Journal of Modern Applied Statistical Methods

Volume 2 | Issue 1 Article 27

5-1-2003

A Semiparametric Regression Model For Oligonucleotide Arrays

Jianhua Hu
University of North Carolina, jhu@bios.unc.edu

Guosheng Yin
M. D. Anderson Cancer Center

Journal of Modern Applied Statistical Methods
May, 2003, Vol. 2, No. 1, 256-267

1538 – 9472/03/$30.00

A Semiparametric Regression Model For Oligonucleotide Arrays

Jianhua Hu
Department of Biostatistics
University of North Carolina

Guosheng Yin
Department of Biostatistics
M. D. Anderson Cancer Center

A semiparametric model incorporating the spline smoothing technique is proposed to study oligonucleotide gene expression data. No specific parametric functional form is assumed for mismatch probe intensities, which allows much more flexibility in the fitted model. The new approach improves the model fitting, hence the estimation of expression indexes. The method is applied to a data set of 18 HuGeneFL arrays.

Key words: Affymetrix, gene expression, microarray, semiparametric spline smoothing

Introduction

DNA microarray technologies have been increasingly used and began to play an important role in many areas of biomedical research. There are two most popular types, namely cDNA microarrays and oligonucleotide arrays. The common advantages of them are to monitor the expression levels of very large numbers of genes simultaneously and repeatedly in cell lines, human tissues and a wide range of organisms. Microarrays have the potential and power to advance our knowledge and understanding at a genomic scale. In particular, the high-density oligonucleotide array has been shown to be very promising. Not only does it have the capability of monitoring all yeast genes, mouse and human genes, but it also can identify important genes and classify disease types or states reliably, due to its special design feature.

The distinctive feature of the oligonucleotide array technology is the effective utilization of the probe redundancy. Multiple oligonucleotides of different sequences are hybridized onto different regions of the same RNA that are complementary to the oligonucleotides.

Correspondence should be sent to: Jianhua Hu, graduate student, Department of Biostatistics, McGavran-Greenberg Hall, CB# 7420, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7420. Phone: 919-966-7287. Email: jhu@bios.unc.edu.

It offers us the possibility to test and examine the stability and reliability of gene expression measurements (outlier detection), improve the accuracy of RNA quantification, reduce cross-hybridization effects, and thus reduce the measurement noise and false-positive percentages. Usually, a probe set of around 20 pairs of a particular length (25 nucleotides typically) represents a gene uniquely (Lockhart et al., 1996).

The other source of redundancy is that mismatch (MM) probes are used, which are identical to their correspondent perfect match (PM) except for a single base mutated at the central position (13th position typically). The MM probes can provide some information on background and cross-hybridization signals, and provide the ability to discriminate between "real" signals and those due to non-specific or semi-specific hybridization (Lipshutz et al., 1999). In other words, the design of oligonucleotide arrays with PM/MM probe sets can improve the differentiating ability over the cDNA arrays that use a single spot. It can help to distinguish whether a signal detected is really due to the hybridization onto the intended RNA region or it happens just by chance due to cross-hybridization or other measurement errors.

Obtaining an accurate gene expression index is essential and fundamental for further research and analysis of oligonucleotide arrays, such as differentiating important genes, classifying genes to co-regulated or anti-coregulated groups and categorizing samples. Hence, it is very

important to develop some methodologies to estimate the gene expression indexes as accurately as possible.

In recent years, various statistical methods have been proposed for analyzing oligonucleotide arrays. For example, the GeneChip software computes the "average difference" (AD) (http://www.genechip.org/index.affx).

Affymetrix's average log ratio is based on log(PM/MM) where the log transformation may be helpful in reducing the skewness and the variation. Li and Wong (2001) proposed a parametric regression model to calculate the model-based expression indexes (MBEI) based on the difference (PM-MM). It can improve the fittness of hybridization intensity extracted from PM and MM, and model the probe effects explicitly. Also, MBEIs are closer to the underlying true gene expression indexes than those provided by most of other software. The way of dealing with the relationship between PM and MM for almost all the above methods is to subtract MM from PM or log(MM) from log(PM) directly. The model based on (PM-MM) assumes a linear relationship between PM and MM and the regression coefficient of MM equals one. Although the old Affymetrix pre-5.0 algorithm claims that there is a linear relationship between most PM and MM probes, there are still a certain amount of probes with nonlinearity. Better fitting models to these genes are desired in order to avoid missing some important biological information.

In practice, the paired PM and MM probe expression levels may not be linearly correlated for a specific probe set (Schadt et al., 2001). As shown in Figure 1, we randomly chose the probe set 17 of Gene 2111 and obtained the scatter plot of PM versus MM intensity levels with a smoothing spline curve fitted after normalization. It is clear that the relationship between PM and MM is not simply linear and some curvature pattern needs to be addressed. For the same gene, we also plotted log(PM) versus log(MM) with a smoothing spline fit. Although the log transformation helps clarify the pattern between them, there is still a curve trend. Therefore, there may be some excess non-linearity that cannot be captured by the parametric model simply based on (PM-MM).

