relativistic axis. The axes normal to the relativistic axes are called the virtual axes. The interactions that occur on the relativistic axis are generally time dependent and the interactions on the virtual axes are generally independent of time. Interference effects and repositioning is assumed to occur along the virtual axes. Entanglement phenomena, which occur with particles which travel at light speed, are also assumed to occur along the virtual axes. A description of the interactions that occur along the virtual axes is assumed to be given by the Feynman path integral interpretation, which can be considered time independent at each point of a trajectory.

The question of time dependence might be tested by a variation of the experiment where the emitter is shut down after a single electron emission and allowed to cool and then emitting the second electron and cooling off and so on. The spatial radius always has a non-zero value on all three axes of the body. The inertial variables of the body are distributed along the spatial radius and therefore vary statistically as a function of the statistical variation of the spatial radius. The shape of the particle at any particular time is the three dimensional envelope formed from the tips of the spatial radii. In a minimal sense the spatial radius is a fresh approach and is a unique interpretation of quantum theory.

The application of a radius of a particle was suggested by DeBroglie as a way to avoid the pitfalls of infinite self-energies which occur with point particles (DeBroglie, 1960). Verification of the spatial radius might be obtained from scattering experiments. It may be both interesting and useful to repeat the classical experiments which resulted in size dependent behavior for an electron. Time is regarded here as a secondary variable which is dependent on motion.
Acknowledgements
For my parents and Lawrence Resnick. The author also thanks Grant R. Gerhart and Richard Felch.

References
The two-point mixture index of fit enjoys some desirable features in model fit assessment and model selection, however, a need exists for efficient computational strategies. Applying an NLP algorithm, a program using the SAS matrix language is presented to estimate the two-point index of fit for two-class LCA models with dichotomous response variables. The program offers a tool to compute $\pi^*$ for two-class models and it also provides an alternative program for conducting latent class analysis with SAS. This study builds a foundation for further research on computational approaches for M-class models.

Key words: Pi-star, two-class LCA models, SAS.

Introduction

The two-point mixture index of fit, $\pi^*$, was introduced to address the issue of model fit for frequency data in two-way contingency tables (Rudas, et al., 1994; Xi, 1994; Clogg, et al., 1995; Xi & Lindsay, 1996). This index has been extended to a variety of other theoretical models. For example, Rudas & Zwick (1997) discussed the use of $\pi^*$ in differential item functioning, Rudas (1999) studied applications of $\pi^*$ with regression models involving continuous variables and Dayton (1999; 2003) extended the application of $\pi^*$ to latent class models.

For a two-point mixture, $P = (1 - \pi)\Phi + \pi\Psi$, let $\Phi$ denote the probability distribution of some hypothesized frequency model, $H$, let $\Psi$ represent an unspecified probability distribution, and let $\pi$ indicate the proportion of the population that is not intrinsically described by model $H$. Then, the mathematical model for $\pi^*$ can be written as (Rudas, et al., 1994):

$$\pi^* = \inf\{\pi \mid P = (1 - \pi)\Phi + \pi\Psi, \Phi \in H\} \tag{1}$$

In effect, $\pi^*$ is defined as the smallest value of $\pi$ for which $P$ remains true for model $H$ and can be viewed as "a measure of the proportion of the population measured with error" (Rudas, et al., 1994, p. 628) or as a measure of lack of fit (Rudas, et al., 1994; Xi, 1994; Xi & Lindsay, 1996). In practice, the minimum proportion of cases that must be removed from the frequency table is compared to the remaining cases in order to provide perfect fit for $H$ (Dayton, 2003).

As opposed to conventional approaches, such as the $G^2$ likelihood ratio test and various information criteria such as AIC, $\pi^*$ represents a new perspective with respect to model-fit assessment and provides an easy-to-interpret alternative basis for model comparison and selection. Rudas, et al. (1994) summarized the desirable properties of this new index as: (1) unique; (2) defined on the 0, 1 interval; (3) decreasing in magnitude for increasingly more complex models when comparing nested models; and (4) invariant to multiplicative transformation of the frequency data. This latter
Property is particularly interesting because it means that the magnitude of \( \pi^* \) is not dependent on sample (although its sampling error is).

**Application of \( \pi^* \) to Latent Class Models**

A latent class model with \( T \) classes is, from a mathematical point of view, a finite mixture of product-multinomial probability functions. Considering a four-variable model as an example, the unconditional probability for the response vector, \( Y = \{i j k l\} \), can be defined as:

\[
P(Y) = \sum_{t=1}^{T} \tau_t P(Y | t) \tag{2}
\]

where \( \tau_t \) is the proportion in latent class \( t \), and \( P(Y | t) \) is the product of the conditional response probabilities for the four variables corresponding to the response pattern \( \{i j k l\} \), given membership in latent class \( t \). The latent class model is subject to the restrictions that: (1) the latent class proportions sum to 1; (2) the conditional response probabilities, given latent class membership, sum to 1 for each variable; (3) the variables are conditionally independent within any given class (Lazarsfeld & Henry, 1968; Goodman, 1974; Dayton, 1999; among others).

In latent class analysis, Chi-square goodness-of-fit tests and information criteria are widely applied procedures for assessing model fit and for model selection. These methods are open to the criticisms that: (1) with small sample size or sparse data, the statistics do not asymptotically follow appropriate \( \chi^2 \) distributions; and (2) with large sample size, it is highly likely that the null hypothesis will be rejected for relatively trivial effects. Therefore, Chi-square tests may not be appropriate for model selection under those circumstances. For information criteria, such as \( AIC \), it is not clear how much the effect of sample size persists when the penalty term is applied. In addition, information criteria cannot be used to assess model fit in an absolute sense insomuch as interpretation of magnitudes of information criteria per se is difficult (Rudas, et al., 1994).

For the \( s^{th} \) response vector, the latent class model can be incorporated into the two-point mixture model as follows (Dayton, 2003):

\[
P_s = (1 - \pi) P(y_s) + \pi \psi_s \tag{3}
\]

where \( P(y_s) \) represents the probability distribution for the \( s^{th} \) response vector or response pattern. \( \pi^* \) is obtained as the minimum value of \( \pi \) when the model holds true across all response vectors (Dayton, 2003). The definition of \( \pi^* \) circumvents the drawbacks of Chi-square statistics, thus, the index enjoys some unique advantages in model selection.

