A Two-Way Semi-Linear Model for Normalization and Significant Analysis of cDNA Microarray Data

Jian Huang[1], Deli Wang[2], and Cun-Hui Zhang[3]

1:Dpeartment of Statistics and Actuarial Science, and Program in Public Health Genetics, Univeristy of Iowa, Iowa City, Iowa, 52242 (Email: jian@stat.uiowa.edu)

2:Department of Biostatistics, and Program in Public Health Genetics, Univeristy of Iowa, Iowa City, Iowa, 52242 (Email: deli-wang@uiowa.edu)

3:Department of Statistics, Rutgers University, Piscataway, NJ 08855 (Email: cunhui@stat.rutgers.edu)

January 2004

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 326

ABSTRACT A basic question in analyzing cDNA microarray data is normalization. The purpose of normalization is to remove systematic bias in the observed expression values by establishing a normalization curve across the whole dynamic range. A proper normalization procedure ensures that the normalized intensity ratios provide meaningful measures of relative expression levels. We propose a two-way semi-linear model (TW-SLM) for normalization and significant analysis of microarray data. This method does not make the usual assumptions underlying some of the existing methods. For example, it does not assume that: (i) the percentage of differentially expressed genes is small; or (ii) there is symmetry in the expression levels of up- and down-regulated genes, as required in the *lowess* normalization method. The TW-SLM also naturally incorporates uncertainty due to normalization into significant analysis of microarrays. We use a semiparametric approach based on polynomial splines in the TW-SLM to estimate the normalization curves and the normalized expression values. We also conduct simulation studies to evaluate the TW-SLM method and illustrate the proposed method using a published microarray data set.

KEY WORDS: differentially expressed genes; microarray; high-dimensional data; semiparametric regression; spline; analysis of variance; noise level; variance estimation.

1 Introduction

Microarray technology has become a useful tool for quantitatively monitoring gene expression patterns and has been widely used in functional genomics (Schena et al., 1995; Brown and Botstein, 1999). In a cDNA microarray experiment, cDNA segments representing the collection of genes and expression sequence tags (ESTs) to be probed are amplified by PCR and spotted in high density on glass microscope slides using a robotic system. Such slides are called microarrays. Each microarray contains thousands of reporters of the collection of genes or ESTs. The microarrays are queried in a co-hybridization assay using two fluorescently labeled biosamples prepared from the cell populations of interest. One sample is labeled with fluorescent dye Cy5 (red), and another with fluorescent dye Cy3 (green). Hybridization is assayed using a confocal laser scanner to measure fluorescence intensities, allowing simultaneous determination of the relative expression levels of all the genes represented on the slide (Hedge, Qi, Abernathy, Gay, Dharap, Gaspard, Earle-Hughes, Snesrud, Lee and Quackenbush 2000).

A basic question in analyzing cDNA microarray data is normalization. The purpose of normalization is to remove systematic bias in the observed expression values by establishing a normalization curve across the whole dynamic range. A proper normalization procedure ensures that the normalized intensity ratios provide meaningful measures of relative expression levels. Normalization is needed because many factors, including differential efficiency of dye incorporation, difference in the amount of RNA labeled between the two channels, uneven hybridizations, differences in the printing pin heads, among others, may cause bias in the observed expression values. Therefore, proper normalization is a critical component in the analysis of microarray data and can have important impact on higher level analysis such as detection of differentially expression genes, classification, and cluster analysis.

Yang, Dudoit, Luu, and Speed (2001) systematically considered several normalization methods such as global, intensity-dependent, and dye-swap normalization. The global normalization method assumes a constant normalization factor for all the genes and re-scales the red and green channel intensities so that the mean or median of the intensity log-ratios is zero. For intensity-dependent normalization, Yang et al. proposed using the locally weighted linear scatterplot

smoother (*lowess*, Cleveland 1979) in the scatter plot of log-intensity ratio versus log-intensity product (the M-A plot) and uses the resulting residuals as the normalized log-intensity ratios. The analysis of variance (ANOVA) method (Kerr, Martin, and Churchill 2000) and the mixed linear model method (Wolfinger, Gibson, Wolfinger, Bennett, Hamadeh, Bushel, Afshari, and Paules 2001) takes into account array and dye effects among others in a linear model framework, and assumes constant normalization factors. Fan, Tam, Woude, and Ren (2003) considered a Semi-Linear In-slide Model (SLIM) method using within-array replications. The SLIM method requires replication of a subset of the genes in an array. If the number of replicated genes is small, the expression values of the replicated genes may not cover the entire dynamic range or reflect spatial variation in an array. Park, Yi, Kang, Lee, Lee, and Simon (2003) conducted comparisons of a number of normalization methods, including global, linear and *lowess* normalization methods. All the methods described above, except the ANOVA method, treat normalization as a step separated from the subsequent significant analysis, and the variation due to normalization is not taken into account.

The *lowess* normalization is one of the most widely used normalization methods. It assumes that at least one of the two biological assumptions is satisfied: (i) the proportion of differentially expressed genes should be small, or (ii) there is symmetry in the expression values between up and down regulated genes. These two assumptions reduce the possibility that the differentially expressed genes are incorrectly "normalized." For experiments where these two assumptions are violated, the *lowess* normalization method is not appropriate. Yang et al. (2001) suggested using dye-swap normalization. This approach makes the assumption that the normalization curves in the two dye-swaped slides are the same. Because of slide-to-slide variation, this assumption may not always be satisfied. To alleviate the dependence of the *lowess* normalization method on the assumption (i) or (ii) stated above, Tseng, Oh, Rohlin, Liao, and Wong (2001) proposed using a rank based procedure to first select a set of *invariant genes* that are likely to be constantly expressed, and then carrying out *lowess* normalization using this set of genes. However, they pointed out that the set of selected genes may be relatively small and not cover the whole dynamic range of the expression values, and extrapolation is needed to fill in the gaps that are not covered

by the invariant genes.

We propose a two-way semi-linear model (TW-SLM) for normalization of cDNA microarray data. This model is motivated in part by examining the *lowess* normalization from the semiparametric regression point of view. The proposed TW-SLM normalization method does not make the assumptions underlying the *lowess* normalization method, nor does it require pre-selection of invariant genes or replicated genes in a array. The TW-SLM also provides a framework for incorporating variability due to normalization into significant analysis of microarray data. Below, we first describe the TW-SLM for microarray data. In Section 3, we describe a Gauss-Seidel algorithm for computing the normalization curves and the estimated relative expression levels based on the TW-SLM model. In Section 4, we present a method for detecting differentially expressed genes based on the TW-SLM. In section 5, we illustrate the proposed method by an example and use simulation to compare the proposed method with the *lowess* method. Some concluding remarks are given in Section 6.

2 A two-way semi-linear model for microarray data

To motivate the proposed TW-SLM model for normalization, we first give a description of the *lowess* normalization method from the semiparametric regression point of view. Because the proposed TW-SLM can be considered as an extension of the standard semiparametric regression model (SRM) of Engel, Granger, Rice, and Weiss (1986), we also give a brief description of this model.

2.1 The *lowess* normalization

Suppose that there are J genes and n arrays in the study and that each gene is spotted once in a array. Let u_{ij} and v_{ij} be the intensity levels of gene j in array i from the type 1 and type 2 samples, respectively. Following Chen et al. (1997) and Yang et al. (2001), let y_{ij} be the log-intensity ratio of the jth gene in the ith array, and let x_{ij} be the corresponding average of the log-intensity. That

is,

$$y_{ij} = \log_2 \frac{u_{ij}}{v_{ij}}, \;\; x_{ij} = \frac{1}{2}\log_2(u_{ij}v_{ij}), \; i = 1, \ldots, n, j = 1, \ldots, J. \tag{1}$$

For array $i, i = 1, \ldots, n$, the *lowess* normalization fits the nonparametric regression

$$y_{ij} = f_i(x_{ij}) + \varepsilon^*_{ij}, \; j = 1, \ldots, J. \tag{2}$$

using Cleveland's *lowess* method. Let $\widehat{f_i}$ be the *lowes* estimator of f_i, and let the residuals from the nonparametric curve fitting be

$$\widehat{\varepsilon}^*_{ij} = y_{ij} - \widehat{f_i}(x_{ij}), \; i = 1, \ldots, n, j = 1, \ldots, J.$$

These residuals are defined as the normalized data and used as the input in the subsequent analysis. So usually the overall analysis consists of two steps: (i) normalization; and (ii) analysis based on normalized data $\widehat{\varepsilon}^*_{ij}$. For example, in comparing two DNA samples using a direct comparison design (i.e., the two cDNN samples are competitively hybridized on an array), a typical approach is to first normalize the data using the *lowess* normalization, and then to make inference about differentially expressed genes based on the normalized data. The underlying statistical framework of such a two-step analysis in the direct comparison design can be described using two models. The first is the nonparametric regression for normalization given in (2). The second model concerns the residual:

$$\varepsilon^*_{ij} = \beta_j + \varepsilon_{ij}, \tag{3}$$

where β_j is the underlying relative expression value of gene j. The goal of the significance analysis is to detect genes with $\beta_j \neq 0$. In the two-step approach, (2) and (3) are used as stand-alone models for each of the two steps, and the effects of the approximation $\widehat{\varepsilon}^*_{ij} \approx \varepsilon^*_{ij}$ are typically completely ignored in the analysis.

