Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 5

11-2014

A Comparison of Methods for Group Prediction with High Dimensional Data

Holmes Finch
Ball State University, whfinch@bsu.edu

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 84-105.

ISSN 1538 − 9472

A Comparison of Methods for Group Prediction with High Dimensional Data

Holmes Finch
Ball State University
Muncie, IN

High dimensional data is the situation in which the number of variables included in an analysis approaches or exceeds the sample size. In the context of group classification, researchers are typically interested in finding a model that can be used to correctly place an individual into their appropriate group; e.g. correctly diagnose individuals with depression. However, when the size of the training sample is small and the number of predictors used to differentiate the groups is larger, standard approaches such as discriminant analysis may not work well. In order to address this issue, statisticians have developed a number of tools designed for supervised classification with high dimensional data. The goal of this simulation study was to compare several such approaches for supervised classification with high dimensional data in terms of their ability to correctly classify individuals into groups, and to identify the number of variables associated with group separation. Results of the study showed that the Random Forest ensemble recursive partitioning algorithm was optimal for group prediction, while the Nearest Shrunken Centroid and Regularized Discriminant Analysis methods were optimal for identifying the number of salient predictor variables. The standard linear discriminant analysis approach was generally the worst performer across all high dimensional simulated conditions. Implications of these results to practice and directions for future research are discussed.

Keywords: Group prediction, Discriminant Analysis, High dimensional data, Regularlization methods

Introduction

High dimensional data refers to the case where the number of variables to be included in an analysis is equal to or exceeds the sample size (Bühlmann & van de Geer, 2011), and is written symbolically as $p>>n$. High dimensional data can create a variety of problems for many standard data analytic techniques, including those used in prediction and classification. In particular, when the number of predictors exceeds the sample size it is frequently not possible to obtain model parameter

Holmes Finch is the George and Frances Ball Distinguished Professor of Educational Psychology, and a professor of statistics and psychometrics. Email at whfinch@bsu.edu.

estimates because covariance matrices are often singular. In addition, in the high dimensional case there may be more unknown parameters than known data, leading to indeterminate estimation problems. Finally, in the high dimensional data case the correlations among variables is often very high, making parameter estimation very difficult. The result of all of these problems is that parameters and their corresponding standard errors are frequently not estimable in the high dimensional case. Furthermore, any estimates that are obtained are likely to be ill-conditioned and therefore unreliable (Kriegel, Kröger, & Zimek, 2009).

Given these difficulties, researchers have developed a set of statistical methods for the problem of high dimensional data. These methods are useful in a variety of contexts, including fitting of linear models, clustering observations based on a number of variables (often referred to as features in high dimensional literature), and classificaiton of individuals into one of several groups, using many predictors. The focus of the current research is on this latter application to group classification. Often in standard data problems where $n>p$, such classification is done using discriminant analysis. However, as we will see below, this approach is ill suited for use when $p>>n$, or even when p approaches n (Hastie, Buja, & Tibshirani, 1995). This Monte Carlo simulation study examines several methods that have been proposed for the high dimensional classification problem, including several based on discriminant analysis, as well as a variation of nearest centroid classificiation, and the recursive partitioning random forest methodology. The remainder of the manuscript is organized as follows: First we discuss discriminant analysis and explain why its use when $p>>n$ is problematic, followed by a description of alternative classifiers that have been proposed for this case. After next describing the study goals, we then outline the simulation study design, followed by a description of the simulation results. Finally, we discuss the results of our simulation and place them in the context of the broader high dimensional classification literature.

Goals of the current study

The goal of the current study was to compare the performance of several methods for group classification in the presence of high dimensional data. This comparison was made using both simulated data and an existing data example. This work adds to the literature in the field in three primary ways. First of all, there has not been another published study in which all of these classification methods have been compared with one another using Monte Carlo methods. While prior research has demonstrated the utility of several of these approaches using extant data (e.g.

Clemmensen, Hastie, Witten, & Ersbøll, 2011; Zhang, Dai, Xu, & Jordan, 2010; Hastie, Buja, & Tibshirani, 1995) or small simulations (Hastie, Tibshirani, & Friedman, 2009), no prior published study has systematically compared the performance of all of the methods described here, each of which has been suggested as a possible alternative for the high dimensional case. Thus, the current study should further the literature in this regard by providing researchers with more information about which such tools might be optimal under which conditions. The second goal of this study was to investigate the performance of RF in the presence of high dimensional data. While there have been some initial calls for such simulation research (Xu, Huang, Williams, Wang, & Ye, 2012), and some demonstrations with existing data (e.g. Zhang, Yu, Singer, & Xiong, 2001) there has not been a great deal of work done examining RF in this context. Finally, the third goal of this study was to introduce methods for high dimensional group prediction to social science researchers, in particular. Traditionally these methods have been used primarily with gene expression data, as a review of Bühlmann and van d Geer (2011) demonstrates. However, there are scenarios in the social sciences in which researchers are faced with small samples as well (e.g., Siklos & Kerns, 2007; Palmer, 2006; Sanden, 2008).