**Methodology**

**Computational Approach**

Programs for LCA such as LEM or SAS PROC LCA (Lanza, et al., 2007) do not provide options for computing \( \pi^* \). However, \( \pi^* \) can be estimated using the iterative procedures proposed by Rudas, et al. (1994) and with MLE or nonlinear programming (NLP) algorithms (Xi, 1994; Xi & Lindsay, 1996). Dayton (2003) discusses computational strategies for the fit index applied to latent class and IRT (Rasch) models and presents examples using Microsoft Excel Solver, a program that is based on a NLP algorithm. For latent class models, Dayton (2003) detailed a computational strategy in two stages: in the first stage, the NLP parameters are defined as \( a_{it}, b_{jt}, c_{kt}, d_{lt} \), etc. such that \( \tilde{\pi}_s = a_{it} \times b_{jt} \times c_{kt} \times d_{lt} \times \cdots \). Given the nonlinear constraint that the total expected frequency is equal to the total observed frequency, conventional MLEs of the parameters for an unrestricted latent class model can be estimated by minimizing \( G^2 \) as the objective function.

In the second stage, more nonlinear constraints, which specify the relationship between the expected frequency and the observed frequency for each response vector, are applied in NLP. The objective function is then redefined as maximizing the total expected frequency (or, equivalently, minimizing \( \pi \), which is a function of the expected frequencies). After convergence to some preset criterion, an estimate of \( \pi^* \) is obtained (Dayton, 2003).
Technically, simply applying the second stage alone generates an estimate of $\pi^*$. However, an associated problem, which is increasingly crucial when the number of parameters increases, is the selection of start values because good start values are critical to computational efficiency and accuracy. With inappropriate start values, the optimization procedure may fail to converge, may converge at a local optimum, or may encounter other unexpected difficulties.

Although it is possible to provide different sets of start values and to examine the results in a single stage, a more efficient approach is to first conduct a conventional unrestricted LCA analysis and then start from the resultant parameter estimates which are, in general, closer to the final NLP estimates than arbitrarily selected start values. Although start values still need to be selected for first stage optimization, one benefit of the two-stage approach is that the closer estimates of the parameters are secured with only one (not multiple) constraints, no matter how many parameters are in the model. Hence, in the second stage, computational efficiency is achieved with faster convergence since the number of NLP function calls is greatly reduced.

Two SAS NLP subroutines, NLPNMS and NLPQN, are available to implement nonlinear constraints. The NLPQN subroutine applies quasi-Newton optimization technique that involves computing first-order partial derivatives in the gradient vector or the Jacobian matrix. It is suitable for medium to moderately large problems (NLPQN, SAS 9.1 Documentation, 2007) that contain relatively large numbers of parameters; NLPNMS is suitable for smaller problems. For nonlinearly constrained optimization, the NLPQN subroutine applies a modification of Powell’s (1978, 1982) Variable Metric Constrained WatchDog algorithm (NLPQN call, SAS 9.1 Documentation, 2007). PI-LCA implements the NLPQN subroutine for optimization.

SAS Program Description

As the computation involves relatively complex matrix operations, the current version of the procedure is restricted to two-class LCA models with dichotomous response variables. The SAS program, PI-LCA, is designed to compute $\pi^*$ for models for varying numbers of variables. However, for large numbers of variables computational time may become excessive. Factors influencing the number of function calls include selection of start values, number of parameters, and data structure, such as the number of zero-frequency vectors.

The SAS program PI-LCA has four sections:

1. Macro variables. Specifically, the following quantities are labeled and input as macro variables:
   a. Number of dichotomous variables;
   b. Number of latent classes (set at 2 in current version of program);
   c. Observed sample size;
   d. Start values for the first stage optimization;
   e. Input data file name and location.

In this area of the program, the user must make adjustments in accordance with the data under consideration.

2. Data input for computing the expected frequencies. The data file can be any format (such as ASCII) that is acceptable to SAS. As the NLP procedure involves nonlinear constraints with regard to each response vector, aggregated data by the response pattern must be used as input. Assume that the number of items is numvar (as suggested previously), there should be numvar+1 fields in the dataset, with the first numvar fields representing the response patterns (e.g., 1 1 1 1 for 4 items) - the last field being the observed frequency. For ASCII data input, such as the text data generated by Microsoft Notepad, the fields should be space delimited; for example: 1 1 1 1 freq. For each observation (response pattern), the first numvar fields can either be 1’s and 2’s or 0’s and 1’s (see Table 1).

3. The first stage of the optimization procedure. This stage computes...
conventional, unrestricted two-class LCA parameters using the NLPQN algorithm. The objective function that is minimized is $G^2$, given the constraint that the total observed frequency and the total expected frequency are equal. In addition, boundary constraints are applied to ensure that all parameter estimates are non-negative. Because start values are randomly selected in this stage, detailed NLP options (items 4-8 in the option vector for NLPQN, which may vary from case to case) are specified to obtain accurate estimates. The options may increase the number of function calls and make the convergence slower, especially when there are large numbers of parameters. When the procedure converges, the start values for the second stage are obtained. It is suggested that distinct sets of start values for the first stage should be tried to ensure that a global optimum has been obtained.

4. The second stage of the optimization procedure. In this stage, both the objective function and nonlinear constraints are redefined. In most cases, the convergence is relatively fast as the start values are close to optimum. In general, items 4-8 in the option vector for NLPQN do not need to be changed from default values. At convergence, the estimate of $\pi^*$ is obtained.

Results: Exemplary Data
Example 1: Academic Cheating Data (Four Items)

Dayton (2003) used Microsoft ExcelSolver to compute $\pi^*$ for a two-class LCA model with frequency data for four dichotomous (2 = yes, have engaged in this cheating behavior, and 1 = no, have not engaged in this cheating behavior) items from a survey concerned with academic cheating behavior by college students (see Table 1).