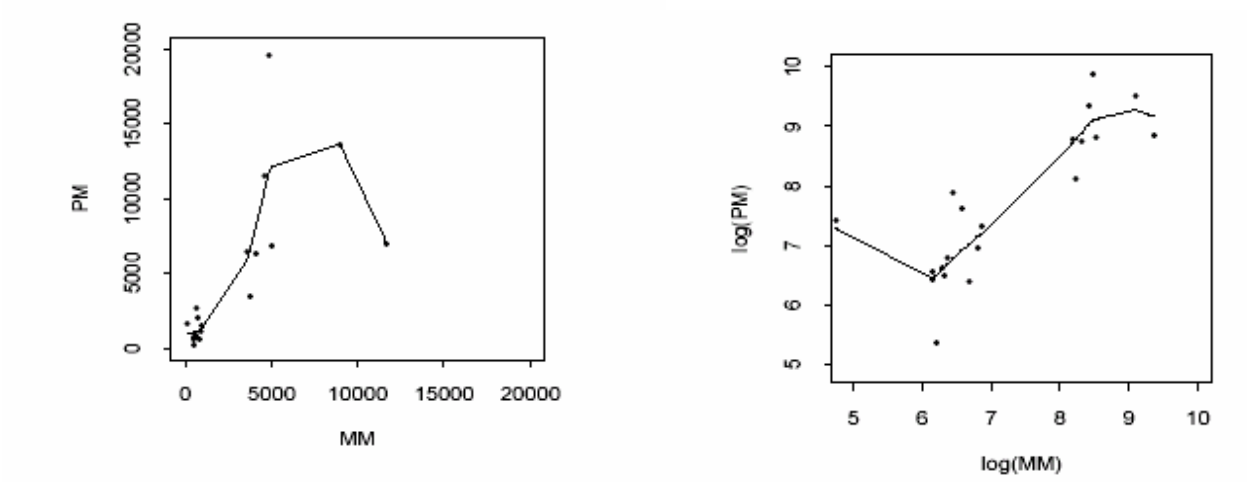

Figure 1: Smoothing spline fitting curves of PM versus MM and log(PM) versus log(MM) for probe set 17 of Gene 2111.

Another notable feature is that it is not rare for MM to be bigger than PM expression intensities after some are removed as outliers. The old Affymetrix pre-5.0 algorithm sets the expression levels of probes to be positive only if PM-MM $\geq$ SDT or PM/MM $\geq$ SRT, where SDT is the statistical difference threshold and SRT is the statistical ratio threshold. By this brutal truncation, it throws away many probes such that some useful biological information might be lost. Current Affymetrix MAS 5.0 handles this situation by setting MM always lower than its paired PM, which is similar to the approach of truncation

(Irizarry et al., 2001). But in many situations, the phenomenon of intensities of MM larger than PM may be caused by some sensible biological variations. Thus researchers still want to keep the features in the data analysis. Moreover, the algorithm is not as flexible and adjustable as model-based approaches.

Li-Wong's reduced model has been proved to be simple, feasible and popular with collaborating biologists and have several aspects of superior behavior. It can produce better estimation for the gene expression indexes, which is one of the most critical steps for further analysis. Since MM probes are used to eliminate the background and hybridization noise as much as possible, the one of most interest to researchers is still PM probes. Validity and goodness-of-fit of a model is essential to obtain accurate parameter estimates and statistical inferences.

We propose a semiparametric regression model to study PM probes with adjustment for MM probes in this article. After normalizations and dropping outliers, we keep the original feature for each gene and seek to obtain a better model-fitting by capturing the nonlinear relationship between PM and MM probes with a semiparametric approach based on Li-Wong's reduced model. We do not assume any parametric functional form of MM while the multiplicative relationship between the gene expression index (q) and the increasing rate (the probe sensitivity index, f) is still kept as in Li-Wong's reduced model. The approach involves three stages and relaxes the restriction of the regression coefficient of PM on MM being one, which is completely data-driven. We apply the proposed mothed to the analysis of HuGeneFL oligonucleotide arrays for Antibody Stain CEL data (http://thinker.med.ohio-state.edu/projects/fbss/index.html).

Methodology

Let θ_i be the expression index for the gene in the ith sample which is the primary target of interest. The full model proposed by Li and Wong (2001) for each gene is given by

$$\begin{aligned} PM_{ij} &= n_j + q_i(a_j + f_j) + e_{ij} \\ MM_{ij} &= n_j + q_i a_j + e_{ij}, \end{aligned} \qquad (1)$$

where PM_{ij} and MM_{ij} are the PM and MM intensity values for the ith array and the jth probe pair for this gene, $i=1, \ldots, I; j=1, \ldots, J$. Note that n_j is the reference response due to nonspecific hybridization, a_j is the increasing rate of MM response, f_j is the additional increasing rate of PM response, and e_{ij} represents a random error. There are many parameters in the full model, whereas a parsimonious statistical model may be preferred with the smaller sample size. A simpler reduced model (LWR) for the difference PM-MM is strongly supported by collaborating biologists. The model is given by

$$PM_{ij} - MM_{ij} = q_i f_j + e_{ij} \qquad (2)$$

It states that the PM and MM intensity differences have a multiplicative relation between q and f.