The *lowess* normalization is usually applied using all the genes in a study. In general, if all the genes are used, the differentially expressed genes may be incorrectly "normalized," since such genes tend to pull the normalization curve towards themselves. Thus the two-step analysis approach may yield biased estimators of both f_i and β_j and inefficient test statistics for the inference of β_j (e.g. relatively large p-values for two-sided tests compared with more efficient procedures).

2.2 The semiparametric regression model

Suppose that the data consists of n triplets $(y_i, x_i, z_i), i = 1, \ldots, n$, where y_i is the response variable, and (x_i, z_i) is the covariate. The SRM is

$$y_i = f(x_i) + z_i'\beta + \varepsilon_i,\ i = 1, \ldots, n, \tag{4}$$

where f is an unknown function, β is the regression parameter, and ε_i is the residual. This model is useful in many situations, for example, when z_i is a dichotomous variable representing two conditions (treatment versus placebo etc.) and we are interested in the treatment effect β but need to adjust for the effect of the continuous covariate x_i. For a p-dimensional covariate $x_i = (x_{i1}, \ldots, x_{ip})'$, it is useful to impose an additive structure on f (Hastie and Tibshirani 1990). A semiparametric generalized additive model is

$$y_i = f_1(x_{1i}) + \cdots f_p(x_{pi}) + z_i'\beta + \varepsilon_i,\ i = 1, \ldots, n, \tag{5}$$

Models (4) and (5) are two basic semiparametric models. There are two important considerations about parameter estimation in (4) and (5). First, both f and β should be estimated jointly. For example, it is incorrect to fix β at 0, obtain an estimate of f, then treat this estimate of f as a known quantity, substitute it back into (4), and then estimate β. Second, the uncertainty due to estimation of f generally needs to be taken into account in estimating β, according to the semiparametric information theory, see for instance, Bickel, Klaassen, Ritov, and Wellner (1993), Example 5, pages 107-109.

2.3 The two-way semi-linear model

We first describe the proposed model for the special case of a direct comparison design, in which two cDNA samples from the respective cell populations are competitively hybridized on the same array. Let y_{ij} and x_{ij} be the log-intensity ratio and product defined in (1). The proposed TW-SLM is

$$y_{ij} = f_i(x_{ij}) + \beta_j + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, J, \tag{6}$$

where f_i is the intensity-dependent normalization curve for the ith array, $\beta_j \in R$ represents the normalized relative expression values of gene j, and ε_{ij} has mean 0 and variance σ_{ij}^2.

The TW-SLM can be considered as a combination of the two models that are implicitly used in the *lowess* normalization (2) and (3). Specifically, we obtain (6) by simply substituting (3) into (2). Combining these two models enables us to estimate normalization curves and gene effects simultaneously. This is desirable, since we typically do not know which genes are constantly expressed (i.e., with $\beta_j = 0$). Approximately unbiased normalization could be carried out using only constantly expressed genes if a large set of such genes can be identified, but this is rarely the reality.

We call (6) a two-way model because it also can be considered as a semiparametric generalization of the two-way ANOVA model. That is, when $f_i = \alpha_i, i = 1, \ldots, n$, where α_i is a constant parameter, (6) simplifies to the two-way ANOVA. The TW-SLM is an extension of but different from the standard semiparametric regression model (4). Clearly it is also different from the semiparametric generalized additive model (5). In particular, in models (4) and (5), the number of finite- and infinite-dimensional parameters is fixed and is independent of the sample size, and they do not include the standard two-way ANOVA as a submodel. In contrast, in the TW-SLM, the number of finite-dimensional parameters is J, which is the sample size for estimating f_i, and the number of infinite-dimensional parameters is n, which is the sample size for estimating β_j.

In general, let $z_i \in R^d$ be a covariate vector associated with the ith array. The general form of the TW-SLM is:

$$y_{ij} = f_i(x_{ij}) + z_i'\beta_j + \varepsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, J, \tag{7}$$

where $\beta_j \in R^d$ is the effect associated with the jth gene, and where f_i and ε_{ij} are the same as in (6).

The covariate vector z_i can be used to code various design schemes, such as the loop, reference, and factorial designs (Kerr and Churchill 2001). For example, for the two-sample direct comparison design, $z_i = 1, i = 1, \ldots, n$, which is model (6). For an indirect comparison design using a common reference, we can introduce a two-dimensional covariate vector $z_i = (z_{i1}, z_{i2})'$. Let $z_i = (1, 0)'$ if the ith array is of the type 1 sample versus the reference, and $z_i = (0, 1)'$ if the ith

array is of the type 2 sample versus the reference. Now $\beta_j = (\beta_{j1}, \beta_{j2})'$ is a two-dimensional vector and $\beta_{j1} - \beta_{j2}$ represents the difference in the expression levels of gene j after normalization. The covariate vector z_i can also include other factors that contribute to the variations of the observed expression values.

In model (7), it is only made explicit that the normalization curve f_i is array-dependent. It is straightforward to extend the model so that f_i also depends on the printing-pin blocks within an array. Specifically, suppose that in each array, there are K printing-pin blocks, and in the kth block, there are J_k genes printed. Let y_{ikj} and x_{ikj} be the log-intensity ratio and log-intensity product of gene j in the kth block of array i, respectively. The model can be written as

$$y_{ikj} = f_{ik}(x_{ikj}) + z_i\beta_j + \varepsilon_{ikj}, \tag{8}$$

$j = 1, \ldots, J_k, i = 1, \ldots, n$, and $k = 1, \ldots, K$.

The TW-SLM also can be easily extended to accommodate the design where a gene is printed multiple times in an array. Such a design is helpful for improving the precision and for assessing the quality of an array using the coefficient of variation (Tseng et al. 2001). Suppose that on the ith array, in the kth printing-pin block, there are J_k genes and the jth gene in this block is printed R_{ijk} times. The TW-SLM can be written as

$$y_{ikjr} = f_{ik}(x_{ikjr}) + z_i\beta_j + \varepsilon_{ikjr}, \tag{9}$$

$r = 1, \ldots, R_{ijk}$, $i = 1, \ldots, n, j = 1, \ldots, J_k$, and $k = 1, \ldots, K$.

We note that we can also adapt the TW-SLM to other designs and incorporate spiked control genes in the TW-SLM. Such genes are often used for the purpose of calibration and normalization in a microarray experiment.

3 TW-SLM normalization

We now define the semiparametric least squares estimator (SLSE) in TW-SLM and describe an algorithm for computing the estimated normalization curves and gene expression values using the TW-SLM. Many nonparametric smoothing procedures can be used for this purpose. We use the

method of polynomial splines (Schumaker 1981). This method is easy to implement, and has similar performance as other nonparametric curve estimation methods such as local polynomial regression and smoothing splines (Hastie, Tibshirani, and Friedman 2001).