<table>
<thead>
<tr>
<th>Item</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A B C D</td>
</tr>
<tr>
<td>1</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>1</td>
<td>1 1 1 2</td>
</tr>
<tr>
<td>1</td>
<td>1 2 1 2</td>
</tr>
<tr>
<td>1</td>
<td>1 2 1 1</td>
</tr>
<tr>
<td>1</td>
<td>1 2 1 2</td>
</tr>
<tr>
<td>1</td>
<td>1 2 1 1</td>
</tr>
<tr>
<td>1</td>
<td>2 2 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1 2</td>
</tr>
<tr>
<td>2</td>
<td>1 2 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 2 1 2</td>
</tr>
<tr>
<td>2</td>
<td>1 2 1 1</td>
</tr>
<tr>
<td>2</td>
<td>1 1 2 2</td>
</tr>
<tr>
<td>2</td>
<td>2 2 1 1</td>
</tr>
<tr>
<td>2</td>
<td>2 2 2 2</td>
</tr>
</tbody>
</table>

Table 1: Academic Cheating Data

Input to Section (1) of the SAS Program

***********************************************************************
PI-LCA: A SAS PROGRAM COMPUTING THE TWO-POINT MIXTURE INDEX OF FIT FOR TWO-CLASS LCA MODELS WITH DICHOTOMOUS VARIABLES
***************************************************************;

* SECTION 1: PROVIDE VALUES FOR FOLLOWING 5 MACRO VARIABLES;

%let numvar=4;     * NUMBER OF ITEMS (MANIFEST VARIABLES);
%let numcl=2;      * NUMBER OF CLASSES;
%let numsap=319;    * NUMBER OF SAMPLE SIZE;
%let start=1;      * START VALUES FOR THE FIRST STAGE OPTIMIZATION;
%let datafile = "C:\cheat4.txt";  * LOCATION OF THE INPUT DATA FILE;
***************************************************************;
PI-LCA SAS PROGRAM FOR TWO-CLASS LCA DICHOTOMOUS VARIABLE MODELS

Selected Output: SAS output 1- Call NLPQN Subroutine in the First Stage to Conduct Latent Class Analysis

Computing Pi-star, the Two-Point Mixture Fit Index

The First Stage: Latent Class Analysis
The Objective Function Is to Minimize G-square

Optimization Results

<table>
<thead>
<tr>
<th>Iterations</th>
<th>98</th>
<th>Function Calls</th>
<th>107</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Calls</td>
<td>100</td>
<td>Active Constraints</td>
<td>1</td>
</tr>
<tr>
<td>Objective Function</td>
<td>3.8821212398</td>
<td>Maximum Constraint</td>
<td>1.4590216E-7</td>
</tr>
<tr>
<td>Maximum Projected Gradient</td>
<td>0.0009804183</td>
<td>Value Lagrange Function</td>
<td>3.8821210939</td>
</tr>
<tr>
<td>Maximum Gradient of the Lagran Func</td>
<td>0.0008997954</td>
<td>Slope of Search Direction</td>
<td>-3.02492E-7</td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Gradient Estimate</th>
<th>Gradient Function</th>
<th>Objective Function</th>
<th>Lagrange Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>8.023430</td>
<td>-32.820114</td>
<td>0.000030386</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>4.030813</td>
<td>-64.491314</td>
<td>0.000059027</td>
<td></td>
</tr>
<tr>
<td>X3</td>
<td>2.585951</td>
<td>-99.709946</td>
<td>0.000146</td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>2.407726</td>
<td>-96.980825</td>
<td>0.000146</td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0.135555</td>
<td>-32.819815</td>
<td>0.000329</td>
<td></td>
</tr>
<tr>
<td>X6</td>
<td>0.121356</td>
<td>-64.490988</td>
<td>0.000386</td>
<td></td>
</tr>
<tr>
<td>X7</td>
<td>0.099626</td>
<td>-99.710992</td>
<td>-0.000900</td>
<td></td>
</tr>
<tr>
<td>X8</td>
<td>0.535523</td>
<td>-96.980898</td>
<td>-0.000119</td>
<td></td>
</tr>
<tr>
<td>X9</td>
<td>1.361351</td>
<td>-15.918088</td>
<td>0.000058771</td>
<td></td>
</tr>
<tr>
<td>X10</td>
<td>1.203668</td>
<td>-17.486176</td>
<td>0.00010131</td>
<td></td>
</tr>
<tr>
<td>X11</td>
<td>1.237312</td>
<td>-32.453413</td>
<td>0.000193</td>
<td></td>
</tr>
<tr>
<td>X12</td>
<td>2.147240</td>
<td>-14.876575</td>
<td>0.000191</td>
<td></td>
</tr>
<tr>
<td>X13</td>
<td>1.856461</td>
<td>-15.917976</td>
<td>0.000171</td>
<td></td>
</tr>
<tr>
<td>X14</td>
<td>1.725583</td>
<td>-17.486067</td>
<td>0.000181</td>
<td></td>
</tr>
<tr>
<td>X15</td>
<td>0.340991</td>
<td>-32.453140</td>
<td>0.000466</td>
<td></td>
</tr>
<tr>
<td>X16</td>
<td>1.295821</td>
<td>-14.876776</td>
<td>-0.00010208</td>
<td></td>
</tr>
</tbody>
</table>

Value of Objective Function = 3.8821212398
Value of Lagrange Function = 3.8821210939

Latent Class Analysis

<table>
<thead>
<tr>
<th>Observed Frequency</th>
<th>Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>207</td>
<td>205.71667</td>
</tr>
<tr>
<td>46</td>
<td>47.414163</td>
</tr>
<tr>
<td>7</td>
<td>8.9574477</td>
</tr>
<tr>
<td>5</td>
<td>2.4494936</td>
</tr>
<tr>
<td>13</td>
<td>12.303603</td>
</tr>
<tr>
<td>4</td>
<td>5.1148321</td>
</tr>
<tr>
<td>1</td>
<td>1.9535633</td>
</tr>
<tr>
<td>2</td>
<td>1.0893932</td>
</tr>
<tr>
<td>10</td>
<td>9.3388155</td>
</tr>
<tr>
<td>3</td>
<td>4.3394233</td>
</tr>
<tr>
<td>1</td>
<td>1.7671869</td>
</tr>
<tr>
<td>2</td>
<td>1.0165214</td>
</tr>
<tr>
<td>11</td>
<td>8.6134428</td>
</tr>
<tr>
<td>4</td>
<td>5.1590278</td>
</tr>
<tr>
<td>1</td>
<td>2.3494963</td>
</tr>
<tr>
<td>2</td>
<td>1.4163751</td>
</tr>
</tbody>
</table>