For the purpose of identifiability, a constraint is set as $\Sigma_j f_j^2 = J$. The error terms are assumed to be independent and identically normally distributed, i.e. $e_{ij} \sim N(0, s^2)$. Depending on the value of f_j, the least square estimate for q_i is

$$\hat{\theta}_i = \sum_j \frac{\left(PM_{ij} - MM_{ij}\right) f \varphi_j}{J} \qquad (3)$$

and the approximate standard error is given by,

$$\begin{aligned} S.E.\left(\hat{\theta}_i\right) &= \sqrt{\sigma^2 / J}, \\ \sigma^2 &= \sum\nolimits_j \left(fitted - observed\right)^2 \Big/ \left(J - 1\right) \end{aligned} \qquad (4)$$

An iterative least square algorithm is carried out for the estimation of the parameters. A software DNA-Chip Analyzer (dChip) has been developed to fit the parametric regression model that Li and Wong proposed (http://www.dchip.org/).

However, Li-Wong's reduced model (LWR) is analogous to the usual regression model for the difference between the pre-treatment (baseline) and post-treatment effects in clinical trials. In some sense, it forces the regression parameter of MM to be one which is a very stringent restriction and may affect the goodness-of-fit of the model tremendously. Moreover, there

is strong evidence of a non-linear relationship between PM and MM intensities (see Figure 1). Therefore, we propose a semiparametric approach to model the expression intensity data for each gene. Inspired by the additive partially linear models (Heckman, 1986; Hastie & Tibshirani, 1990), we model MM based on a nonparametric spline smoothing technique (LWS), namely,

$$PM_{ij} = g(MM_{ij}) + q_i f_j + e_{ij} \qquad (5)$$

where $g(.)$ is an unknown smooth function and is estimated with the cubic spline smoothing method. In many instances, rather than modeling every covariate nonparametrically or parametrically, a semiparametric partially linear regression model is more desirable. The model specification for the oligonucleotide array data is particularly appealing since the gene expression index q is the major interest, while the effects of MM are nuisance.

We can draw statistical inferences and estimate q by making minimal assumptions about the effects of MM with a fully nonparametric function. LWR does not have the same computational issue (too many parameters for sample sizes of practical use) as Li-Wong's full model that involves too many parameters. Basically, we relax the relationship between PM and MM to get a better fitted model and expect to have a more accurate estimate of the expression indexes. Hence, it is practically applicable to oligonucleotide gene expression data analysis.

Our estimating procedure involving three stages of iterative algorithms is described as follows:

Stage 1**:** Take LWR estimates as the initial values of $q_i^{(0)}$ and $f_j^{(0)}$. Note that LWR itself iteratively fits the sets of q_i and f_j while treating one of the two sets as known and fixed. We calculate the initial values using the dChip software.

Stage 2**:** Use the cubic spline smoothing technique to fit a nonparametric model with PM_{ij}-$q_i^{(0)} f_j^{(0)}$ as the response and MM_{ij} as the predictor, and thereby get the predicted values of $\hat{g}(MM_{ij})$.

Stage 3: Calculate the updated PM values $PM_{ij}^{new} = PM_{ij}^{old} - \hat{g}(MM_{ij})$, then regress the new estimates of PM on q's and f's, namely, $PM_{ij}^{new} = q_i f_j + e_{ij}$. The new estimates of q's and f's have been obtained. Go back to Stage 2, and continue till the prescribed convergence criteria are met.

Spline smoothing methods consisting of piecewise cubic polynomials are popular because they provide great flexibility for fitting the data and model non-linearities without specifying a functional form, with fewer parameters than higher-degree splines. To reduce the undesirable instability in the tails, one may restrict the function to be linear before the first knot and after the last knot. Fitting a cubic spline model which minimizes the residual sum of squares while

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int \{g''(x)\}^2 dx \qquad (6)$$

adjusting the smoothness of the resulting spline can be achieved by minimizing the penalized residual sum of squares

The smoothing parameter λ controls the trade-off between bias and variance and may be estimated by the cross-validation procedure. Excellent reviews of nonparametric regression and spline smoothing are available in the literature (Silverman, 1985; Eubank 1999).