3.1 Semiparametric LS estimator in TW-SLM

Let $\Omega_0^{J\times d}$ be the space of all $J \times d$ matrices $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_J)'$ satisfying $\sum_{j=1}^J \beta_j = 0$. It is clear from the definition of the TW-SLM model (7) that $\boldsymbol{\beta}$ is identifiable only up to a member in $\Omega_0^{J\times d}$, since we may simply replace β_j by $\beta_j - \sum_{k=1}^J \beta_k/J$ and $f_i(x)$ by $f_i(x) + \sum_{k=1}^J \beta_k' z_i/J$ in (7). In what follows, we assume

$$\boldsymbol{\beta} \in \Omega_0^{J\times d} \equiv \Big\{\boldsymbol{\beta} : \sum_{j=1}^J \beta_j = 0\Big\}. \tag{10}$$

Let $b_{i1}, \ldots, b_{i,K_i}$ be K_i B-spline base functions. Let

$$S_i \equiv \overline{\{b_{i0}(x) \equiv 1, b_{ik}(x), k = 1, \ldots, K_i\}} \tag{11}$$

be the spaces of all linear combinations of the basis functions. We approximate f_i by

$$\alpha_{i0} + \sum_{k=1}^{K_i} b_{ik}(x)\alpha_{ik} \equiv \boldsymbol{b}_i(x)'\alpha_i, \in S_i$$

where $\boldsymbol{b}_i(x) = (1, b_{i1}(x), \ldots, b_{i,K_i}(x))'$, and $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iK_i})'$ are coefficients to be estimated from the data. Let $\boldsymbol{f} = (f_1, \ldots, f_n)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$. The LS objective function is

$$D^2(\boldsymbol{\beta}, \boldsymbol{f}) = \sum_{i=1}^n \sum_{j=1}^J [y_{ij} - f_i(x_{ij}) - z_i'\beta_j]^2.$$

We define the semiparametric least squares estimator (SLSE) of $\{\boldsymbol{\beta}, \boldsymbol{f}\}$ to be the $\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{f}}\} \in \Omega_0^{J\times d} \times \prod_{i=1}^n S_i$ that minimizes $D^2(\boldsymbol{\beta}, \boldsymbol{f})$. That is,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{f}}) = \underset{(\boldsymbol{\beta},\boldsymbol{f})\in \Omega_0^{J\times d}\times\prod_{i=1}^n S_i}{\arg\min} D^2(\boldsymbol{\beta}, \boldsymbol{f}). \tag{12}$$

Let $B_{ij} = (1, b_{i1}(x_{ij}), \ldots, b_{i,K_i}(x_{ij}))'$ be the spline basis functions evaluated at $x_{ij}, 1 \le i \le$

$n, 1 \leq j \leq J$. The spline basis matrix for the ith array is

$$B_i = \begin{pmatrix} B_{i1}' \\ \vdots \\ B_{iJ}' \end{pmatrix} = \begin{pmatrix} 1 & b_{i1}(x_{i1}) & \dots & b_{iK_i}(x_{i1}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & b_{i1}(x_{iJ}) & \dots & b_{iK_i}(x_{iJ}) \end{pmatrix}.$$

Let $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)'$. We can write $D^2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = D^2(\boldsymbol{\beta}, \boldsymbol{f})$. Then the problem of minimizing $D^2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is equivalent to solving the linear equations:

$$\widehat{\boldsymbol{\beta}} \sum_{i=1}^{n} (z_i z_i') + \sum_{i=1}^{n} B_i \widehat{\alpha}_i z_i' = \sum_{i=1}^{n} \boldsymbol{y}_i z_i', \quad B_i B_i' \widehat{\alpha}_i + B_i' \widehat{\boldsymbol{\beta}} z_i = B_i' \boldsymbol{y}_i.$$

Let $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$ be the solution. We define $\hat{f}_i(x) \equiv \boldsymbol{b}_i(x)' \widehat{\alpha}_i, i = 1, \dots, n$.

3.2 Computation

Our approach for minimizing $D^2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is to use the Gauss-Seidel method, also called the back-fitting algorithm (Hastie, Tibshirani, and Friedman 2001), that alternately updates $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Set $\boldsymbol{\beta}^{(0)} = \mathbf{0}$. For $k = 0, 1, 2, \dots$,

Step 1: Compute $\boldsymbol{\alpha}^{(k)}$ by minimizing $D^2(\boldsymbol{\beta}^{(k)}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$. The explicit solution is

$$\alpha_i^{(k)} = (B_i' B_i)^{-1} B_i' (\boldsymbol{y}_i - \boldsymbol{\beta}^{(k)} z_i), i = 1, \dots, n.$$

Step 2: Given the $\boldsymbol{\alpha}^{(k)}$ computed in Step 1, let $f_i^{(k)}(x) = \boldsymbol{b}_i(x)' \alpha_i^{(k)}$, compute $\boldsymbol{\beta}^{(k+1)}$ by minimizing $D_w(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(k)})$ with respect to $\boldsymbol{\beta}$. The explicit solution is

$$\widehat{\beta}_j^{(k+1)} = \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} \sum_{i=1}^{n} z_i \Big(y_{ij} - f_i^{(k)}(x_{ij}) \Big), j = 1, \dots, J. \tag{13}$$

Iterate between Steps 1 and 2 until the desired convergence criterion is satisfied. Because the objective function is strictly convex, the algorithm converges to the sum of residual squares. Suppose that the algorithm meets the convergence criterion at step K. Then the estimated values of β_j are $\widehat{\beta}_j = \beta_j^{(K)}, j = 1, \dots, J$, and the estimated normalization curves are

$$\widehat{f}_i(x) = \boldsymbol{b}_i(x)' \alpha_i^{(K)} = \boldsymbol{b}_i(x)' (B_i' B_i)^{-1} B_i' (\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i), \ i = 1, \dots, n. \tag{14}$$

The algorithm described above can be conveniently implemented in the statistical computing environment R (Ihaka and Gentleman, 1996). Specifically, Steps 1 and 2 can be solved by the function lm in R. The function bs can be used to create a basis matrix for the polynomial splines.

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})'$ and $f_i(\boldsymbol{x}_i) = (f_i(x_{i1}), \ldots, f_i(x_{iJ}))'$. Let $Q_i = B_i(B_i'B_i)^{-1}B_i'$. By (14), the estimator of $f_i(\boldsymbol{x}_i)$ is

$$\widehat{f}_i(\boldsymbol{x}_i) = Q_i(\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i).$$

Thus the normalization curve is the result of the linear smoother Q_i operating on $\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i$. The gene effect $\widehat{\boldsymbol{\beta}} z_i$ is removed from $\boldsymbol{y}_i$. In comparison, the *lowess* normalization method does not remove the gene effect. An analogue of the *lowess* normalization, but using polynomial splines, is

$$\widetilde{f}_i(\boldsymbol{x}_i) = Q_i \boldsymbol{y}_i = B_i \boldsymbol{\alpha}_i^{(0)}.$$

Comparing $\widehat{f}_i(\boldsymbol{x}_i)$ with $\widetilde{f}_i(\boldsymbol{x}_i)$, if there is a relatively large percentage of differentially expressed genes, the difference between this two normalization curves can be large. The magnitude of the difference also depends on the magnitude of the gene effects.

4 TW-SLM for significant analysis of microarray data

In addition to being a stand-alone model for normalization, the TW-SLM can also be naturally used for detection of differentially expressed genes. Indeed, the Gauss-Seidel algorithm described above already yields an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, which represents gene effects. Therefore, for the purpose of making inference about $\boldsymbol{\beta}$, we need to estimate the variance of $\widehat{\boldsymbol{\beta}}$. Below, we first consider the structure of $\widehat{\boldsymbol{\beta}}$, and then describe an intensity dependent variance estimator.

4.1 Structure of the semiparametric LS estimator

We give the expression of $\widehat{\boldsymbol{\beta}}$ and define the observed information matrix for $\boldsymbol{\beta}$ in the presence of the normalization curves $f_i, i = 1, \ldots, n$. Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})', \boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$ and $f(\boldsymbol{x}_i) \equiv (f(x_{i1}), \ldots, f(x_{iJ}))'$ for a univariate function f. We write the TW-SLM (7) in vector

notation as

$$\boldsymbol{y}_i = \boldsymbol{\beta} z_i + f_i(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{15}$$

Using (15), it can be shown that the SLSE (12) equals

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - (I_J - Q_i)\boldsymbol{\beta} z_i \right\|^2. \tag{16}$$

In the special case of model (6), $d = 1$ (scalar β_j) and $\boldsymbol{\beta}$ is a vector in $\mathbb{R}^J$, (16) is explicitly

$$\widehat{\boldsymbol{\beta}} = \widehat{\Lambda}_{J,n}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i) \boldsymbol{y}_i z_i' \right), \tag{17}$$

since $I_J - Q_i$ are projections in $\mathbb{R}^J$, where $z_i = 1$ (scalar) and, where

$$\widehat{\Lambda}_{J,n} \equiv \frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i). \tag{18}$$

We note that $\widehat{\Lambda}_{J,n}$ can be considered as the observed information matrix. Here and below, A^{-1} denotes the generalized inverse of matrix A, defined by $A^{-1}\boldsymbol{x} \equiv \arg\min \{ \|\boldsymbol{b}\| : A\boldsymbol{b} = \boldsymbol{x} \}$. If A is a symmetric matrix with eigenvalues λ_j and eigenvectors $\mathbf{v}_j$, then $A = \sum_j \lambda_j \mathbf{v}_j \mathbf{v}_j'$ and $A^{-1} = \sum_{\lambda_j \neq 0} \lambda_j^{-1} \mathbf{v}_j \mathbf{v}_j'$.