Total Expected Frequency | 319

<table>
<thead>
<tr>
<th>LC1 Proportion</th>
<th>LC2 Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.839431</td>
<td>0.160569</td>
</tr>
</tbody>
</table>

Conditional Probabilities

<table>
<thead>
<tr>
<th>CP Positive Response (1)</th>
<th>0.9833050 0.4230674</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.970773 0.4190132</td>
</tr>
<tr>
<td>3</td>
<td>0.9629034 0.783951</td>
</tr>
<tr>
<td>4</td>
<td>0.8180504 0.6236428</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CP Negative Response (2)</th>
<th>0.0166142 0.5769326</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.829227 0.5890868</td>
</tr>
<tr>
<td>3</td>
<td>0.0370966 0.216049</td>
</tr>
<tr>
<td>4</td>
<td>0.1819496 0.3763572</td>
</tr>
</tbody>
</table>
SAS Output 2: Call NLPQN Subroutine in the Second Stage to Compute $\pi^*$

Computing $\pi^*$, the Two-Point Mixture Fit Index

The Second Stage: Pi Optimization

The Objective Function Is to Maximize the Total Expected Frequency

<table>
<thead>
<tr>
<th>Optimization Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterations</td>
</tr>
<tr>
<td>Function Calls</td>
</tr>
<tr>
<td>Gradient Calls</td>
</tr>
<tr>
<td>Active Constraints</td>
</tr>
<tr>
<td>Objective Function</td>
</tr>
<tr>
<td>Maximum Constraint</td>
</tr>
<tr>
<td>Violation</td>
</tr>
<tr>
<td>Maximum Projected Gradient</td>
</tr>
<tr>
<td>Value Lagrange Function</td>
</tr>
<tr>
<td>Maximum Gradient of the Lagran Func</td>
</tr>
<tr>
<td>Slope of Search Direction</td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>N Parameter</th>
<th>Gradient Objective Function Estimate</th>
<th>Gradient Objective Function Lagrange Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 X1</td>
<td>8.033189</td>
<td>32.624665</td>
</tr>
<tr>
<td>2 X2</td>
<td>4.037358</td>
<td>64.709589</td>
</tr>
<tr>
<td>3 X3</td>
<td>2.592207</td>
<td>100.059213</td>
</tr>
<tr>
<td>4 X4</td>
<td>2.414567</td>
<td>90.971362</td>
</tr>
<tr>
<td>5 X5</td>
<td>0.176867</td>
<td>32.624664</td>
</tr>
<tr>
<td>6 X6</td>
<td>0.101910</td>
<td>64.709588</td>
</tr>
<tr>
<td>7 X7</td>
<td>0.084711</td>
<td>100.059213</td>
</tr>
<tr>
<td>8 X8</td>
<td>0.529770</td>
<td>90.971362</td>
</tr>
<tr>
<td>9 X9</td>
<td>1.324498</td>
<td>13.361448</td>
</tr>
<tr>
<td>10 X10</td>
<td>0.937841</td>
<td>15.143763</td>
</tr>
<tr>
<td>11 X11</td>
<td>1.350402</td>
<td>28.603220</td>
</tr>
<tr>
<td>12 X12</td>
<td>2.385143</td>
<td>12.948459</td>
</tr>
<tr>
<td>13 X13</td>
<td>1.838894</td>
<td>13.361447</td>
</tr>
<tr>
<td>14 X14</td>
<td>1.846183</td>
<td>15.143764</td>
</tr>
<tr>
<td>15 X15</td>
<td>0.123579</td>
<td>28.603222</td>
</tr>
<tr>
<td>16 X16</td>
<td>0.870889</td>
<td>12.948459</td>
</tr>
</tbody>
</table>

Value of Objective Function = 310.01091238
Value of Lagrange Function = 310.01091224

Pi-Star Results

<table>
<thead>
<tr>
<th>Observed Frequency Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>207 207</td>
</tr>
<tr>
<td>46  46</td>
</tr>
<tr>
<td>7   7</td>
</tr>
<tr>
<td>5   1.5891942</td>
</tr>
<tr>
<td>13  13</td>
</tr>
<tr>
<td>4   4</td>
</tr>
<tr>
<td>1   0.8881988</td>
</tr>
<tr>
<td>2   0.2999071</td>
</tr>
<tr>
<td>10  10</td>
</tr>
<tr>
<td>3   3</td>
</tr>
<tr>
<td>1   0.6521739</td>
</tr>
<tr>
<td>2   0.2168445</td>
</tr>
<tr>
<td>11  11</td>
</tr>
<tr>
<td>4   4</td>
</tr>
<tr>
<td>1   1</td>
</tr>
<tr>
<td>2   0.3645937</td>
</tr>
</tbody>
</table>

Total Expected Frequency: 310.01091
Pi-Star: 0.028179

LC1 Proportion LC2 Proportion
0.8640029 0.1359971

Conditional Probabilities

<table>
<thead>
<tr>
<th>CP Positive Response (1)</th>
<th>1 0.9784573 0.419757</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 0.9753798 0.336851</td>
</tr>
<tr>
<td></td>
<td>3 0.9683548 0.9161599</td>
</tr>
<tr>
<td></td>
<td>4 0.8200717 0.7325305</td>
</tr>
<tr>
<td>CP Negative Response (2)</td>
<td>1 0.0215427 0.580243</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td></td>
<td>2 0.0246202 0.6631349</td>
</tr>
<tr>
<td></td>
<td>3 0.0316452 0.0883401</td>
</tr>
<tr>
<td></td>
<td>4 0.1799283 0.2674695</td>
</tr>
</tbody>
</table>
In this example, the start values for all the parameters are set equal to 1. In general, distinct sets of start values should be employed to ensure a global maximum. In this stage, there are 98 iterations and 107 function calls. The maximum constraint violation is in the range of 1E-6, which is acceptable. The objective function \( G^2 \) is minimized at 3.88. With the NLP parameters, the latent class proportions and the conditional probabilities (CP) for the LCA model are computed.