Results

Description of Experiment and Data set

The data set is from an experiment conducted by the Division of Human Cancer Genetics at the Ohio State University (Lemon et al., 2002). There are 18 HuGeneFL arrays, each of which was loaded with 11 ug/200uL labeled cRNA. As shown by the graph in the Appendix, the process is described as the following. Human fibroblast cells were grown in media supplemented with 20% FBS for 5 passages (27 flasks) according to the distributor's recommendations. After 48 hours of placing cultures in serum-reduced media (0.1% FBS), 9 flasks (Stimulated) were returned to a 20% serum condition for 24 more hours and were then placed in RNA-Stat60. Cells from the other flasks (Starved) were placed in RNA-Stat60 directly after being placed in

serum-reduced media for 48 hours. Finally total RNA was extracted and purified according to a certain criterion. Based on the above steps, a set of stimulated and starved samples is produced. Another RNA sample was produced as a balanced mixture of simulated and starved samples, which is called the 50:50 sample.

For each condition (serum stimulated, serum starved and a 50:50 mixture of serum stimulated and starved), two aliquots of RNA were drawn and processed separately on three consequent days. Meanwhile, spiked-in genes were added in the following way: *Lys* and *Phe* RNAs at 0.08 ng/8μg total RNA were added to Stimulated RNA samples. The Starved samples received the same amount of *Dap* and *Thr* and all the four spiked-in genes at 0.04 ng/8μg were assigned to the 50:50 samples. Another set of control genes were added as well, which were *BioB*, *BioC*, *BioD* and *Cre* with final concentrations of 1.5, 5, 25 and 100 pM, respectively. For each group (Stimulated, Starved and 50:50), six replicated HuGeneFL arrays were produced. Eighteen arrays were produced in total. The technical variability was minimized through using a single fluidics station and a same lot for the 18 arrays. Multiple experiments or arrays for each gene allows researchers to evaluate the potentially different variability of genes.

There are 7129 probe sets in each array. Among them, a total of 149 genes are represented twice or more although they might not be in the same probe set. Most of the probe sets have 20 probe pairs. However, there are 330 probe sets with probe pairs less or more than 20. To compare Li-Wong's reduced model with our new proposal, the 330 probe sets were left out without losing any practical meaning.

The experimental design has very appealing features that the relationship among the arrays are known in advance and control genes are spiked in. Hence, it is suitable to use the data set to make comparisons among different estimation approaches.

Normalization, Variance and Goodness-of-fit

Because scanned images may have different overall brightness, it is important to normalize arrays such that they have comparable brightness before any analysis on expression levels. A traditional Average Difference (AD) method analyzes one array at a time, thus normalization among the different arrays can be done after calculating the quantities of interest. Because the model-based expression index analysis involves different arrays simultaneously, the comparable brightness of the arrays needs to be assured. As a very important issue, normalization has been extensively discussed and studied in the literature, and it is still an active research area.

We use the normalization method based on an "invariant set" (Li & Wong, 2001; Schadt et al., 2002). Normalization is based on probe values of non-differentially expressed genes that are identified through an iterative procedure (called the "invariant set"). Keeping the array which has the median overall brightness (the baseline array) as the invariant one, all the other arrays are normalized to it. The two arrays are drawn on the y-axis and x-axis, respectively. A straight line through the origin or a curve (i.e. smoothing spline) is fitted to the scattered points, which shows the normalization relationship between the two arrays.

If the variance of the model based expression index is overestimated, it may be possible not to differentiate some important genes that are supposed to express significantly, especially for genes with low expression levels. Hence, the model which yields smaller variances of the estimated expression indexes is desired. On average, LWS reduces the standard error of θ by 22% with respect to LWR. It indicates that LWS gives the more stable estimated expression index in terms of the 20 probe pairs than LWR. Figure 2 shows the histogram plots of standard errors of all the expression index estimates from both LWR and LWS. Obviously there are shifting differences between the distributions of S.E.'s from the two models (LWR and LWS). Most of the S.E.'s from LWS are within the range of (0, 500) while those from LWR even exceed beyond 1000.