For general z_i and $d \geq 1$, (16) is still given by (17) with

$$\widehat{\Lambda}_{J,n} \equiv \frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i) \otimes z_i z_i'. \tag{19}$$

The information operator (19) is an average of tensor products, i.e. a linear mapping from $\Omega_0^{J \times d}$ to $\Omega_0^{J \times d}$ defined by $\widehat{\Lambda}_{J,n}\boldsymbol{\beta} \equiv n^{-1} \sum_{i=1}^{n} (I_J - Q_i)\boldsymbol{\beta} z_i z_i'$.

From the expression of $\widehat{\boldsymbol{\beta}}$ given in (17), we see that, because Q_i is a linear smoother, $Q_i \boldsymbol{y}_i$ is an estimated curve through the M-A plot in the ith array, and $(I_J - Q_i)\boldsymbol{y}_i = \boldsymbol{y}_i - Q_i \boldsymbol{y}_i$ is the residual from this estimated curve. In the *lowess* normalization method, such residuals are used as the normalized data, except that there the local regression smoother is used instead of polynomial splines. In the proposed TW-SLM method, the normalized data for the ith array is

$$\widehat{\Lambda}_{J,n}^{-1} (I_J - Q_i)\boldsymbol{y}_i = \widehat{\Lambda}_{J,n}^{-1} (\boldsymbol{y}_i - Q_i \boldsymbol{y}_i).$$

The simple residual $\boldsymbol{y}_i - Q_i \boldsymbol{y}_i$ is corrected multiplicatively by the inverse of the information operator $\widehat{\Lambda}_{J,n}$.

4.2 Variance estimation and inference for β

Based on (17), we have, conditional on $\{x_{ij}\}$,

$$Var(\widehat{\boldsymbol{\beta}}) = \widehat{\Lambda}_{J,n}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(I_J - Q_i)Var(\boldsymbol{\varepsilon}_i)(I_J - Q_i)\otimes z_i' z_i\right)\widehat{\Lambda}_{J,n}^{-1}. \quad (20)$$

The variance matrix $Var(\boldsymbol{\varepsilon}_i)$ can be estimated based on the residuals. Therefore, in principle, $Var(\widehat{\boldsymbol{\beta}})$ can be estimated based on (20). However, direct computation involves inverting a $dJ \times dJ$ matrix. When $J = O(10^4)$, as in many microarray experiments, inverting such a large matrix is difficult. Therefore, we derive an approximation to $Var(\boldsymbol{\beta}_j)$ that is easier to compute. Let $Z_n = \sum_{i=1}^{n} z_i z_i'$. Based on (13), we have

$$\begin{aligned} Z_n\widehat{\beta}_j &= \sum_{i=1}^{n} z_i(y_{ij} - \widehat{f}_i(x_{ij})) \quad (21)\\ &= \sum_{i=1}^{n} z_i(\varepsilon_{ij} + z_i'\beta_j + f_i(x_{ij}) - \widehat{f}_i(x_{ij})). \end{aligned}$$

Therefore,

$$Z_n(\widehat{\beta}_j - \beta_j) = \sum_{i=1}^{n} z_i\varepsilon_{ij} + \sum_{i=1}^{n} z_i[f_i(x_{ij}) - \widehat{f}_i(x_{ij})].$$

This leads to:

$$Var(Z_n\widehat{\beta}_j) \approx \sum_{i=1}^{n} z_i z_i' E(\varepsilon_{ij})^2 + \sum_{i=1}^{n} z_i z_i' E[f_i(x_{ij}) - \widehat{f}_i(x_{ij})]^2.$$

So we have

$$\begin{aligned} Var(\widehat{\beta}_j) &\approx Z_n^{-1}\left[\sum_{i=1}^{n} z_i z_i' Var(\varepsilon_{ij})\right] Z_n^{-1} + Z_n^{-1}\left[\sum_{i=1}^{n} z_i z_i' Var(\widehat{f}_i(x_{ij}))\right] Z_n^{-1}\\ &\equiv \Sigma_{\varepsilon,j} + \Sigma_{f,j}. \end{aligned}$$

The variance of $\widehat{\beta}_j$ consists of two components. The first component represents the variation due to the residual errors in the TW-SLM, and the second component is due to the variation in the estimated normalization curves.

For the first term $\Sigma_{\varepsilon,j}$, we have

$$\Sigma_{\varepsilon,j} = Z_n^{-1}\left[\sum_{i=1}^{n} z_i z_i' \sigma_{ij}^2\right] Z_n^{-1}.$$

Suppose that $\hat{\sigma}_{ij}^2$ is a consistent estimator of σ_{ij}^2, which will be given below. We estimate $\Sigma_{\varepsilon,j}$ by

$$\widehat{\Sigma}_{\varepsilon,j} = Z_n^{-1} \left[\sum_{i=1}^{n} z_i z_i' \hat{\sigma}_{ij}^2 \right] Z_n^{-1}.$$

For the second term $\Sigma_{f,j}$, we approximate $\widehat{\boldsymbol{f}}_i$ by the ideal normalization curve, that is,

$$\widehat{f}_i(\boldsymbol{x}_i) = Q_i(\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i)] \approx Q_i(\boldsymbol{y}_i - \boldsymbol{\beta} z_i) = Q_i(\boldsymbol{\varepsilon}_i + f_i(\boldsymbol{x}_i)).$$

Therefore, conditional on $\boldsymbol{x}_i$, we have,

$$Var(\widehat{f}_i(\boldsymbol{x}_i)) \approx Q_i Var(\boldsymbol{\varepsilon}_i) Q_i,$$

and

$$Var(\widehat{f}_i(x_{ij})) = \mathbf{e}_j' Q_i Var(\boldsymbol{\varepsilon}_i) Q_i \mathbf{e}_j,$$

where $\mathbf{e}_j$ is the unit vector whose jth element is 1. Let $\widehat{\Sigma}_i$ be an estimator of $Var(\boldsymbol{\varepsilon}_i)$. We estimate $\Sigma_{f,j}$ by

$$\widehat{\Sigma}_{f,j} = Z_n^{-1} \mathbf{e}_j' \left[\sum_{i=1}^{n} Q_i \widehat{\Sigma}_i Q_i \right] \mathbf{e}_j Z_n^{-1}.$$

Finally, we estimate $Var(\widehat{\beta}_j)$ by

$$\widehat{\Sigma}_{\beta,j} = \widehat{\Sigma}_{\varepsilon,j} + \widehat{\Sigma}_{f,j}. \tag{22}$$

Then a test for the contrast $c'\beta_j$, where c is a known contrast vector, is based on the statistic

$$t_j = \frac{c'\widehat{\beta}_j}{\sqrt{c'\widehat{\Sigma}_{\beta,j} c}}.$$

We use a t distribution with $n-1$ degrees of freedom to approximate the null distribution of t_j when $c'\beta_j = 0$. Resampling methods such as permutation or bootstrap can also be used to evaluate the distribution of t_j.

We now consider two models for σ_{ij}.

(i) The residual variances are different for each gene but do not change across the arrays. That is, for $j = 1, \ldots, J$,

$$\sigma_{ij}^2 = \sigma_j^2, \ i = 1, \ldots, n.$$

We estimate σ_j^2 by

$$\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\hat{\varepsilon}_{ij}^2. \tag{23}$$

One problem with this variance estimation approach is that, because the number of genes in a microarray study is usually large, there may be many small $\hat{\sigma}_j^2$ values just by chance, which can result in large t statistic values even if the differences in expression values are small. One solution to this problem is to add a suitable constant to the value of $\hat{\sigma}_j^2$ (Tusher, Tibshirani, and Chu 2001). However, it is not clear what is the impact of such an adjustment on the false negative rate.