The start values are imported from the first stage output. The objective function is redefined as maximizing the total expected frequency, which converges at 310.01 (in contrast to the total observed frequency of 319). There are only 6 iterations and 8 function calls prior to convergence (compared to 98 and 107 in the first stage). The estimated value of \( \pi^* \) converges at 0.028. Thus, only 2.8% of the cases in the population are estimated as not described by the two-class model; this suggests adequate model-data fit. (See SAS Output 2.)

Example 2: Drug Use Data (Five Items)

Five dichotomous (2 = yes, have used this drug and 1 = no, have not used this drug) items in the drug use data set with a large number of zero frequencies (see Table 2). Following the approach of Clogg, et al. (1991) in applying flattening constants to deal with the sparse data that do not support conventional maximum likelihood analysis, zero frequencies are replaced with 0.5, which enables the NLP optimization to converge. This increased the total frequency from 7,224 to 7,233.

<table>
<thead>
<tr>
<th>Item</th>
<th>Frequency</th>
<th>Item</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Data continues in next table

| Totals |    | 7224 | 7233 |
PI-LCA: A SAS PROGRAM COMPUTING THE TWO-POINT MIXTURE INDEX OF FIT FOR TWO-CLASS LCA MODELS WITH DICOTOMOUS VARIABLES

* SECTION 1: PROVIDE VALUES FOR FOLLOWING 5 MACRO VARIABLES;

%let numvar=5;    * NUMBER OF ITEMS (MANIFEST VARIABLES);
%let numcl=2;    * NUMBER OF CLASSES;
%let numsap=7224;   * NUMBER OF SAMPLE SIZE;
%let start=1.2; * START VALUES FOR THE FIRST STAGE OPTIMIZATION;
%let datafile = "c:\druguse.txt";  * LOCATION OF THE INPUT DATA FILE;

SAS output 1- Call NLPQN Subroutine in the First Stage to Conduct Latent Class Analysis

Computing Pi-star, the Two-Point Mixture Fit Index
The First Stage: Latent Class Analysis
The Objective Function Is to Minimize G-square

Optimization Results

| Iterations | 267 | Function Calls | 326 |
| Gradient Calls | 269 | Active Constraints | 1 |
| Objective Function | 469.21431307 | Maximum Constraint | 4.316415E-8 |
| Maximum Projected Gradient | 3.7700252447 | Value Lagrange Function | 469.21431303 |
| Maximum Gradient of the Lagran Func | 2.8297208894 | Slope of Search Direction | -9.946578E-8 |

Latent Class Analysis

Total Expected Frequency
7233

LC1 Proportion LC2 Proportion
0.6394273 0.3605727

Conditional Probabilities

| CP Positive Response (1) | 1 0.2155591 0.0027008 |
| 2 0.3561529 0.0612726 |
| 3 0.9700971 0.0074087 |
| 4 0.9981559 0.6917313 |
| 5 0.9994005 0.9550512 |

| CP Negative Response (2) | 1 0.7844409 0.9972992 |
| 2 0.6438471 0.9387274 |
| 3 0.8299029 0.9925913 |
| 4 0.3082687 0.0005995 |
| 5 0.0449488 0.0005995 |
A vector of start values equal to 1.2 provides better start values than 1’s as used in the first example, although the NLP call required comparatively more iterations before convergence. In the first stage, the objective function converges at 469.21. In the second stage, the value of the maximized objective function is 6,444.75 (total expected frequency), which corresponds to a $\pi^*$ value of 0.108. The result suggests that in order to provide perfect fit for the two-class model, about 11% of the cases in the population are not described by the model $H$.

**Example 3: Abortion Data (Six Items)**

The 6-item General Social Survey (GSS) abortion attitude data set (1=Yes, approve abortion for this reason, and 2=No, do not approve abortion for this reason), was collected between 1972 and 1998 and analyzed by Dayton (2006). As shown in Table 3, the total sample size is 27,151. Because there is a zero frequency for the response vector \{212121\}, it is replaced with .5 as was done in Example 2. The matrix combining the parameters and response patterns is 64x12, which requires relatively a long computational time.
Table 3: Abortion Data