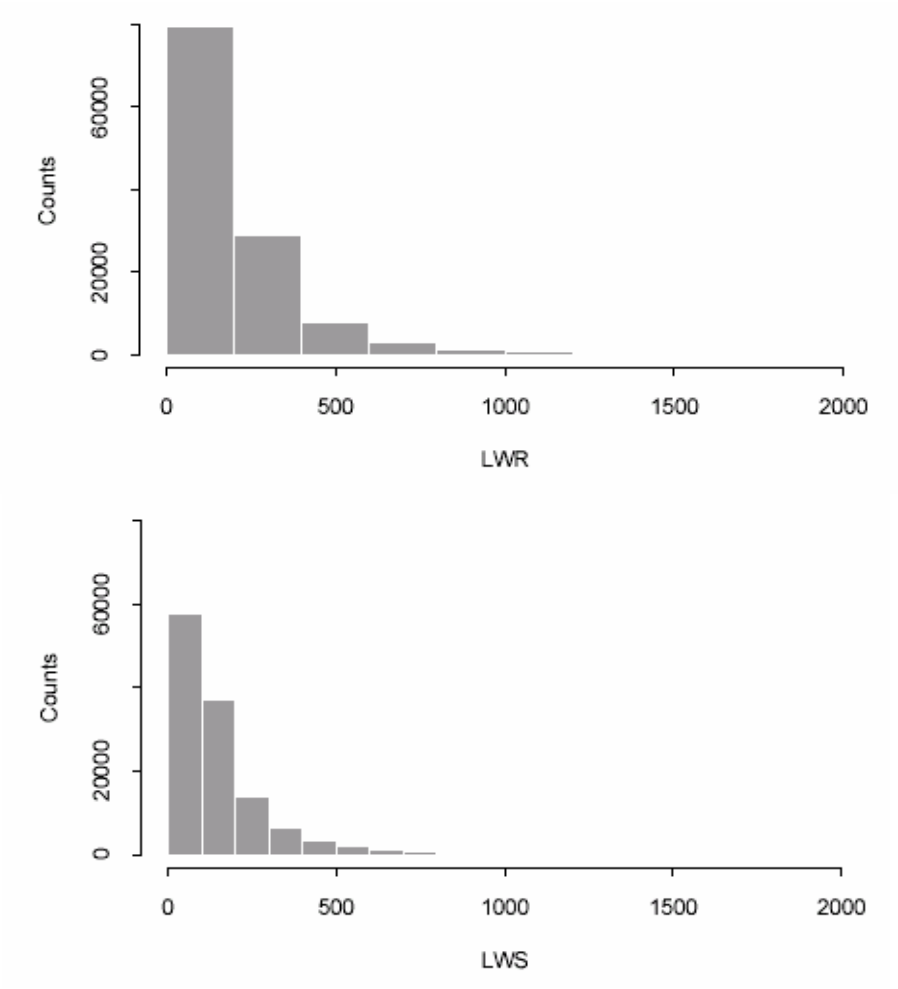

Figure 2: Histograms for standard errors of the estimated gene expression indexes.

Figure 3 presents the plot of residues of the fitted model versus predicted values for Gene 1007 (chosen randomly) from the two models, respectively. The horizontal line is the reference with the residue being zero. It is clear that the scatter plot from LWS gives a more random and symmetric pattern around the reference line, while LWR has more points further deviated away from the zero-line.