(ii) The residual variances depend smoothly on the total intensity values, and such dependence may vary from array to array. So the model is

$$\sigma_{ij}^2 = \sigma_i^2(x_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, J,$$

where σ_i^2 is a smooth positive function. This model takes into account the possible array to array variations in the variances. Because of the smoothness assumption on σ_i^2, this model says that, in each array, the genes with similar expression intensity values also have similar residual variances. This is a reasonable assumption, for in many microarray data, the variability of the log-intensity ratio depends on the total intensity. In particular, it is often the case that the variability is higher in the lower range of the total intensity than that in the higher range.

We use the method proposed by Ruppert, Wand, Holst, and Hössjet O (1997) and Fan and Yao (1998) in estimating the variance function in a nonparametric regression model. For each $i = 1, \ldots, n$, we fit a smooth curve through the scatter plot $(x_{ij}, \hat{\varepsilon}_{ij}^2)$, where $\hat{\varepsilon}_{ij}^2 = (y_{ij} - \widehat{f}_i(x_{ij}) - z_i'\widehat{\beta}_j)^2$. This is equivalent to fitting the nonparametric regression model

$$\hat{\varepsilon}_{ij}^2 = \sigma_i^2(x_{ij}) + \tau_{ij},\ j = 1, \ldots, J,$$

for $i = 1, \ldots, n$, where τ_{ij} is the residual term in this model. We use the same spline bases as in the estimation of f_i (14). The resulting spline estimator $\hat{\sigma}_i^2$ can be expressed as

$$\hat{\sigma}_i^2(x) = \boldsymbol{b}_i'(x)(B_i'B_i)^{-1}B_i\widehat{\boldsymbol{\varepsilon}}_i^2, \tag{24}$$

where $\widehat{\boldsymbol{\varepsilon}}_i^2 = (\hat{\varepsilon}_{i1}^2, \ldots, \hat{\varepsilon}_{iJ}^2)'$. The estimator of σ_{ij}^2 is then $\hat{\sigma}_{ij}^2 = \hat{\sigma}_i^2(x_{ij})$.

5 An example and simulation studies

5.1 Apo A1 data

We now illustrate the TW-SLM for microarray data by the Apo A1 data set of Callow, Dudoit, Gong, Speed, and Rubin (2000). The purpose of this experiment is to identify differentially expressed genes in the livers of mice with very low HDL cholesterol levels compared to inbred mice. The treatment group consists of 8 mice with the apo A1 gene knocked-out and the control group consists of 8 C57BL/6 mice. For each of these mice, target cDNA is obtained from mRNA by reverse transcription and labeled using a red fluorescent dye (Cy5). The reference sample (green-fluorescent dye Cy3) used in all hybridizations was obtained by pooling cDNA from the 8 control mice. The target cDNA is hybridized to microarrays containing 5,548 cDNA probes. This data set was analyzed by Callow et al. (2000) and Dudoit, Yang, Callow, and Speed (2000). Their analysis uses *lowess* normalization and the two-sample t-statistic. Eight genes with multiple comparison adjusted permutation p-value ≤ 0.01 are identified.

We apply the proposed normalization and analysis method to this data set. As in Dudoit et al. (2000), we use printing-tip dependent normalization. The TW-SLM model used here is

$$y_{ikj} = f_{ik}(x_{ikj}) + z_i'\beta_j + \varepsilon_{ikj},$$

where $i = 1, \ldots, 16, k = 1, \ldots, 16$, and $j = 1, \ldots, 399$. Here i and j index arrays and genes as before, k indexes the printing-tip blocks in an array. ε_{ikj} are residuals with mean 0 and variance σ^2_{ikj}. We use the model

$$\sigma^2_{ikj} = \sigma^2_{ik}(x_{ikj}),$$

where σ^2_{ik} are unknown smooth functions. We apply the printing-pin dependent normalization and estimation approach described in Section 4.2. The covariate $z_i = (1, 0)'$ for the treatment group (apo A1 knock out mice) and $z_i = (0, 1)'$ for the control group (C57BL/6 mice). The coefficient $\beta_j = (\beta_{j1}, \beta_{j2})$. The contrast $\beta_{j1} - \beta_{j2}$ measures the expression difference for the jth gene between the two groups.

As examples of the normalization results, Figure 1 displays the M-A plots and printing-tip dependent normalization curves in the 16 printing-pin blocks of the array from one knock-out mouse. The solid green line is the normalization curve based on the TW-SLM model, and the dashed red line is the *lowess* normalization curve. The degrees of freedom used in the spline basis function in the TW-SLM normalization is 12, and following Dudoit et al. (2000), the span used in the *lowess* normalization is 0.40. We see that, there are differences between the normalization curves based on the two methods. The *lowess* normalization curve attempts to fit each individual M-A scatter plot, without taking into account the gene effects. In comparison, the TW-SLM normalization curves do not follow the plot as closely as the lowess normalization.

Figure 2 displays the volcano plots of $-\log 10$ p-values versus the mean log-differences of expression values between the knock-out and control groups. In the first (left) volcano plot, both the normalization and estimation of $\boldsymbol{\beta}$ are based on the TW-SLM. The variances are estimated based on (24) that assumes that the residual variances depend smoothly on the total log-intensities. The second plot is based on the *lowess* normalization method and use the two-sample t-statistics as in Dudoit et al. (2000), but the p-values are obtained based on Welch's correction for the degrees of freedom. The 8 solid circles are the significant genes that were identified by Dudoit et al. (2000). These 8 genes are also significant based on the proposed method, as can be seen from the volcano plot. Comparing the two volcano plots, we see that the $-\log 10$ p-values based on the TW-SLM method tend to be higher than those based on the *lowess* and the t-test method, as discussed at the end of Section 2.1.

The differences between the two volcano plots are due to different normalization methods and two difference approaches for estimating the variances. We first examine the differences between the TW-SLM normalization values and the *lowess* normalization values. We plot the estimated mean expression differences based on the TW-SLM versus those calculated based on the *lowess* normalization, see Figure 3. The solid line is the fitted linear regression line, which is

$$y = 0.00029 + 1.09x.$$

The standard error of the intercept is 0.0018, so the intercept is negligible. The standard error of the slope is 0.01. Therefore, on average, the mean expression differences based on the TW-SLM

normalization method are about 10% higher than those based on the *lowess* normalization method.

Figure 4 shows the histograms of the standard errors obtained based on intensity-dependent smoothing defined in (24) and the standard errors calculated for individual genes. The standard errors based on the individual genes have a relatively large range of variation, but the range of standard errors based on intensity-dependent smoothing shrinks towards the middle. The SE's based on the smoothing method are more tightly centered around the median value of about 0.13. Thus, the analysis based on the smooth estimate of the error variances is less susceptible to the problem of artificially small p-values resulting from random small $\widehat{\sigma}_j$.

5.2 Simulation studies

We use simulation to compare the TW-SLM and *lowess* normalization methods with regard to mean square errors (MSE) in estimating expression levels β_j. Let α_1 and α_2 be the percentages of up- and down-regulated genes, respectively, and let $\alpha = \alpha_1 + \alpha_2$. We consider four models in our simulation.

Model 1: There is no dye bias. So the true normalization curve is set at the horizontal line at 0. That is $f_i(x) \equiv 0, 1 \leq i \leq n$. In addition, the expression levels of up- and down-regulated genes are symmetric and $\alpha_1 = \alpha_2$.

Model 2: As in Model 1, the true normalization curves $f_i(x) \equiv 0, 1 \leq i \leq n$. But the percentages of up- and down-regulated genes are different. We set $\alpha_1 = 3\alpha_2$

Model 3: There are non-linear and intensity dependent dye biases. The expression levels of up- and down-regulated genes are symmetric and $\alpha_1 = \alpha_2$.

Model 4: There is non-linear and intensity dependent dye bias. The percentages of up- and down-regulated genes are different. We set $\alpha_1 = 3\alpha_2$.

Models 1 and 2 can be considered as baseline ideal case in which there is no channel bias. The data generating process is as follows:

(i) Generate β_j. For most of the genes, we simulate $\beta_j \sim N(0, \tau_j^2)$. The percentage of such genes is $1 - \alpha$. For up-regulated genes, we simulate $\beta_j \sim N(\mu, \tau_{Uj}^2)$ where $\mu > 0$. For down-regulated genes, we simulate $\beta_j \sim N(\mu, \tau_{Dj}^2)$. We use $\tau_j = 0.6, \mu = 2, \tau_{Uj} = \tau_{Dj} = 1$.