<table>
<thead>
<tr>
<th>Item</th>
<th>Frequency</th>
<th>Item</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F</td>
<td>10728</td>
<td>A B C D E F</td>
<td>61</td>
</tr>
<tr>
<td>1 1 1 1 1 1</td>
<td>732</td>
<td>2 1 1 1 1 1</td>
<td>24</td>
</tr>
<tr>
<td>1 1 1 1 2 1</td>
<td>12</td>
<td>2 1 1 1 2 1</td>
<td>2</td>
</tr>
<tr>
<td>1 1 1 1 2 2</td>
<td>24</td>
<td>2 1 1 1 2 2</td>
<td>6</td>
</tr>
<tr>
<td>1 1 1 2 1 1</td>
<td>413</td>
<td>2 1 1 2 1 1</td>
<td>7</td>
</tr>
<tr>
<td>1 1 1 2 1 2</td>
<td>503</td>
<td>2 1 1 2 1 2</td>
<td>25</td>
</tr>
<tr>
<td>1 1 1 2 2 1</td>
<td>7</td>
<td>2 1 2 2 1 2</td>
<td>5</td>
</tr>
<tr>
<td>1 1 1 2 2 2</td>
<td>53</td>
<td>2 1 2 2 2 2</td>
<td>11</td>
</tr>
<tr>
<td>1 1 2 1 1 1</td>
<td>29</td>
<td>2 1 2 1 1 1</td>
<td>15</td>
</tr>
<tr>
<td>1 1 2 1 1 1 2</td>
<td>11</td>
<td>2 1 2 1 1 1 2</td>
<td>7</td>
</tr>
<tr>
<td>1 1 2 1 2 1 1</td>
<td>1</td>
<td>2 1 2 1 2 1 1</td>
<td>0</td>
</tr>
<tr>
<td>1 1 2 1 2 2 1</td>
<td>1</td>
<td>2 1 2 1 2 2 1</td>
<td>9</td>
</tr>
<tr>
<td>1 1 2 2 1 1 1</td>
<td>7</td>
<td>2 1 2 2 1 1 1</td>
<td>6</td>
</tr>
<tr>
<td>1 1 2 2 1 2 1</td>
<td>9</td>
<td>2 1 2 2 1 2 1</td>
<td>7</td>
</tr>
<tr>
<td>1 1 2 2 2 1 1</td>
<td>4</td>
<td>2 1 2 2 2 1 1</td>
<td>2</td>
</tr>
<tr>
<td>1 1 2 2 2 2 1</td>
<td>3</td>
<td>2 1 2 2 2 2 1</td>
<td>12</td>
</tr>
<tr>
<td>1 2 1 1 1 1 1</td>
<td>774</td>
<td>2 2 1 1 1 1 1</td>
<td>48</td>
</tr>
<tr>
<td>1 2 1 1 1 1 2</td>
<td>1059</td>
<td>2 2 1 1 1 1 2</td>
<td>91</td>
</tr>
<tr>
<td>1 2 1 1 2 1 1</td>
<td>18</td>
<td>2 2 1 1 2 1 1</td>
<td>4</td>
</tr>
<tr>
<td>1 2 1 1 2 2</td>
<td>60</td>
<td>2 2 1 1 2 2</td>
<td>34</td>
</tr>
<tr>
<td>1 2 1 2 1 1</td>
<td>641</td>
<td>2 2 1 2 1 1</td>
<td>46</td>
</tr>
<tr>
<td>1 2 1 2 1 2</td>
<td>5643</td>
<td>2 2 1 2 1 2</td>
<td>1100</td>
</tr>
<tr>
<td>1 2 1 2 2 1</td>
<td>21</td>
<td>2 2 1 2 2 1</td>
<td>3</td>
</tr>
<tr>
<td>1 2 1 2 2 2</td>
<td>1181</td>
<td>2 2 1 2 2 2</td>
<td>1040</td>
</tr>
<tr>
<td>1 2 2 1 1 1 1</td>
<td>7</td>
<td>2 2 2 1 1 1 1</td>
<td>6</td>
</tr>
<tr>
<td>1 2 2 1 1 1 2</td>
<td>14</td>
<td>2 2 2 1 1 1 2</td>
<td>8</td>
</tr>
<tr>
<td>1 2 2 1 2 1 1</td>
<td>1</td>
<td>2 2 2 1 2 1 1</td>
<td>3</td>
</tr>
<tr>
<td>1 2 2 1 2 2</td>
<td>3</td>
<td>2 2 2 1 2 2</td>
<td>6</td>
</tr>
<tr>
<td>1 2 2 2 1 1</td>
<td>10</td>
<td>2 2 2 2 1 1</td>
<td>8</td>
</tr>
<tr>
<td>1 2 2 2 1 2</td>
<td>153</td>
<td>2 2 2 2 1 2</td>
<td>264</td>
</tr>
<tr>
<td>1 2 2 2 2 1</td>
<td>2</td>
<td>2 2 2 2 2 1</td>
<td>1</td>
</tr>
<tr>
<td>1 2 2 2 2 2</td>
<td>121</td>
<td>2 2 2 2 2 2</td>
<td>2045</td>
</tr>
</tbody>
</table>

Total: 27,151
Input to Section (1)
*************************************************************
PI-LCA: A SAS PROGRAM COMPUTING THE TWO-POINT MIXTURE INDEX OF FIT FOR TWO-CLASS LCA MODELS WITH DICHOTOMOUS VARIABLES
*************************************************************;

* SECTION 1: PROVIDE VALUES FOR FOLLOWING 5 MACRO VARIABLES;

%let numvar=6;    * NUMBER OF ITEMS (MANIFEST VARIABLES);
%let numcl=2;    * NUMBER OF CLASSES;
%let numsap=27151;  * NUMBER OF SAMPLE SIZE;
%let start=2.5; * START VALUES FOR THE FIRST STAGE OPTIMIZATION;
%let datafile = "c:\abortion6.txt"; * LOCATION OF THE INPUT DATA FILE;
*************************************************************

SAS output 1 - Call NLPQN Subroutine in the First Stage to Conduct Latent Class Analysis

Computing Pi-star, the Two-Point Mixture Fit Index
The First Stage: Latent Class Analysis
The Objective Function Is to Minimize G-square
Optimization Results

<table>
<thead>
<tr>
<th>Iterations</th>
<th>67</th>
<th>Function Calls</th>
<th>91</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Calls</td>
<td>69</td>
<td>Active Constraints</td>
<td>1</td>
</tr>
<tr>
<td>Objective Function</td>
<td>5356.558615</td>
<td>Maximum Constraint</td>
<td>0.0000207942</td>
</tr>
<tr>
<td>Maximum Projected Gradient</td>
<td>8.5964821829</td>
<td>Value Lagrange Function</td>
<td>5356.5585942</td>
</tr>
<tr>
<td>Maximum Gradient of the Lagran Func</td>
<td>8.7492309364</td>
<td>Slope of Search Direction</td>
<td>-0.000050218</td>
</tr>
</tbody>
</table>

Latent Class Analysis

Total Expected Frequency
27203.5

LC1 Proportion LC2 Proportion
0.4834715      0.5165285

Conditional Probabilities

CP Positive Response (1)     1 0.9923362 0.6579097
                               2 0.9209919 0.0488761
                               3 0.9962743 0.8059632
                               4 0.9546303 0.0921695
                               5 0.9987107 0.6670409
                               6 0.9257254 0.0547449

CP Negative Response (2)     1 0.0076638 0.3420093
                               2 0.0790081 0.9519239
                               3 0.0037257 0.1940368
                               4 0.0453697 0.9078305
                               5 0.0012893 0.3329591
                               6 0.0742746 0.9452551
A vector of values equal to 2.5 was selected as start values. While the latent class proportions are 58% and 42%, respectively, the value of $\pi^*$ is near 0.188, indicating that in order to provide perfect fit, around 19% of the cases in the population are not taken into account. This suggests that the two-class model does not provide adequate fit; Dayton (2006) considered more complex models for these data.