Methods for classification with high dimensional data

Linear Discriminant Analysis

Perhaps one of the most widely used classification methods is linear discriminant analysis (LDA). This technique, which is based upon a multivariate linear model, is used when there exists a grouping variable and a set of predictors that are believed to distinguish members of the 2 or more groups. The algorithm identifies weights for each of the predictors such that their linear combination maximally separates the groups from one another (Huberty & Olejnik, 2006). This linear combination appears in equation (1).

$$C_{ji} = \beta_{j0} + \sum \beta_{jm} x_{mi} + ln\left(\frac{n_j}{N}\right) \quad (1)$$

The terms in (1) are defined as:

C_{ji} = Classification score for group j for subject i
β_{j0} = Constant for group j
β_{jm} = Weight for predictor m for group j
x_{mi} = Value of predictor m for subject i
n_j = Sample size for group j
N = Total sample size

The natural log of the ratio of group size to the total sample serves as prior information about the relative frequency of the group in the population. In many applications, this prior probability of group membership is calculated using the values of n_j and N from the sample, as described above. However, the prior probability can also be provided directly by the researcher, bypassing the use of relative group size in the sample. This might be a useful strategy if it is known that the sample is not representative of the population in terms of the relative frequency with which members of each group appear. Determination of the coefficients in (1) is made so as to maximize the following criterion:

$$\beta_j^T \Sigma_b \beta_j \text{ subject to } \beta_j^T \Sigma_w \beta_j \leq 1 \quad (2)$$

Here, the coefficients (β) are as defined previously, with Σ_b being the between class covariance matrix, and Σ_w the within class covariance matrix. The resulting linear combination in (1) can be used to determine category membership for each observation in the original training data or in a cross-validation sample. Values of C_j are calculated for each member of the sample, and individuals are classified into the group for which they have the largest such score.

In terms of determining variable importance in terms of group classification, researchers typically rely on the structure matrix, which can be interpreted as the correlation matrix between the individual predictors and *C*. While there is not a hypothesis test for these values, recommendations for cut values have been suggested, including 0.32 (Tabachnick & Fidell, 2013), which will be used in the current study. Thus, absolute values of the structure matrix elements greater than 0.32 are taken as indicative that a predictor contributes to a classification solution.

In the case when $P > N$, which is the focus of this study, LDA is not typically a good choice for group classification because the within class covariance matrix

estimated using the sample is likely to be singular (Witten & Tibshirani, 2011). Even when this is not the case, Witten and Tibshirani have shown that the classifier in (1) will most likely exhibit a high variance, thereby degrading the resulting prediction model. Finally, in the case when *P* itself is large (irrespective of the size of *N*), LDA can prove problematic because the classifier in (1) will, by definition include all of the predictor variables, potentially leading to problems with interpreting the classifier function (Witten & Tibshirani). Therefore, while it is a popular and frequently used tool for researchers interested in classification, LDA may not be appropriate for cases in which the number of predictor variables is almost as large as, or larger than the number of subjects in the sample (Hastie, Tibshirani, & Friedman, 2009). Given these limitations, we will need to turn to alternative methods of classification better suited to the high dimensional problem.

Penalized LDA

One alternative for high dimensional classification is penalized LDA (PLDA), as proposed by Hastie, Buja, and Tibshirani (1995). PLDA is based upon a regularization of the discriminant function; i.e. a reduction in the number of predictors (sometimes referred to as features) used to develop the prediction algorithm. By limiting the number of predictors, the resultant discriminant function should not suffer from the problems associated with LDA when $P > N$. The key to this method working optimally is the use of an appropriate regularization strategy. PLDA shares the basic methodology described above for LDA, including the form of (1) for the prediction algorithm. However, (2) is adjusted to the following:

$$\beta_j^T \Sigma_b \beta_j - P\left(\beta_j\right) \text{ subject to } \beta_j^T \Sigma_w \beta_j \leq 1 \tag{3}$$

In (3), *P* is a penalty function designed to regularize the set of predictor variables, by reducing it to only those that are most salient in differentiating the groups.

Witten and Tibshirani (2011) describe the penalty function to be used in (3) as the PLDA-L1 algorithm appearing in (4):

$$\beta_j^T \Sigma_b \beta_j - \lambda \sum \left|\sigma_j \beta_j\right| \text{ subject to } \beta_j^T \Sigma_w \beta_j \leq 1 \tag{4}$$

In this case, σ_j is the within class standard deviation for predictor *j*, and λ is a tuning parameter that is set by the researcher. When λ is large, the relative

importance of individual predictors is reduced, and some will even go to 0, meaning that they do not contribute to the classification function in (1) at all. The value of λ is determined through the use of jackknife cross-validation, in which each member of the sample is removed in turn and various values of the tuning parameter are used with each jackknife sample. The optimal value is determined to be the one that minimizes classification error across the cross-validation samples. In addition, the inclusion of σ_j means that predictors with greater variation within classes will contribute less to the overall classification function than those with less such variability. Witten and Tibshirani assert that using PLDA-L1 will result in a function involving a subset of the predictors, and is most appropriate if the researcher desires a relatively sparse classifier function. It should be noted that Witten and Tibshirani also describe a second method for determining the penalty in (3), based on the fused Lasso method of regularization (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005). However, this approach was not employed in the current study because it assumes a linear ordering of the predictors, which was felt to be a limitation to its use in many applied settings.