(ii) Generate x_{ij}. We simulate $x_{ij} \sim 16 * Beta(a, b)$, where $a = 1, b = 2.5$.

(iii) Generate ε_{ij}. We simulate $\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$, where $\sigma_{ij} = \sigma(x_{ij})$. Here $\sigma(x) = 0.3 * x^{-1/3}$. So the error variance is higher at lower intensity range than at higher intensity range.

(iv) The log-intensity ratios are computed as $y_{ij} = f_i(x_{ij}) + \beta_j + \varepsilon_{ij}$. In Cases 3 and 4, for the ith printing-pin block in an array, we use

$$f_i(x) = \frac{a_{i1} x^2 \sin(x/\pi)}{1 + a_{i2} x^2}, \quad x \in [0, 16],$$

where a_{i1} and a_{i2} are generated independently from the uniform distribution $U(0.6, 1.4)$. Thus the normalization curves vary from block to block within an array and between arrays.

The number of printing-pin blocks is 16, and in each block, there are 400 spots. The number of arrays in each data set is 10. The number of replications for each simulation is 10. Based on these 10 replications, we calculate bias, variance, and mean square error of estimated expression values relative to the generating values. In each of the four cases, we consider three levels of the percentage of differentially expressed genes: $\alpha = 0.01, 0.06, 0.12$.

Tables 1 to 4 give the summary statistics of the MSEs for estimating the relative expression levels β_j in the four models described above. In Table 1 for simulation Model 1, in which the true normalization curve is the horizontal line at 0 and the expression levels of up- and down-regulated genes are symmetric, the TW-SLM normalization tends to have slightly higher MSEs. In Table 2, when there is no longer symmetry in the expression levels of up- and down-regulated genes, the TW-SLM method has smaller MSEs. In Table 3 for simulation Model 3, there is non-linear intensity dependent dye bias, but the percentages of up- and down-regulated genes are the same, the TW-SLM has slightly smaller MSEs. The Table 4 for simulation Model 4, there is non-linear intensity dependent dye bias, and the percentages of up- and down-regulated genes are different, the TW-SLM has considerably smaller MSEs. We have also examined biases and variances. There are only small differences in variances between the TW-SLM and *lowss* methods. However, the biases of the *lowess* method are generally higher.

6 Discussion

We have proposed a TW-SLM for normalization and significant analysis of microarray data. The basic idea of TW-SLM normalization is to estimate the normalization curves and the relative gene effects simultaneously. The TW-SLM normalization does not assume that the normalization is constant as in the global normalization method, nor does it make the assumptions that the percentage of differentially expressed genes is small or that the up- and down-regulated genes are distributed symmetrically, as are required in the *lowess* normalization method (Yang et al. 2001). This model puts normalization and significant analysis of gene expression in the framework of a high dimensional semiparametric regression model. We used the Gauss-Seidel algorithm to compute the semiparametric least squares estimates of the normalization curves using polynomial splines and the gene effects. For identification of differentially expressed genes, we used an intensity-dependent variance model, and applied the nonparametric regression method based on squared residuals (Ruppert et al. 1997, Fan and Yao 1998, and Fan et al. 2003) to estimate the variance function. This variance model is a compromise between the constant residual variance assumption used in the ANOVA method and the approach in which the variances of all the genes are treated as being different. For the example we considered in Section 4, the proposed method yields reasonable results when compared with the published results. Our simulation studies show that the TW-SLM normalization has better performance in terms of the mean squared errors than the *lowess* normalization method. Thus the proposed TW-SRM for microarray data is an interesting alternative to the existing normalization and analysis methods.

The TW-SLM is a "two-way" generalization of the semiparametric regression model proposed by Engle et al. (1986). If $J = 1$ and $f_1 = \cdots = f_n \equiv f$, then the SRM simplifies to the model of Engle et al. (1986). However, the TW-SLM is qualitatively different from the standard semiparametric regression model. For microarray data, the number of genes J is always much greater than the number of arrays n. This fits the description of the well-known "small n, large p" problem. Furthermore, in the TW-SLM, both n (the number of arrays) and J (the number of genes) play the dual role of sample size and number of parameters. That is, for estimating $\boldsymbol{\beta}$, J is the number of parameters, n is the sample size. But for estimating $\boldsymbol{f}$, n is the number of (infinite-

dimensional) parameters, J is the sample size for $\boldsymbol{f}$. We are not aware of any other semiparametric models (Bickel, Klaassen, Ritov and Wellner 1993) in which both n and J play such dual roles of sample size and number of parameters.

There are many interesting and challenging theoretical and computational questions arising from the TW-SLM that have not been addressed in the present paper. For example, it is of interest to consider the asymptotic properties of the SLSE of $\beta_j, 1 \leq j \leq J$ and $f_i, 1 \leq i \leq n$ when $J \to \infty$ but for fixed n. This is a natural setting for microarray data because J is usually large and n small. It is clear that the existing methods and results for semiparametric models (Bickel et al. 1993) do not apply directly to the TW-SLM. Another question of interest is the behavior of the variance-function estimator (24) given in Section 4 for the array and intensity-dependent residual variance model $\sigma_{ij}^2 = \sigma_i^2(x_{ij})$. It appears that the consistency and efficiency results of Ruppert et al. (1997) and Fan and Yao (1998) in the usual nonparametric regression setting do not cover the present case, because the structure of the TW-SLM and that of the usual nonparametric regression model are different. A third question involves computation and properties of robust estimation procedures in the TW-SLM, such as least absolute deviation regression, Huber's M-estimation, and other robust methods.

Acknowledgments: The research of Huang is supported in part by the NIH grants MH001541 and HL72288-01 and an Iowa Informatics Initiative grant. The research of Zhang is partially supported by the NSF grants DMS-0102529 and DMS-0203086. The authors thank Professor Terry Speed and his collaborators for making the Apo A1 data set avaiable online. This data set is used as an example in this paper.

References

1. Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Effi cient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore

2. Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with microarrays. Nat. Genet., 21 (suppl. 1), 33-37.

3. Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. Genome Research, Vol. 10: 2022-2029.

4. Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics, 2: 364-374.

5. Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc., 74, 829-836.

6. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistical Sinica, 12: 111-140.

7. Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. J. Amer. Statist. Assoc., 81: 310-320.

8. Fan, J., Tam, P., Vande Woude, G. and Ren, Y. (2004). Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a Cytokine. Proc. Natl Acad. Sci. USA, to appear.

9. Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. Biometrika, 85: 645-660.

10. Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

11. Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer, New York.

12. Hedge, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Earle-Hughes, J., Snesrud, E., Lee, N., and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. Biotechniques, 29: 548-562.

13. Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5: 299-314.

14. Kerr, M. K., Martin, M., and Churchill, G. A. (2000). Analysis of variance for gene expression microarray data. Journal of Computational Biology, 7: 819-837.

15. Kerr, M. K. and Churchill, G. A. (2001). Experimental design for gene expression microarrays. Biostatistics, 2: 183-201.

16. Park, T., Yi, S-G, Kang, S-H, Lee, S. Y., Lee, Y. S., Simon, R. (2003). Evaluation of normalization methods for microarray data. BMC Bioinformatics, 4: 33-45.

17. Ruppert, D., Wand, M. P., Holst, U., and Hössjet, O. (1997). Local polynomial variance-function estimation. Technometrics, 39: 262-273.

18. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary cDNA microarray. Science, 270: 467-470.

19. Schumaker, L. (1981). *Spline functions: Basic theory.* Wiley, New York.

20. Tseng, G. C., Oh, M-K, Rohlin, L., Liao, J. C. and Wong, W-H (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variation and assessment of gene effects. Nucleic Acids Research, 29: 2549-2557

21. Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significant analysis of microarrays applied to transcriptional response to ionizing radiation. Proc Natl Acad Sci, 98: 5116-5121.

22. Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology, 8: 625-637.

23. Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), *Microarrays: Optical Technologies and Informatics, Vol. 4266 of Proceedings of SPIE*, pages 141-152.