References


Appendix: The SAS Program to Compute Pi-star with the Cheat4 Data

```sas
options nodate pageno=1 linesize=80 pagesize=60;

******************************************************************************
A SAS PROGRAM COMPUTING THE TWO-POINT MIXTURE INDEX OF FIT FOR THE
TWO-CLASS LCA MODELS
******************************************************************************

* SECTION 1: PROVIDE VALUES FOR FOLLOWING 5 MACRO VARIABLES;

%let numvar=4;   * NUMBER OF ITEMS (MANIFEST VARIABLES);
%let numcl=2;     * NUMBER OF CLASSES;
%let numsap=319;  * NUMBER OF SAMPLE SIZE;
%let start=1; * START VALUES FOR THE FIRST STAGE OPTIMIZATION;
%let datafile = "c:\cheat4.txt"; * LOCATION OF THE INPUT DATA FILE;

******************************************************************************

* SECTION 2: PREPARE DATA TO COMPUTE EXPECTED FREQUENCY;

* READ IN DATA FILE;

data lca;
infile &datafile;
input x1-x&numvar count;
run;

*NLP MACRO;
%macro Twoclasspistar;
```
Appendix: The SAS Program to Compute Pi-star with the Cheat4 Data (continued)

* CREATE A DATA SET WITH BINARY (0/1) DATA, EACH OBSERVATION CORRESPONDING TO THE CONDITIONAL JOINT DISTRIBUTION PATTERN OF ONE RESPONSE VECTOR;

```sas
data bin1 (drop=i j);
  do i= &numvar to 1 by -1;
    do j= (2**i) to 1 by -1;
      binary1 = putn((j-1),"binary&numvar.");
      if i=&numvar then
        output;
    end;
  end;
run;

data bin2 (drop= i j);
  do i=1 to &numvar;
    do j=1 to (2**i);
      binary2 = putn((j-1),"binary&numvar.");
      if i=&numvar then
        output;
    end;
  end;
run;

data bin (drop=i j binary1 binary2);
merge bin1 bin2;
  array x[&numvar] x1-x&numvar;
  do i=1 to &numvar;
    x[i]=substr(binary1,i,1);
  end;
  array y[&numvar] y1-y&numvar;
  do j=1 to &numvar;
    y[j]=substr(binary2,j,1);
  end;
run;

* CALL SAS PROC IML;
proc iml;

* CONVERT SAS DATAFILES INTO PROC IML MATRICES;

*WRITE THE BINARY DATA INTO THE MATRIX A;
use bin;
read all into a;

* WRITE THE COUNTS OF THE RESPONSE VECTORS INTO THE MATRIX OBSF;
use lca;
read all var {count} into obsf;

* CREATE A MACRO TO COMPUTE THE EXPECTED FREQUENCY FOR EACH CLASS;
```
Appendix: The SAS Program to Compute Pi-star with the Cheat4 Data (continued)

```sas
%macro expf;
   b=x[1:2*&numvar];
   b= b``;
   c=x[2*&numvar+1:4*&numvar];
   c= c``;
   p=j(2**&numvar,2*&numvar,0);
   q=j(2**&numvar,2*&numvar,0);
   do i=1 to &numcl**&numvar;
      do j=1 to 2*&numvar;
         p[i,j]=a[i,j]*b[j,];
         if p[i,j]=0 then
            p[i,j]=1;
         else p[i,j]=p[i,j];
         q[i,j]=a[i,j]*c[j,];
         if q[i,j]=0 then
            q[i,j]=1;
         else q[i,j]=q[i,j];
      end;
   end;
   pjoint=p[,#]; * EXPECTED FREQUENCY FOR EACH RESPONSE VECTOR IN LC 1;
   qjoint=q[,#]; * EXPECTED FREQUENCY FOR EACH RESPONSE VECTOR IN LC 2;
%mend;

*******************************************************************************
* SECTION 3: THE FIRST STAGE - CONVENTIONAL LATENT CLASS ANALYSIS;
* DEFINE THE BLOCK OF PARAMETER BOUNDS;

bounds=j(2,2*&numvar*&numcl, .);

* SPECIFY POSITIVE BOUNDS;

bounds[1,1:2*&numvar*&numcl]=1.e-6;

* DEFINE THE SUBROUTINE OF THE OBJECTIVE FUNCTION;

start F_objective(x) global (a, obsf, pjoint, qjoint);
   %expf;
   expf=pjoint+qjoint;
   ins1= obsf/expf;
   ins2=log(ins1);
   g=obsf#ins2;
   gsquare=g[+,#];
   return(gsquare); * DEFINE THE OBJECTIVE FUNCTION AS G-SQUARE/2;
finish F_objective;

start C_nlin(x) global(a, obsf,pjoint,qjoint);
   %expf;
   expf=pjoint+qjoint;
   Tot_expf=expf[+,#]; * AGGREGATE THE EXPECTED FREQUENCY;
   Tot_obsf=obsf[+,#]; * AGGREGATE THE OBSERVED FREQUENCY;
   c=Tot_obsf-Tot_expf; * THE TOTAL EXPECTED FREQUENCY IS EQUAL TO THE TOTAL OBSERVED FREQUENCY;
   return (c); * APPLY NONLINEAR CONSTRAINTS;
finish C_nlin;
```

328
Appendix: The SAS Program to Compute Pi-star with the Cheat4 Data (continued)

* NLP PROCEDURE;

x=j(1,2*&numvar*&numcl,&start);   * EXTRACT START VALUES;
optn= j(1,11,.) ;   * DEFINE THE VECTOR OF NLP OPTIONS;
optn[1]=0;   * SPECIFY A MINIMIZATION FOR THE OBJECTIVE FUNCTION;
optn[2]=2;   * SPECIFY THE AMOUNT OF OUTPUT PRINTED BY THE SUBROUTINES;