Regularized Discriminant Analysis (RDA)

Guo, Hastie, and Tibshirani (2007) introduced an alternative to LDA in the high dimensional case that focuses on shrinking the within class covariance matrix (Σ_w) in the sample toward the diagonal matrix, through the use of a tuning parameter, γ. This shrunken version of Σ_w takes the following form

$$\Sigma_{w\gamma} = \gamma\Sigma_w + (1-\gamma)\,diag\left(\Sigma_w\right) \tag{5}$$

The value of γ ranges between 0 and 1, where $\gamma \to 1$ corresponds to standard LDA, and $\gamma \to 0$ yields highly regularized discriminant functions, in which a small number of predictors contribute to the within class covariance matrix. As with PLDA, jackknife cross-validation is used to identify the optimal value of γ. When $\Sigma_{w\gamma}$ is obtained, it is applied to (2), and the classifier in (1) is developed based upon this shrunken within class covariance matrix. In practice, RDA yields a classification function that utilizes many fewer predictors than are present in the data, or than would be used in standard LDA, thus avoiding problems associated with complex classifiers containing many correlated predictors, as cited above.

Sparse Discriminant Analysis (SDA)

RDA regularized the set of predictors by applying a penalty to the within class covariance matrix. Alternatively, regularization could be achieved through applying the penalties directly to the discriminant function coefficients for the predictors (β_j). This approach, known as sparse discriminant analysis (SDA) was described in Clemmensen, Hastie, Witten, and Ersboll (2011). It is based on the elastic net (Zou & Hastie, 2005), which is used with linear models in the presence of high dimensional and/or highly collinear data. In the context of discriminant analysis, this elastic net approach seeks to minimize the following function:

$$Y\theta_j - X\beta_j^{\,2} + \gamma\beta_j^T\Omega\beta_j + \lambda\beta_j \tag{6}$$

The parameters γ and λ are tuning parameters, Y is an indicator variable for whether an individual belongs to a particular group, θ_j is a score matrix, and Ω is a positive definite penalty matrix. In the current study, we use the elastic net approach suggested by Clemmensen, et al., such that $\Omega = \gamma I$, where I is the identity matrix. Jackknife cross-validation is used to determine the optimal values of γ and λ . The elastic net approach to regularization has some theoretical advantages over other approaches, including the lasso based PLDA method described above. Chief among these advantages are that highly correlated predictors tend to have similar coefficients in the final equation, and a greater number of predictors might be included in the final equation (Zou & Hastie, 2005).

Nearest Shrunken Centroids (NSC)

Another approach that we will examine for dealing with the high dimensionality classification problem is based upon an approach known as diagonal-covariance LDA. This method is based upon centroid classification, in which a multivariate mean (centroid) across all predictors is estimated for each group, and then new cases are placed in the class to whose centroid their scores are closest. NSC is a variant of this approach in which the class centroids are shrunken toward the overall centroid of the sample by an amount equal to a predetermined threshold value. The nearest centroid classification rule is expressed as:

$$\delta_{ij} = -\sum \frac{x_{im-}\bar{x}_{jm}}{s_m^2} + 2ln\pi_j \tag{7}$$

In (7) δ_{ij} is the discriminant score of subject *i* for category *j*, $\bar{x}_{jm}$ is the mean for predictor *m* in category *j*, s_m^2 is the variance of predictor *m* pooled across all categories, and x_{im} is the value of predictor *m* for subject *i*. An individual is placed into the group for which their value of δ_{ij} is smallest.

NSC adjusts the nearest centroid classification in the following way. First, the value d_{jm} is calculated, reflecting the difference between each group mean and the overall mean, as seen in (8).

$$d_{jm} = \frac{\bar{x}_{jm} - \bar{x}_m}{m_j\left(s_m + s_0\right)} \tag{8}$$

Here, terms are as defined in (7), with the addition that $\bar{x}_m$ is the overall mean across categories for predictor *m*, $m_j^2 = \frac{1}{n_k} - \frac{1}{n}$, and s_0 is the median of the s_m values. Its purpose is to ensure that d_{jm} does not become too large if the value of a given predictor is close to 0. The second step in NSC involves determining the degree to which the individual predictors' group means should be shrunk toward the overall mean across groups. In order to do so, the value of the threshold parameter, Δ must be made. This is typically done using jackknife cross-validation in which a potential Δ is used and predictions are made for each jackknife sample. The threshold value that yields the most accurate cross-validation predictions is the one to be used in the final NSC algorithm. Shrinkage occurs by adjusting d_{jm} as in (9):

$$d'_{jm} = sign\left(d_{jm}\right)\left(\left|d_{jm}\right|\right) - \Delta \tag{9}$$

Finally, (8) is solved for $\bar{x}_{jm}$ and shrunken versions of the predictor means are calculated as in (10).