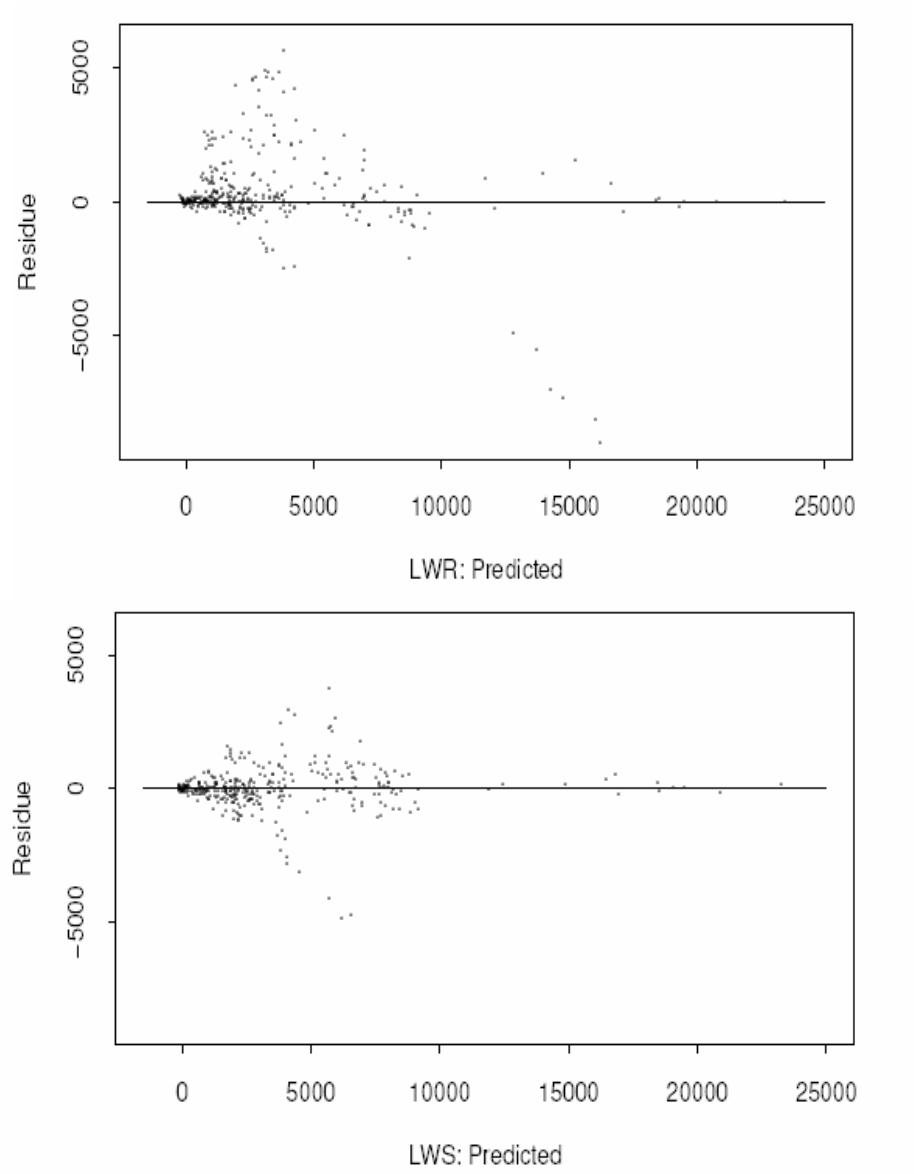

Figure 3: Residuals versus predicted values for gene 1007.

The better the model fits, the higher correlation of the predicted and observed PM values is supposed to be. Thus, correlations for all the probe sets are calculated for LWR and LWS. The histograms of the correlations obtained from the two models are shown in Figure 4, respectively. Note that most of the correlations obtained from LWS concentrate within 0.92 to 1, while the correlations from LWR even go below 0.90.

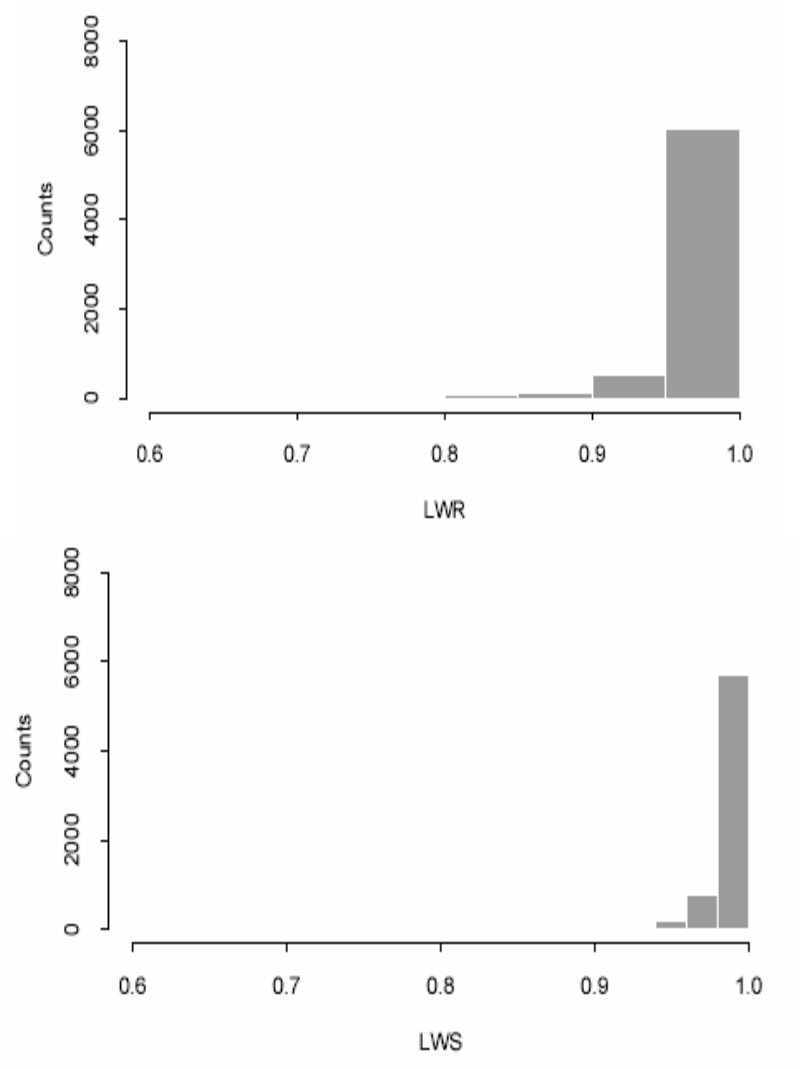

Figure 4: Histograms for correlations between observed and predicted PM intensities.

During the three consecutive days of the experiment, six replicated arrays for each group (Stimulated, 50:50, Starved) were produced. The manufacturing process and analytical methods, including normalization, assure the biological variation among the six independent arrays as low as possible. The variation of the gene expression indexes across the six replicated arrays may serve as a good statistic for comparing the two different regression models. A better model should be able to produce a smaller variation of the gene expression indexes among the six duplicates. In Table 1, the simple descriptive statistics of the sample variances of the expression indexes among the six arrays in each condition (Stimulated, 50:50, Starved) are given to compare LWR with LWS. The result shows that the relationship generally holds that $\mathrm{Var}(\hat{q}_{\mathrm{LWS}}) < \mathrm{Var}(\hat{q}_{\mathrm{LWR}})$. In the Stimulated and 50:50 conditions, LWS yields much smaller variation among the six replicated arrays than LWR, while LWR and LWS perform roughly the same at the Starved condition. In other words, LWS gives more stable results such that the expression indexes from the six arrays in each condition (Stimulated, 50:50 and Starved) have a smaller variation than LWR.