Table 1: Simulation Model 1: $10,000\times$ Summary of MSE of TW-SRM and *Lowess* Normalization

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	5.47	46.50	67.47	85.56	98.20	3428.00
	Lowess	5.54	48.97	70.71	91.96	102.50	4885.00
$\alpha = 0.06$	TW-SRM	4.47	52.87	77.74	91.76	112.80	3563.00
	Lowess	6.48	50.76	73.74	88.60	106.00	5674.00
$\alpha = 0.12$	TW-SRM	7.25	60.88	91.32	106.60	136.30	1704.00
	Lowsss	4.20	51.25	75.49	91.38	111.50	2390.00

Table 2: Simulation Model 2: $10,000\times$ Summary of MSE of TW-SRM and *Lowess* Normalization

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	7.16	51.14	77.69	94.96	120.90	1161.00
	Lowess	8.94	70.39	105.90	124.20	156.80	1650.00
$\alpha = 0.06$	TW-SRM	7.75	47.73	68.68	81.07	98.10	4564.00
	Lowess	7.20	60.46	89.74	105.60	127.70	8583.00
$\alpha = 0.12$	TW-SRM	4.44	50.78	73.45	85.01	105.00	1584.00
	Lowsss	10.85	86.99	132.10	153.20	192.20	3253.00

Table 3: Simulation Model 3: $10,000\times$ Summary of MSE of TW-SRM and *Lowess* Normalization

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	8.31	52.45	75.24	86.51	106.50	1956.00
	Lowess	6.30	54.05	79.42	92.92	113.20	2687.00
$\alpha = 0.06$	TW-SRM	4.32	51.81	75.87	88.84	108.00	3012.00
	Lowess	6.26	53.86	76.72	92.81	108.90	4715.00
$\alpha = 0.12$	TW-SRM	7.74	52.85	78.10	92.71	115.30	2502.00
	Lowsss	7.74	57.30	83.63	100.80	123.00	3131.00

Table 4: Simulation Model 4: $10,000\times$ Summary of MSE of TW-SRM and *Lowess* Normalization

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	6.12	56.96	85.25	105.50	132.60	2767.00
	Lowess	10.20	99.95	153.10	187.60	231.80	3730.00
$\alpha = 0.06$	TW-SRM	5.59	51.77	74.43	84.72	105.70	1067.00
	Lowess	9.78	78.90	119.80	137.30	175.70	1441.00
$\alpha = 0.12$	TW-SRM	6.29	53.49	77.78	91.57	112.10	10070.00
	Lowsss	19.10	121.40	174.50	193.80	236.60	12550.00

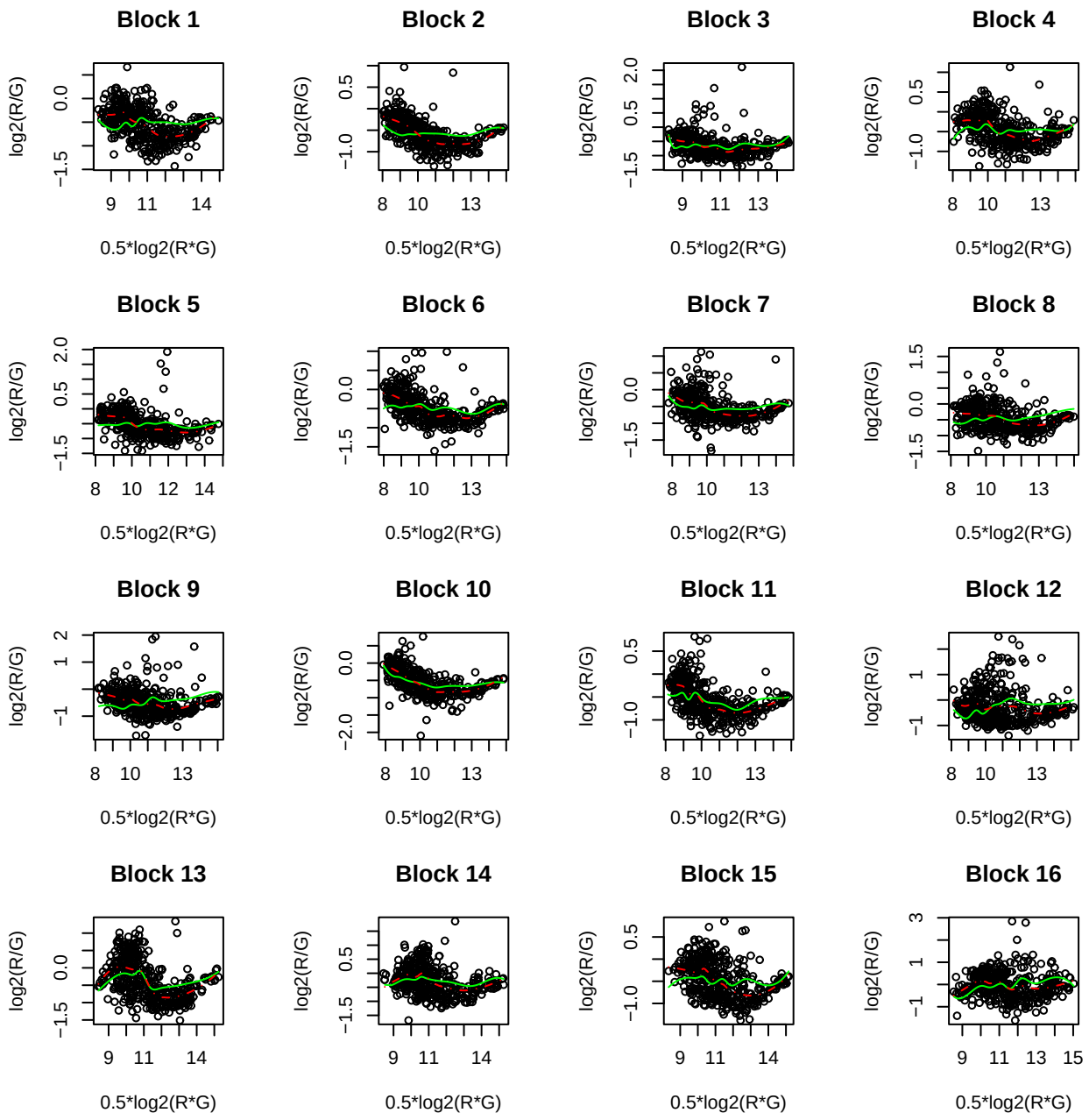

Figure 1: Apo AI data: Comparison of normalization curves in the 16 blocks of the array from one knock-out mouse in the treatment group. Solid Green line: normalization curve based on TW-SLM; Dashed Red line: normalization curve based on *lowess*