$$\bar{x}'_{jm} = \bar{x}_m + m_j\left(s_m + s_0\right)d'_{jm} \tag{10}$$

The shrunken centroids obtained in (10) are then used in (7). An important point to note here is that if for a given predictor, the shrinkage takes its centroid value down to (or past) 0, the centroid is assigned the value of 0. As an example, if

a predictor centroid is 1 and the amount of shrinkage determined using (9) and (10) is -2, then the shrunken centroid value would be 0, because 1-2 takes the value down to (and past) 0. NSC is known to have two advantages when used with high dimensional data. First, it reduces the impact of predictors with high variances, thus also reducing the amount of noise in the predictions themselves. Second, it creates a *de facto* predictor selection algorithm by removing the impact of variables that contribute relatively less to group separation (Tibshirani, Hastie, Narasimhan, & Chu, 2002).

Random Forest

The final method of classification to be considered in this study is the Random Forest (RF) of Brieman (2001), which is based upon the classification and regression tree (CART) recursive partitioning algorithm that Breiman, Friedman, Olshen, & Stone (1984) described. For CART with a categorical outcome variable, predictors are used to partition members of the sample in ever more heterogeneous groups, with respect to the outcome. The partitioning continues until a predetermined stopping rule has been reached such that no further divisions of the sample will yield appreciable gains in prediction accuracy.

A problem with CART is that it has a tendency to overfit the training data, making the resultant prediction algorithm less generalizable to the general population. However, it is also true that CART solutions are unbiased so that if they are averaged across a great many samples from the population, the results should provide very accurate prediction heuristics (Dietterich, 2000; Bauer & Kohavi, 1999). Brieman used this unbiasedness property in developing RF, which consists of an ensemble of CART results applied to a sample, and then averaged to create a single prediction algorithm. RF works by randomly selecting *B* subsamples of the original sample, either with replacement and therefore being of size *n*, or only a portion of the total sample without replacement so that the subsample is less than *n*. Those individuals not included in a subsample for a given tree are referred to as the out of bag sample. In addition, a subsample of the predictors is also randomly selected, and used with CART to create a prediction tree for the subsample of individuals. This process is completed a large number of times (e.g. 1000), and the resulting trees are saved after each analysis. Each tree is then applied to members of the training sample, or to new individuals, and a predicted outcome (e.g. classification) is obtained. These results are then averaged across the *B* trees for each individual in order to obtain a RF predicted value. The diversity of solutions introduced through the large number of trees based on subsamples of both

individuals and predictors results in a final solution that is more generalizable than any individual CART model.

Variable importance in prediction is determined through permutation tests (Nicodemus, Malley, Strobl, & Ziegler, 2010). For RF, the permutation importance of an individual predictor variable is calculated by comparing the number of correct predictions made by the actual data (i.e. the predictor ordered as it appears in the original dataset) with the number of correct predictions made when the variable has been permuted (i.e. randomly shuffled), averaged across all trees in the ensemble. The classification accuracy rate across trees for the original variable with no permutation is then compared with that of the mean accuracy rate for the permuted trees. If the difference in prediction accuracy is large, and presumably in favor of the tree based on the original data, we would conclude that the variable is important in accurately predicting group membership. On the other hand, if the difference in classification accuracy between the actual and permuted values is very small, then we would conclude that the variable does not contribute much more to determining group membership than if it were random and thus totally unrelated to the outcome. More formally, importance for variable x_m for a single tree (t) is calculated as:

$$VI_t\left(x_m\right) = \frac{\sum I\left(y_i = \hat{y}_{iO}\right)}{\left|\bar{B}\right|} - \frac{\sum I\left(y_i = \hat{y}_{iP}\right)}{\left|\bar{B}\right|} \tag{11}$$

where

$\hat{y}_{iO}$ = Predicted class for observed data
$\hat{y}_{iP}$ = Predicted class for permuted data
B = out-of-bag sample

If variable x_m is not included in the tree, then VI=0. In order to obtain the overall variable importance measure for the RF, we then calculate

$$VI\left(X_m\right) = \frac{\sum_{t=1}^{T} VI_t\left(x_m\right)}{T} \tag{12}$$

where T is the total number of trees in the ensemble.

Methods

The research questions outlined above were addressed through the use of a Monte Carlo simulation study carried out with in the R software system, version 2.15.1 (R Foundation for Statistical Computing, 2011). The variables that were manipulated in the simulation study were selected in order to mirror conditions that researchers, particularly in the social sciences, might see when faced with a high dimensional dataset. For all conditions, data were simulated for two groups and unless otherwise noted the data were from a standard multivariate normal distribution.

Manipulated variables

Method A total of 6 methods were examined in this study, including LDA, RF, PLDA, SDA, NSC, and RDA. For methods relying on the setting of tuning parameters for optimal performance, the jackknife cross-validation methods described above were incorporated into the simulation code.

Sample size Sample size conditions included in the study were 10, 20, 30, 40, and 50. In all cases group sizes were held equal.