* WHEN OPTIONS 4-8 ARE SPECIFIED, MORE FUNCTION CALLS MAY BE REQUIRED TO
OBTAIN ACCURATE ESTIMATES;

optn[4]=3;   * DEFINE THE UPDATE TECHNIQUE FOR (DUAL) QUASI-NEWTON AND
CONJUGATE GRADIENT TECHNIQUES;
optn[5]=7;   * DEFINE THE LINE-SEARCH TECHNIQUE FOR THE NLPQN SUBROUTINE;
optn[6]=1;   * DEFINE THE VERSION OF THE ALGORITHM USED TO UPDATE THE
VECTOR OF THE LAGRANGE MULTIPLIERS;
optn[7]=1;   * DEFINE THE TYPE OF START MATRIX, G(0), USED FOR THE HESSIAN
APPROXIMATION;
optn[8]=21;  * DEFINE THE TYPE OF FINITE DIFFERENCE APPROXIMATION;

* NUMBER OF NONLINEAR CONSTRAINTS;

optn[10]=1;   * SPECIFY TOTAL NUMBER OF NONLINEAR CONSTRAINTS;
optn[11]=1;   * SPECIFY NUMBER OF EQUALITY CONSTRAINTS;

* MAXIMUM NUMBER OF ITERATIONS AND FUNCTION CALLS;

tc=j(1,10,.) ;
tc[1]=800;
tc[2]=1000;

* ADD TITLES FOR THE LATENT CLASS ANALYSIS;

title 'Computing Pi-star, the Two-Point Mixture Fit Index';
title2 'The First Stage: Latent Class Analysis';
title3 'The Objective Function Is to Minimize G-square';

* CALL NLPQN;

call nlpqn(rc, xr, "F_objective",x,optn,bounds) nlc="C_nlin" tc=tc;

* AGGREGATE THE TOTAL EXPECTED FREQUENCY AND COMPUTE THE LC PROPORTIONS;

%macro tef;
expf=pjoint+qjoint;
Tot_expf=expf[+1];
ppjoint=pjoint[+1];
qqjoint=qjoint[+1];
prop1=ppjoint/Tot_expf;
prop2=qqjoint/Tot_expf;
pistar=1-Tot_expf/&numsap;
%mend;

* RUN THE MARCO;

%tef;
Appendix: The SAS Program to Compute Pi-star with the Cheat4 Data (continued)

* CREATE A MARCO TO COMPUTE CONDITIONAL PROBABILITIES;

`%macro cp;
xr=xr`;
x1=xr[\(1\) : &numvar, ]; x2=xr[&numvar+1:2* &numvar, ]; x3=xr[2* &numvar+1:3* &numvar, ]; x4=xr[3* &numvar+1:4* &numvar, ];
p1=x1/(x1+x2); p2=x2/(x1+x2);
p3=x3/(x3+x4); p4=x4/(x3+x4);
cp=(p1\//p2)\|| (p3\//p4);
%end;

* RUN THE MARCO;

%cp;

* PRINT OUTPUT;

Print 'Latent Class Analysis';
print obsf [label='Observed Frequency'] expf [label='Expected Frequency'];
Print Tot_expf [label='Total Expected Frequency'];
print prop1 [label='LC1 Proportion'] prop2 [label='LC2 Proportion'];
print cp1 [label='Conditional Probabilities' rowname='CP Positive Response (1)'];
print cp2 [label=' ' rowname='CP Negative Response (2)'];

*************************************************************
* SECTION 4: THE SECOND STAGE - COMPUTE PISTAR;

* REDEFINE THE SUBROUTINE OF THE OBJECTIVE FUNCTION TO MAXIMIZE THE EXPECTED FREQUENCY;

start F_objective(x) global (a, pjoint,qjoint);
  %expf;
  expf=pjoint+qjoint;
  Tot_expf=expf[+]; * AGGREGATE THE EXPECTED FREQUENCY;
  return(Tot_expf); * REDEFINE THE OBJECTIVE FUNCTION AS TOTAL EXPECTED FREQUENCY;
finish F_objective;

* REDEFINE THE SUBROUTINE OF NONLINEAR CONSTRAINTS;

start C_nlin(x) global(a,obsf,pjoint,qjoint);
  %expf;
  expf=pjoint+qjoint;
  c=obsf-expf; * FOR EACH RESPONSE VECTOR, THE EXPECTED FREQUENCY IS EQUAL TO OR SMALLER THAN THE OBSERVED FREQUENCY;
  return (c); * APPLY NONLINEAR CONSTRAINTS;
finish C_nlin;

330
Appendix: The SAS Program to Compute Pi-star with the Cheat4 Data (continued)

* CALL NLP PROCEDURE;

x=xr;       * EXTRACT START VALUES;
optn= j(1,11, .); * DEFINE THE VECTOR OF NLP OPTIONS;
optn[1]=1;   * SPECIFY A MAXIMIZATION FOR THE OBJECTIVE FUNCTION;
optn[2]=2;   * SPECIFY THE AMOUNT OF OUTPUT PRINTED BY THE SUBROUTINES;
optn[10]=2**&numvar;   * SPECIFY TOTAL NUMBER OF NONLINEAR CONSTRAINTS;
optn[11]=0;   * SPECIFY NUMBER OF EQUALITY CONSTRAINTS;

* ADD TITLES FOR THE PI-STAR COMPUTATION;
title 'Computing Pi-star, the Two-Point Mixture Fit Index';
title2 'The Second Stage: Pi Optimization';
title3 'The Objective Function Is to Maximize the Total Expected Frequency';

* CALL NLPQN;
call nlpqn(rc, xr, "F_objective", x, optn, bounds) nlc="C_nlin";

* RUN THE MACRO TO AGGREGATE THE TOTAL EXPECTED FREQUENCY;
%tef;

* RUN THE MACRO TO COMPUTE THE CONDITIONAL PROBABILITIES;
%cp;

* PRINT OUTPUT;
Print " Pi-Star Results";
print obsf [label='Observed Frequency'] expf [label='Expected Frequency'];
Print Tot_expf [label='Total Expected Frequency'];
print Pistar [label='Pi-Star'];
print prop1 [label='LC1 Proportion'] prop2 [label='LC2 Proportion'];
print cp1 [label='Conditional Probabilities' rowname='CP Positive Response (1)'];
print cp2 [label=' ' rowname='CP Negative Response (2)'];

* EXIT SAS IML;
quit;

* CLOSE NLP MACRO;
%Mend;

* RUN MACRO;
%TwoclassPistar;
run;