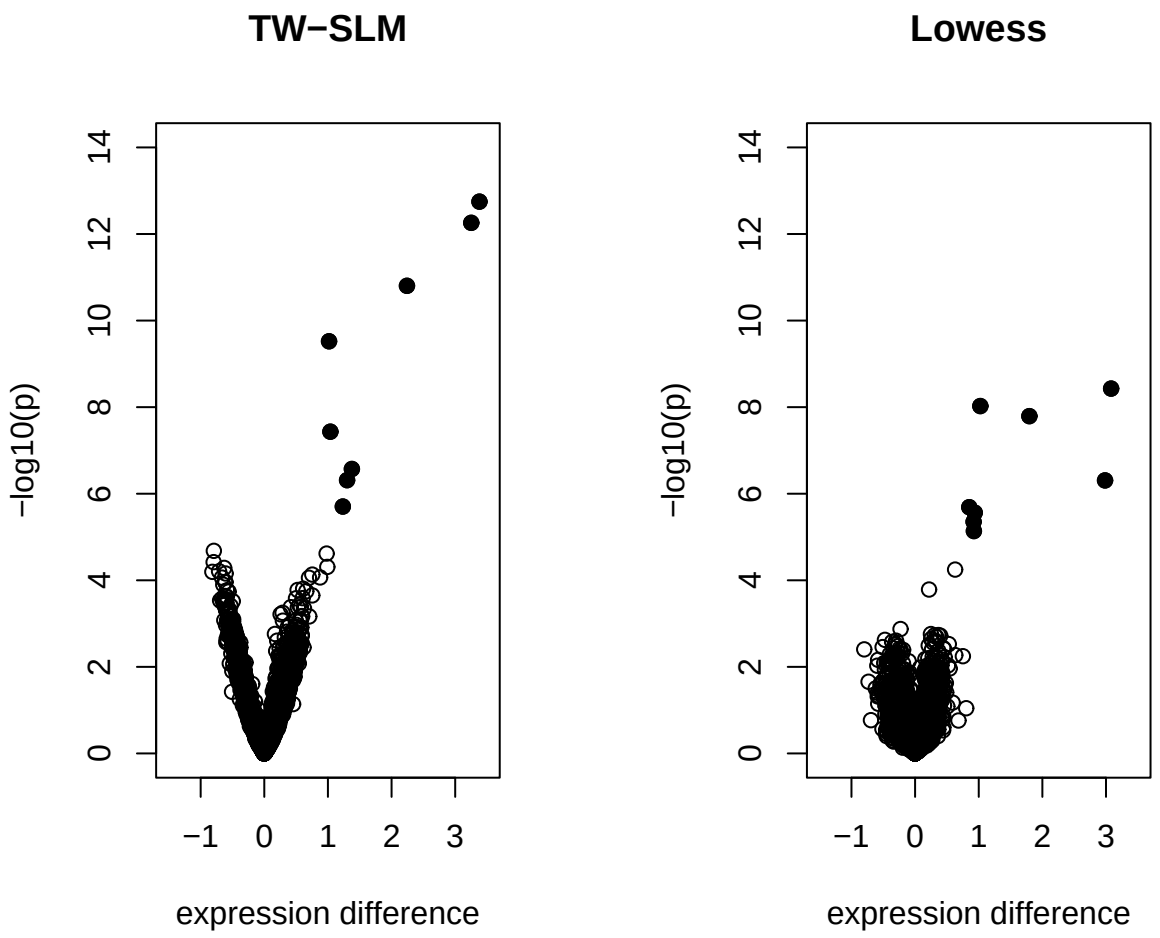

Figure 2: Volcano plot: Scatter plot of $-\log_{10}(\text{p-value})$ versus estimated mean expression value

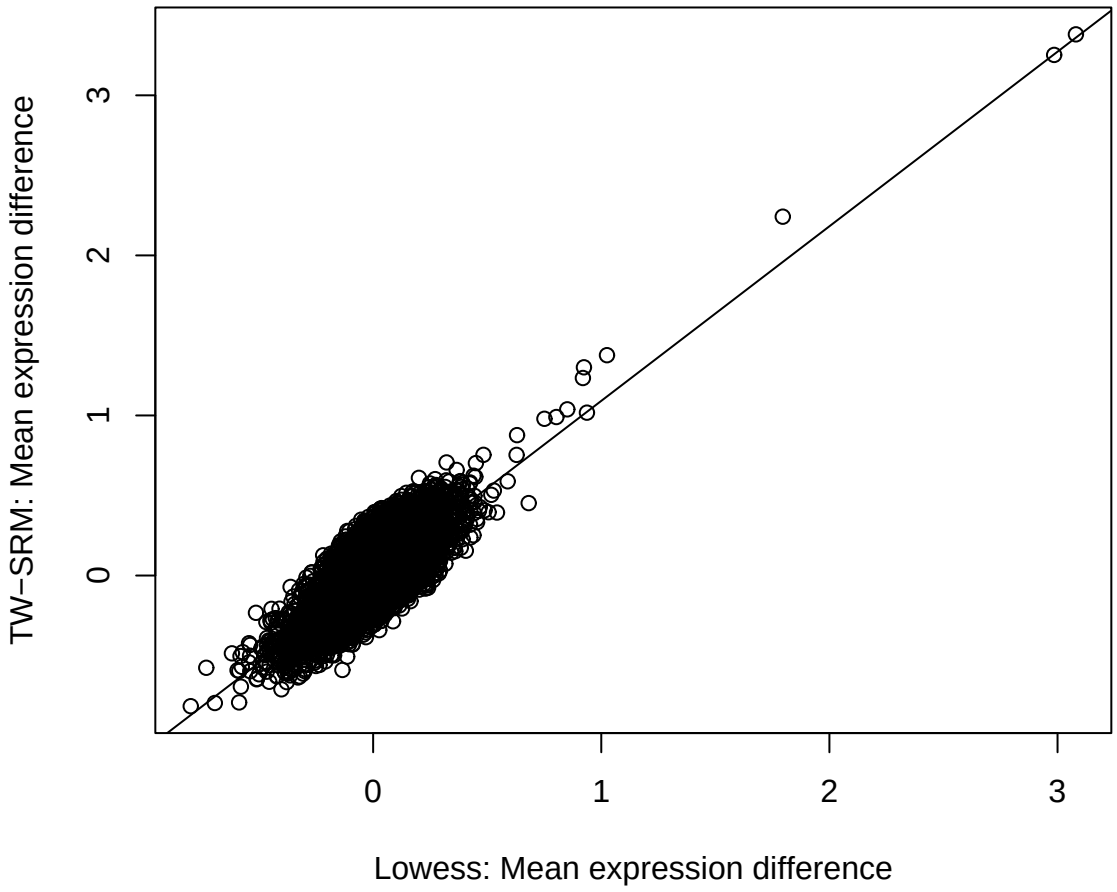

Figure 3: Comparison of TW-SLM and *Lowess* normalized expression values: Scatter plot of normalized mean expression differences

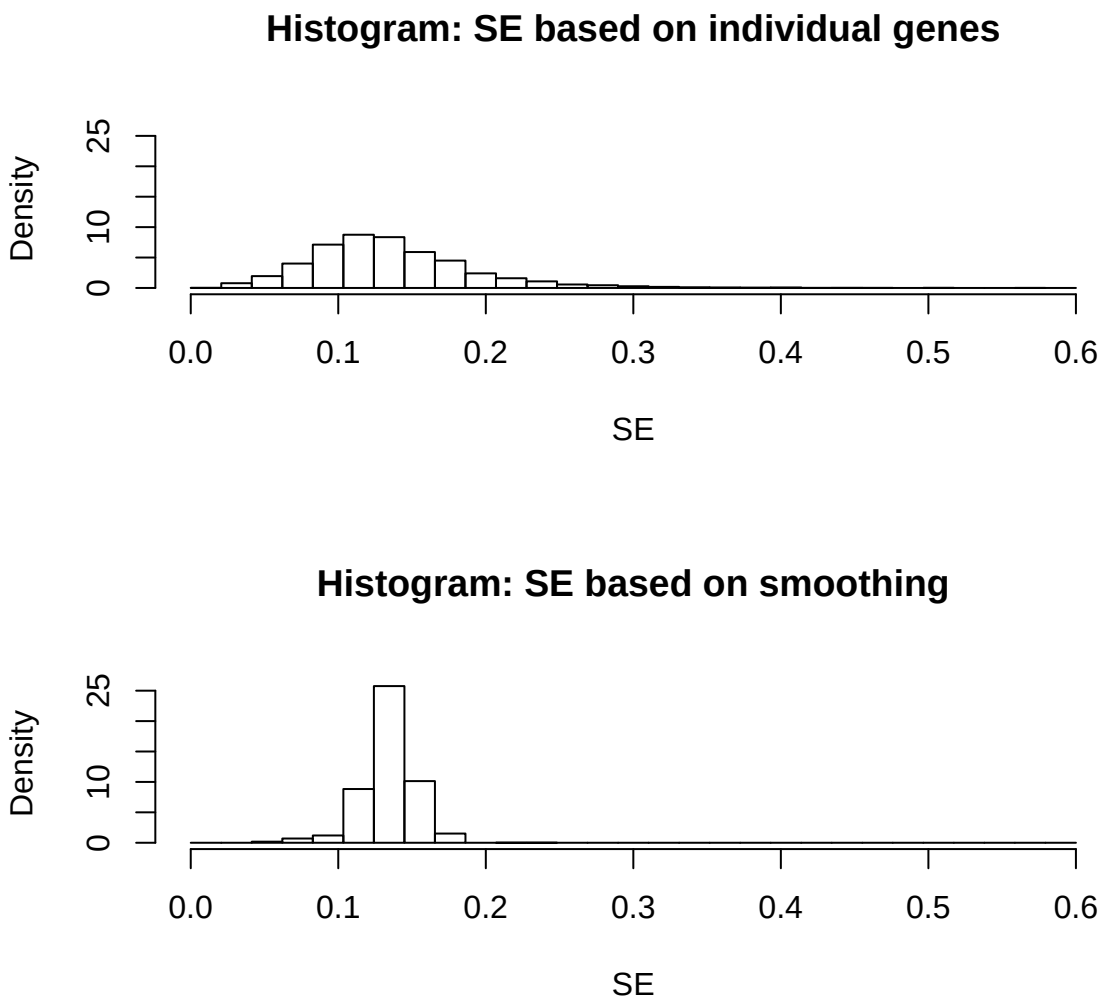

Figure 4: Comparison of variance estimation methods. Top panel: SE estimated based on individual genes. Bottom panel: SE estimated based on smoothing