Number of predictors The number of predictors simulated in this study were 14, 28, and 50. Taken together with the sample size conditions discussed above, the ratios of *P* to *N* ranged from 5/1 to just over 1/4. While these conditions would not be considered terribly high dimensional in genetics, or another science where extreme high dimensionality is common, they do represent relatively high dimensional data in the context of psychology, education, and other social sciences, in which researchers typically strive to have many more subjects than variables.

Group mean separation The separation between group means was quantified in terms of Cohen's *d* effect size. For all predictors the groups' means differed by the same amount, either 0.2, 0.5, or 0.8. Thus, for example, in the *P*=50 group mean difference 0.5 case, all 50 variables were simulated to differ by 0.5 between the two groups. Group 1 was simulated to have means of 0 and standard deviations of 1 across conditions, and group 2 was simulated with means of 0.2, 0.5, or 0.8 and standard deviations of 1 for all predictors, depending on the group mean separation condition.

Correlation among predictors The predictors were simulated to have correlations among one another of 0, 0.5, or 0.8. These values were selected in order to assess performance of the methods in two relatively extreme cases (no correlation, very high correlation), and when the correlation was in the middle.

Distribution of the predictors In order to investigate the performance of the methods when data were normal and when they were not, two distribution conditions were simulated: multivariate normal and skewed with skewness of 2.5. Given the reliance of some of the methods on the assumption of normality, in particular LDA, it was of some interest to ascertain the impact that violating the assumption would have on performance of the methods.

Simulation outcomes Two outcome variables were examined in this study. First, the overall misclassification rate for a cross-validated sample drawn from the identical distribution as the training sample for a given combination of simulation conditions was recorded. This rate simply represents the proportion of cases that were incorrectly classified by each method. The second outcome variable of interest was the proportion of predictor variables that were correctly identified as being associated with group separation. As noted above, all predictors were simulated to differ between the groups, so in the population this proportion was 1 for every simulation condition. Therefore, this outcome variable reflects the proportion of predictors that each method correctly found to contribute to group differences. For LDA, a variable was considered to contribute to the classification solution if the absolute value of its structure value was 0.32 or greater (Tabachnick & Fidell, 2013). With respect to RF, variables were considered to be important if the permutation test statistic described above was statistically significant at α=0.05. With regard to RDA, SDA, and PLDA variable importance was determined through the standardized discriminant weights. Based on findings in Cao, Boitard, and Besse (2011), variables were considered to be important predictors if these standardized values were greater than or equal to 0.1. Finally, with respect to NSC, a predictor was considered to contribute to the prediction if its weights were not shrunken to 0, again in keeping with recommendations in the literature (Christin, Hoefsloot, Smilde, Hoekman, Suits, Bischoff, & Horvatovich, 2013).

All simulation conditions were completely crossed with one another for a total of 324 different simulations. For each of these simulations, 1000 replications were generated and analyzed. In order to ascertain which main effects and interactions of the manipulated conditions contributed significantly to the outcome variables, repeated measures analysis of variance (ANOVA) models were used. For each

combination of simulation conditions, the outcome variables were calculated for each of the methods studied here, for each replication. These outcomes were then averaged across the 1000 replications in order to create individual values for the two outcomes of interest. These values then served as the dependent variables in two separate ANOVA models (one for misclassification and one for proportion of predictors correctly identified). The within subjects variable was method of classification, and the between subjects variables were the other manipulated factors described above. In addition to the statistical significance of the main effects and interactions of these factors, the η^2 effect size was also used to identify model effects worthy of post hoc investigation. Main effects and interactions that were statistically significant, and which had η^2 of 0.1 or greater were considered "important", because they were associated with at least 10% of the variance in the outcome variable.

Results

Classification accuracy

The ANOVA used to determine which of the manipulated factors or their interactions were related to overall classification accuracy. The interaction of method (*M*) by sample size (*N*) by correlation (*C*) was significantly associated with classification accuracy ($F_{40,952} = 3.621,\ p < 0.001,\ \eta^2 = 0.132$), as was the interaction of *M* by number of predictors (*P*) ($F_{8,472} = 11.937,\ p < 0.001,\ \eta^2 = 0.168$), and the interaction *M* by mean difference (*D*) ($F_{68,952} = 10.514,\ p < 0.001,\ \eta^2 = 0.429$). The overall misclassification rates by method, sample size, and correlation among the predictors appear in Table 1.