Table 1: Descriptive statistics of sample variances among six arrays at each condition.

	Stimulated		50:50		Starved	
	LWR	LWS	LWR	LWS	LWR	LWS
Minimum	143.620	37.663	39.357	77.091	29.797	8.655
Maximum	3.029e7	3.653e7	1.210e8	8.310e7	6.217e7	4.129e7
Median	89881.3	76892.1	138872.6	118526.1	98829.2	88985.3
Mean	384779.9	369334.5	414866.1	395693.4	327800.8	329822.8

Assessing Gene Expression Measurements

In the experiment, the genes *Lys* and *Phe* were not spiked in starved samples, while *Dap* and *Thr* were not in stimulated sample. Therefore, 12 probe sets and 18 samples of the four spiked-in genes are known to be expressed or not in advance. Totally 144 probe sets should be expressed and 72 should be unexpressed. We obtained the number of expressed and unexpressed genes using the criterion of $\hat{q}/\text{S.E.}(\hat{q}) > 6.0$. The two methods (LWR and LWS) can detect the same number of expressed probe sets (132) and unexpressed probe sets (66). However, regarding the median standard error of the control probe sets, LWS gives a much smaller variation (S.E. of 177.2) of the estimated expression indexes than LWR (S.E. of 307.9). Hence, LWS is more reliable and stable for the estimation of the gene expression indexes.

Focusing on the four spiked-in genes, each gene known to be unexpressed should have a rank as low as possible among all the control genes. One probe set of *Thr* in a Stimulated condition that should be unexpressed has a unexpectedly high expression level. It is considered as an outlier and left out from our analysis. After averaging the expression indexes of each spiked-in probe set over their own six replicated arrays and calculating their ranks, the results are shown in Table 2. The ranks of the 11 unexpressed probe sets are listed with respect to the two models. The comparison between LWR and LWS based on the ranks is summarized with descriptive statistics as follows: LWS has the smaller median rank (6) and the smaller sum of ranks (68) with the smaller variance (13) while LWR has the median rank (8) and the sum of ranks (82) with the variance (17), respectively.

Table 2: Ranks of unexpressed genes among the control genes.

	Dap1	*Dap2*	*Dap3*	*Lys1*	*Lys2*	*Lys3*	*Phe1*	*Phe2*	*Phe3*	*Thr1*	*Thr2*
LWR	2	1	6	13	12	11	10	9	7	8	3
LWS	2	3	10	9	6	1	8	13	4	7	5

Moreover, we examined the ranks of the 11 probe sets of unexpressed control genes among all the genes in our study. Because we put no RNAs for these 11 probe sets, their measured expression levels should be close to zero and their ranks among all the genes should be among the lowest. As shown in Table 3, the ranks of the 11 probe sets detected from LWS are much lower than those from LWR. In summary, LWS has the median rank (19) and the sum of ranks (312) with the variance (979) while LWR has the median rank (99) and the sum of ranks (2482) with the variance (79754), respectively. Based on the ranks of expression levels of the unexpressed control genes, LWS gives much better results than LWR.

Table 3: Ranks of unexpressed genes among all the genes in the study.

	Dap1	*Dap2*	*Dap3*	*Lys1*	*Lys2*	*Lys3*	*Phe1*	*Phe2*	*Phe3*	*Thr1*	*Thr2*
LWR	14	6	33	821	616	472	213	152	36	99	20
LWS	15	16	28	22	19	14	21	122	17	20	18

Among the spiked-in genes, *Dap* and *Thr* in 50:50 samples obtained 0.04 ng/8μg total RNA, 0 in stimulated and 0.08 for starved samples, while *Lys* and *Phe* in 50:50 samples obtained 0.04, 0.8 in stimulated and 0 for starved samples. Better gene expression index estimates should have the ability of differentiating between samples in which the underlying true gene expression levels vary. Hence, a sensible criterion is to assess an estimated expression index according to its correlation with the underlying true expression.

Intuitively, the true expression index should be proportional to the mRNA concentration. Thus higher correlation between the estimated expression indexes and mRNA concentrations is expected if the indexes are closer to the true expression levels. The correlation from LWR is 0.608 and from LWS is 0.609 where LWS is slightly higher than LWR. Similar results are obtained from the study of the correlations among the hybridization genes (*BioB, BioC, BioD, Cre*) and quantities of mRNA (2.5, 5, 25, 100).

To this end, we have made comparisons between the two regression models from several different perspective. LWR is a parametric regression model while LWS is a semiparametric model that is more robust in terms of model mis-specification.