Table 1. Misclassification rates by Method, Sample Size (*N)*, and Correlation among the predictor variables (*C*)

N	C	LDA	RF	PLDA	SDA	NSC	RDA
10	0	0.34	0.09	0.24	0.24	0.21	0.21
	0.5	0.39	0.13	0.23	0.23	0.35	0.37
	0.8	0.42	0.16	0.20	0.21	0.40	0.41
20	0	0.29	0.10	0.19	0.24	0.19	0.19
	0.5	0.37	0.12	0.18	0.24	0.35	0.35
	0.8	0.42	0.09	0.18	0.16	0.40	0.40
30	0	0.29	0.04	0.20	0.31	0.20	0.20
	0.5	0.37	0.07	0.17	0.13	0.36	0.36
	0.8	0.41	0.07	0.18	0.14	0.39	0.40
40	0	0.26	0.05	0.18	0.35	0.18	0.18
	0.5	0.36	0.08	0.18	0.17	0.35	0.35
	0.8	0.40	0.09	0.19	0.19	0.39	0.39
50	0	0.24	0.08	0.18	0.36	0.17	0.17
	0.5	0.37	0.10	0.16	0.21	0.36	0.36
	0.8	0.42	0.10	0.20	0.23	0.41	0.41
200	0	0.19	0.07	0.16	0.33	0.16	0.16
	0.5	0.36	0.08	0.18	0.30	0.36	0.36
	0.8	0.39	0.09	0.20	0.33	0.39	0.39

The results in Table 1 show that RF uniformly had the lowest misclassification rates of the methods studied here, across both sample size and correlation among the predictor variables. The highest misclassification rates belonged to LDA, particularly for the combination of *N* less than 40, and *C* of 0.5 or 0.8. For the combination of *N* less than 50 and *C* of 0.5 or 0.8, SDA had among the lowest misclassification rates, after RF, though when the predictors were uncorrelated, these rates were among the highest, particularly for larger sample sizes. Finally, PLDA did not exhibit increases in misclassification rates with increasing sample sizes, unlike SDA, and it generally had lower misclassification rates for *C* of 0.5 or 0.8 than any other method except for RF, and SDA with *N* less than 50. In short, PLDA generally maintained consistent misclassification rates at or just under 0.2 for the conditions simulated here.

Table 2: Overall Misclassification Rates by Method and number of Predictors (*P*)

P	LDA	RF	PLDA	SDA	NSC	RDA
14	0.37	0.14	0.31	0.22	0.34	0.33
28	0.34	0.09	0.23	0.21	0.32	0.31
50	0.33	0.10	0.15	0.20	0.31	0.30

Table 2 includes the misclassification rates for *M* by *P*. Each of the approaches exhibited lower misclassification rates in the presence of more predictors. This effect was muted, however, for all of the methods except PLDA. In the latter case, the decrease in the proportion of misclassified cases was 0.16 from 14 to 50 predictors, whereas for the other methods, the decline in misclassification was never more than 0.04. In other words, the number of predictors included in the analysis had a much greater impact on the performance of PLDA than it did on any of the other methods studied here. Finally, Table 3 includes the overall misclassification rates for *M* by *D*. Across all methods, misclassification rates declined as differences in group means increased. This decline was particularly notable for PLDA, which produced a difference in misclassification of 0.29 between *D*=0.2 and *D*=0.8. Similarly, LDA, NSC, and RDA also evinced declines in misclassification of more than 0.2 between the smallest and largest group separation conditions. On the other hand, both RF and SDA displayed much smaller such declines, though for these methods as well, the rates declined with increasing group separation.

Table 3: Overall Misclassification Rates by Method and Difference in Group Means (*D*)

D	LDA	RF	PLDA	SDA	NSC	RDA
0.2	0.47	0.15	0.45	0.26	0.45	0.43
0.5	0.35	0.11	0.19	0.21	0.31	0.30
0.8	0.23	0.07	0.16	0.17	0.21	0.21

Correct Identification of Predictors Contributing to Group Separation

As with the misclassification rates, ANOVA used to determine which of the manipulated factors or their interactions were related to the proportion of predictors correctly identified as being associated with group separation. The interaction of *M* by *P* was significantly associated with the proportion of predictors correctly

identified as related to group differences ($F_{40,392} = 2.818,\ p < 0.001,\ \eta^2 = 0.223$). In addition, the interaction of *M* by *D* ($F_{68,952} = 10.514,\ p < 0.001,\ \eta^2 = 0.457$), and *M* by predictor distribution (*PD*) ($F_{4,94} = 17.556,\ p < 0.001,\ \eta^2 = 0.428$) were also significantly related to the proportion of predictors identified as important.

Table 4: Proportion of Predictors Associated with Group Differences Correctly Identified by Method and Number of Predictors (P)

P	LDA	RF	PLDA	SDA	NSC	RDA
14	0.19	0.53	0.18	0.53	0.85	0.83
28	0.003	0.41	0.08	0.38	0.66	0.73
50	0.001	0.08	0.05	0.24	0.57	0.55

Table 4 includes the proportion of the number of predictors by *M* and *P*. LDA consistently displayed among the lowest, if not the lowest proportion of predictors correctly identified. The next lowest rates belonged to PLDA, which performed similarly to LDA with *P*=14, and somewhat better for *P*=28 and 50. RF and SDA had comparable predictor identification rates for *P*=14 and 28, but the performance of RF fell more dramatically for *P*=50 than was the case for SDA. The best performers in terms of correctly identifying predictor variables associated with the group differences were NSC and RDA, each of which had proportions that were 0.2 or higher than their nearest competitors. For example, when *P*=14, both methods accurately identified over 80% of the predictors as being associated with group separation. This value dropped to 57% and 55%, respectively, when *P*=50, which represented more accurate performance than any of the other methods, even at their best, when *P*=14.

The proportion of predictors correctly identified by the method (*M*) and group mean separation (*D*) appears in Table 5.

Table 5: Proportion of Predictors Correctly Identified by Method and Difference in Group Means (D)

D	LDA	RF	PLDA	SDA	NSC	RDA
0.2	0.47	0.15	0.45	0.26	0.45	0.43
0.5	0.35	0.11	0.19	0.21	0.31	0.30
0.8	0.23	0.07	0.16	0.17	0.21	0.21

As was evident in Table 4, across methods LDA had the lowest correct proportion of predictors, except for *D*=0.8, in which case PLDA had the lowest proportion. Indeed, the ability of PLDA to correctly identify the number of predictors associated with group membership did not seem to be associated with group separation, as its rate stayed large constant. RF and SDA had similar rates to one another for *D*=0.2 and *D*=0.8, but SDA performed somewhat better when *D*=0.5. Neither of these methods performed as well as RDA or NSC, however. RDA had the highest proportion of predictors correctly identified for both *D*=0.2 and 0.5, and was slightly lower than NSC for *D*=0.8. Furthermore, the rates for RDA were largely unaffected by the degree of group separation, making it almost as accurate for low mean differences as for high ones. On the other hand, the performance of NSC was much more strongly influenced by *D*, as is evidenced by the change in the proportion of predictors from 0.2 to 0.8.

Table 6 includes the proportion of predictors by PD.

Table 6: Proportion of Features Correctly Identified as Important by Method and Predictor Distribution

PD	LDA	RF	PLDA	SDA	NSC	RDA
Normal	0.13	0.38	0.20	0.43	0.66	0.98
S2.5	0	0.30	0.07	0.33	0.65	0.42

Several of the methods were deleteriously impacted by the presence of skewness in the distribution of predictors, in particular RDA, which was nearly perfect in identifying the correct number of important predictor variables when the data were normal, but did so less than half the time for skewed data. Similarly, LDA, RF, PLDA, and SDA all had proportions of predictor rates for the S2.5 condition 0.08 or more lower than was the case with normal data. On the other hand, the performance of NSC in terms of correctly identifying the number of predictors was virtually unaffected by predictor distribution.

Discussion

The goal of this simulation study was to compare several methods for supervised group classification in the presence of high dimensional data. Prior work in this area has tended to focus on a small number of such methods using applied examples with single datasets, or small simulation studies with relatively few manipulated conditions. The goal of this study was to expand upon these earlier efforts in several ways. First, by utilizing a larger set of simulated conditions than has been seen previously, we were able to test the various methods across a wider array of scenarios. In addition, we included a number of methods in this study that had not been previously compared with one another, including RF, which has never been systematically studied in the high dimensional case. Finally, this study examined the performance of the methods both in terms of their ability to correctly classify individuals into groups, and in terms of their use of salient predictors.

As described above, the results of this simulation study clearly support the use of RF if the primary goal of the researcher is to correctly classify individuals into their appropriate groups. No other method was nearly as effective in this regard, across all conditions simulated here. Conversely, standard LDA was the worst performer in terms of prediction accuracy, across virtually all conditions simulated here. The other approaches, each of which relied on some type of regularization or penalty function, produced misclassification rates between these two methods. In examining why RF might have performed so much better than the alternatives, we might consider its very nature as a recursive partitioning algorithm. As noted above, a problem with many prediction models in the high dimensional case is that the covariance matrices used to obtain model coefficients are ill behaved and sometimes singular. The regularization methods studied here (e.g. RDA, PLDA, SDA, NSC) each attempts to deal with this problem by reducing the number of predictors that are used in the prediction. However, in doing so, they also reduce the number of variables that contribute to group prediction, including those that might be salient. RF, on the other hand, does not use the covariance matrix at all, and thus does not face the problem of poor estimation of model coefficients faced by LDA, and reduction in the number of variables used in prediction that is a part of the regularized approaches. RF simply divides the sample based on the available data, selecting the best predictors at each step of the tree building process. Furthermore, because it relies on a large number of such trees, each of which is based upon a subset of the predictors and members of the sample, it should be more generalizable to the population than perhaps are some of the other methods. And indeed, we found this to be the case in the current study.

While RF provided the most accurate predictions of group membership, it was not particularly effective at identifying the number of salient predictors of group separation. The permutation test used to do so is still fairly new and untried, and so while it has been shown to work reasonably well with larger samples (Nicodemus, Malley, Strobl, & Ziegler, 2010), there has been little work done with small samples, regardless of the number of predictors. Given that significance for a given predictor variable is determined by comparing classification accuracy using it in its natural state, and when it is randomly ordered, it is possible that with small samples and many predictors there is simply little difference in accuracy associated with any one variable. On the other hand, both NSC and RDA were much more accurate in terms of identifying the number of predictors associated with group separation. In considering which of these methods might be optimal if a researcher's goal is to identify variables associated with group separation, the results of this study would suggest that the decision should be based upon the nature of data being used. For example, if the researchers are unsure as to how different the predictor group means are, or if it is known that differences for some of them are relatively small, and the data are normally distributed, then RDA might be the best choice. Its ability to correctly identify the number of salient predictors was optimal when the data were multivariate normal, and it seemed largely uninfluenced by the degree of mean separation. In particular, it was the most effective approach when the effect size separating the groups was small. On the other hand, if the researcher knows that the data are not normally distributed, NSC might be the best approach to use because it was the least affected by the skewness simulated here. RDA performed relatively poorly in the presence of skewed data.