Meanwhile, we notice that LWS gives slightly lower estimation of the expression indexes than LWR does generally. To compare LWR and LWS by combining the mean and variance of the expression indexes, we order all the measures and divide them into 50 quantile groups, then compute the median coefficient of variation (C.V.) for each group. Based on this criterion, LWS gives the average of all the median C.V.'s (0.088), which is smaller than that from LWR (0.094). Figure 5 shows a global and clear picture of the comparison. The median C.V. for each of the 50 groups from LWS is plotted against those from LWR. The straight line is the reference line with unit slope through the origin. It can be seen that most points in the square are above the reference line which indicates that the C.V.'s from LWS are smaller than those from LWR in general.

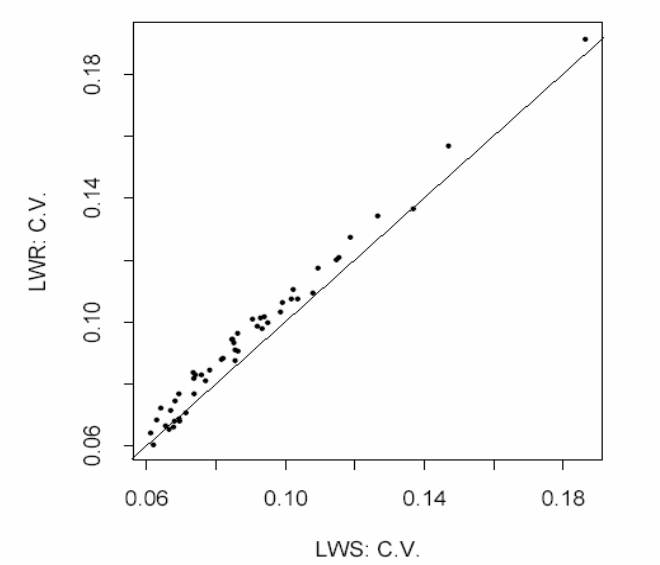

Figure 5: Comparison of coefficients of variation (C. V.) between LWR and LWS.

Conclusion

Recently, much effort has been devoted to obtaining good estimates of the gene expression indexes, where Li-Wong's reduced model is widely used in applications. In this paper, we have proposed a semiparametric model based on Li-Wong's reduced model. The cubic spline smoothing technique allows a flexible functional form for MM expression intensities. Hence, it offers a better model-fitting procedure and captures the important gene expression patterns that might be missed by Li-Wong's reduced model.

From several aspects of comparison, our proposed model outperforms Li-Wong's reduced model. Practically and statistically, our new model is meaningful and easy to implement as well. The reason that we compare the proposed model with Li-Wong's reduced model is that the latter is very popular in practice and proved to perform better than the average expression index, the log-transformed average expression and others.

It is of interest to compare the proposed model with the new Affymetrix MAS 5.0 algorithm and other approaches. The variation of expression indexes changes positively with the intensity, which suggests a certain correlation or a linear trend between them. From the biological point of view, the genes are not independent, especially those that co-regulate. However, so far almost all model-based methods assume the variation has an independent structure. Therefore, a new methodology to incorporate the correlation structures needs to be developed.

For the comparison of measurements, we have extensively utilized the control genes which provided important and helpful information to our study. Control genes can also be used for normalization (Lemon et al., 2002). Hence if possible, we suggest that more control genes, especially those with more replicates should be used under reasonable biological consideration.

As to the model goodness-of-fit, there is no standard criteria available to justify and compare models with regard to the gene expression data where further research is needed. In the proposed model, the cubic spline smoothing is used, while the kernel smoothing (Speckman, 1988) and other nonparametric techniques may be applied to fit MM intensities as well. The proposed method can be improved in an adaptive way as follows. We first test the goodness of fit of LWR based on the likelihood ratios. If there is no enough evidence to reject LWR, we would accept the estimates ($\hat{q}$ and $\hat{f}$) from LWR, otherwise we would proceed to LWS (spline).

References

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing.* (2nd Ed). New York: Marcel Dekker.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society*, *B 48*, 244-248.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models.* London: Chapman and Hall.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis K. J., Scherf, U. & Speed, T. P. (2002). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. (Accepted for publication in Biostatistics).

Lipshutz R., Fodor S., Gingeras T., & Lockhart D. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics*, *21*, 20-24.

Lemon W. J., Palatini J., Krahe R., & Wright F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics*, *18*, 1470-1476.

Li, C., & Wong W. H. (2001). *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.* Proceedings of the National Academy of Science, USA, *98*, 31-36.

Li, C., & Wong W. H. (2001). Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biology* , *2*, 1-11.

Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., & Horton, H. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, *14*, 1675-1680.

Schadt E., Li C., Ellis B., & Wong W. H. (2002). Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, *84*, S37, 120-125.

Schadt E., Li C., Su C. &, Wong W. H. (2001). Analyzing high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, *80*, 192-202.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*, *B 47*, 1-52.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society*, *B 50*, 413-436.

Appendix: Experiment design float chart