Recommendations and directions for future research

The results of this study suggest some recommendations for practice for researchers faced with high dimensional data. First, if the primary goal is to develop some type of prediction algorithm to be used with future cases, then RF seems to be the best choice. It provided much more accurate predictions than any of the other methods, regardless of the nature of the data. On the other hand, if the researcher is most interested in trying to identify which variables are most associated with group separation, then NSC or RDA may be better choices than RF. In particular, if the data are normally distributed, RDA would be recommended, whereas if the data are skewed then NSC is likely the optimal choice. In all cases, LDA is not recommended when the number of predictors approaches, or is larger than the sample size.

The current study represents an extension of prior work in this area in terms of the number of high dimensional prediction methods examined, and the number of conditions simulated. However, it also has limitations that future research should seek to address. First of all, only two groups were simulated here. Future studies in this area need to compare the performance of these methods with three or more groups. In addition, all of the variables were simulated to be related to group separation. However, in reality researchers are often faced with a situation in which only some of the variables are related to group differences. Therefore, future simulation studies should include some predictors that are not different between the groups. Finally, given the clear impact of predictor distribution on the accuracy of some methods, future studies should expand upon the nature of nonnormal data, including some categorical variables.

References

Bauer, E. & Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning, 36*, 105-139.

Breiman, L. (2001). Random Forests. *Machine Learning 45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* New York: Wadsworth.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data*. New York: Springer.

Cao, K-A. L., Boitard, S., & Besse, P. (2011). Sparse PLS Discriminant Analysis: Biologically Relevant Feature Selection and Graphical Displays for Multiclass Problems. *BMC Bioinformatics, 22*, 253-263.

Christin, C., Hoefsloot, H.C., Smilde, A.K., Hoekman, B., Suits, F., Bischoff, R., & Horvatovich, P. (2013). A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics. *Molecular Cell Proteomics, 12*(1), 263-276.

Clemmensen, L., Hastie, T., Witten, D., & Ersbøl, B. (2011). Sparse Discriminant Analysis. *Technometrics, 53*(4), 406-413.

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning, 40*, 139-157.

Guo, Y., Hastie, T. & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics, 8,* 86–100.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized Discriminant Analysis. *Annals of Statistics, 23,* 73-102.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

Huberty, C. J. & Olejnik, S. (2006). *Applied MANOVA and Discriminant Analysis*. Hoboken, NJ: John Wiley & sons, Inc.

Kriegel, H-P, Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A Survey on Subspace Clustering, Pattern-Based Clustering and Correlation Clustering. *ACM Transactions on Knowledge Discovery from Data, 3*(1), 1-58.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The Behavior of Random Forest Permutation Based Variable Importance Measures under Predictor Correlation. *Bioinformatics, 11*, 110-122.

Palmer, G. A. (2006). Neuropsychological Profiles of Persons with Mental Retardation and Dementia. *Research in Developmental Disabilities: A multidisciplinary Journal, 27*(3), 299-308.

R Foundation for Statistical Computing. (2011). R Software, V. 2.15.1. Vienna, Austria. Sanden, J. (2004). Math/FCS Class Boosts Test Scores. *Journal of Family and Consumer Sciences, 96*(1), 18-19.

Sanden, C., Befus, C. & Zhang, J. Z. (2008). Clustering-Based Genre Prediction on Music Data. Paper presented at the annual meeting of the Association for Computing Machinery, Montreal, QC, May.

Siklos, S. & Kerns, K. A. (2007). Assessing the Diagnostic Experiences of a Small Sample of Parents of Children with Autism Spectrum Disorders. *Research in Developmental Disabilities: A multidisciplinary Journal, 28*(1), 9-22.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using Multivariate Statistics.* Boston, MA: Pearson.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Science, USA, 99,* 6567-6572.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, B, 67*, 91-108.

Witten, D. M. & Tibshirani, R. (2011). Penalized Classification using Fisher's Linear Discriminant. *Journal of the Royal Statistical Society, B, 73*, 753-772.

Xu, B., Huang, J. Z., Williams, G. J., Wang, Q., Ye, Y. (2012). Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces. *International Journal of Data Warehousing and Mining, 8*(2), 1-20.

Zhang, H., Yu, C-Y, Singer, B., & Xiong, M. (2001). Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data. *Proceedings of the National Academy of Sciences, 98*(12), 6730-6735.

Zhang, Z., Dai, G., & Jordan, M. I. (2010). Regularized Discriminant Analysis, Ridge Regression, and Beyond. *Journal of Machine Learning Research, 11*, 2199-2228.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, B, 67,* 301–320.