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

Jian Huang, Deli Wang and Cun-Hui Zhang

ABSTRACT A basic question in analyzing cDNA microarray data is normalization, the purpose of which 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 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 significance 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 study the theoretical properties of the proposed estimator in the TW-SLM, including the finite sample distributional properties of the estimated gene effects and the rate of convergence of the estimated normalization curves when the number of genes under study is large. 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.

Jian Huang is Professor, Department of Statistics and Actuarial Science and Program in Public Health Genetics, University of Iowa, Iowa City, IA 52242 (Email: jian@stat.uiowa.edu). Deli Wang is Research Assistant Professor, Biostatistics and Bioinformatics Unit, Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35294 (Email: deli.wang@ccc.uab.edu). Cun-Hui Zhang is Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08855 (Email: cunhui@stat.rutgers.edu). 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 the Editor, the Associate Editor and two anonymous reviewers for their helpful comments that led to substantial improvement in the paper. The authors also thank Professor Terry Speed and his collaborators for making the Apo A1 data set available online. This data set is used as an example in this paper.

1. Introduction Microarray technology has become a useful tool for quantitatively monitoring gene expression patterns and has been widely used in functional genomics (Schena, Shalon, Davis and Brown 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 which 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, including 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. (2001) 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 (2004) proposed a Semi-Linear In-slide Model (SLIM) method that makes use of replications of a subset of the genes in an array. Fan, Peng and Huang (2004) generalized the SLIM method to account for across-array information, resulting in an aggregated SLIM, so that replication within an array is no longer required. 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, in which 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 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 an array. The TW-SLM also provides a framework for incorporating variability due to normalization into significance 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 values 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 provide theoretical results for the proposed estimators of TW-SLM. In Section 6, we illustrate the proposed method by an example. We also use simulation to compare the proposed method with the *lowess* normalization method and an analogue of the *lowess* method where splines are used in the curve fitting instead of local regression. Some concluding remarks are given in Section 7.

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) (Wahba 1984; 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 an 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, Daugherty and Bittner (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 the ith array $i = 1, \ldots, n$, 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 *lowess* 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 cDNA 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 consist 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), 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 (simple) 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 SRM (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 proposed (general) 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 z_i can also include other factors such as covariates and block effects 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. This can be achieved by simply treating each block as an array and apply the TW-SLM at the block level. We can also adapt the TW-SLM to other designs such as multiple spotting and incorporate spiked control genes in the TW-SLM. Multiple spotting is helpful for improving the precision and for assessing the quality of an array using the coefficient of variation (Tseng et al. 2001). Spike genes can be used for the purpose of calibration and for helping with normalization in an experiment.

3. TW-SLM normalization We now define the semiparametric least squares estimator (SLSE) in the 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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)'$ 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

$$\boldsymbol{\beta} \in \Omega_0^{J\times d} \equiv \Big\{\boldsymbol{\beta} : \sum_{j=1}^J \beta_j = 0\Big\}. \tag{8}$$

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

$$S_i = \overline{\{b_{i0}(x) \equiv 1, b_{ik}(x), k = 1, \ldots, K_i\}} \tag{9}$$

be the spaces of all linear combinations of the basis functions. We approximate f_i by $\alpha_{i0} + \sum_{k=1}^{K_i} b_{ik}(x)\alpha_{ik} = \boldsymbol{b}_i(x)'\alpha_i \in S_i$, 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,

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{f}}) = \underset{(\boldsymbol{\beta},\boldsymbol{f})\in\Omega_0^{J\times d}\times\prod_{i=1}^{n} S_i}{\arg\min} D^2(\boldsymbol{\beta}, \boldsymbol{f}). \tag{10}$$

Let $B_{ij} = (1, b_{i1}(x_{ij}), \ldots, b_{i,K_i}(x_{ij}))'$ be the spline basis functions evaluated at $x_{ij}, 1 \le i \le n, 1 \le j \le J$. The spline basis matrix for the ith array is $B_i = (B_{i1}', \ldots, B_{iJ}')$ where $B_{ij} = (1, b_{i1}(x_{ij}), \ldots, b_{iK_i}(x_{ij}))'$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$. 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 $\hat{f}_i(x) = \boldsymbol{b}_i(x)'\widehat{\alpha}_i, i = 1, \ldots, n$.

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 $k = 0, 1, 2, \ldots$,

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

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

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

$$\widehat{\beta}_j^{(k+1)} = \left(\sum_{i=1}^{n} z_i z_i'\right)^{-1}\sum_{i=1}^{n} z_i\Big(y_{ij} - f_i^{(k)}(x_{ij})\Big), j = 1, \ldots, J. \tag{11}$$

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 $\widehat{\beta}_j = \beta_j^{(K)}, j = 1, \ldots, J$, and the estimated normalization curves are

$$\widehat{f}_i(x) = \boldsymbol{b}_i(x)'\alpha_i^{(K)} = \boldsymbol{b}_i(x)'(B_i'B_i)^{-1}B_i'(\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i), \ i = 1, \ldots, n. \tag{12}$$

The algorithm described above can be conveniently implemented in the statistical computing environment R (R Development Core Team 2003). 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 (12), 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

$$\widetilde{f}_i(\boldsymbol{x}_i) = Q_i \boldsymbol{y}_i = B_i \boldsymbol{\alpha}_i^{(0)}. \tag{13}$$

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. 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 $f(\boldsymbol{x}_i) = (f(x_{i1}), \ldots, f(x_{iJ}))'$ for a univariate function f. We write the TW-SLM (7) in vector notation as

$$\boldsymbol{y}_i = \boldsymbol{\beta} z_i + f_i(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{14}$$

Using (14), it can be shown that the SLSE (10) equals

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - (I_J - Q_i)\boldsymbol{\beta} z_i \right\|^2. \tag{15}$$

In the special case of model (6), $d = 1$ (scalar β_j) and $\boldsymbol{\beta}$ is a vector in $\mathbb{R}^J$, (15) is explicitly

$$\widehat{\boldsymbol{\beta}} = \widehat{\Lambda}_{J,n}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i)\boldsymbol{y}_i z_i' \right), \tag{16}$$

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

$$\widehat{\Lambda}_{J,n} = \frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i). \tag{17}$$

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 $A^{-1}\boldsymbol{x} = \arg\min\{\|\boldsymbol{b}\| : A\boldsymbol{b} = \boldsymbol{x}\}$. 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 $d \geq 1$, (15) is still given by (16) with

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

The information operator (18) 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 $\widehat{\Lambda}_{J,n}\boldsymbol{\beta} = n^{-1}\sum_{i=1}^{n}(I_J - Q_i)\boldsymbol{\beta} z_i z_i'$.

From the expression of $\widehat{\boldsymbol{\beta}}$ given in (16), 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 TW-SLM, 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 $\boldsymbol{\beta}$ Based on (16), we have, conditional on $\{x_{ij}\}$,

$$Var(\widehat{\boldsymbol{\beta}}) \;=\; \widehat{\Lambda}_{J,n}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(I_J - Q_i)Var(\boldsymbol{\varepsilon}_i)(I_J - Q_i) \otimes z_i' z_i\right)\widehat{\Lambda}_{J,n}^{-1}. \tag{19}$$

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 (19). However, direct computation involves inverting a $dJ \times dJ$ matrix. When $J = O(10^4)$, as in many microarray experiments, direct inverting such a large matrix is difficult. In Section 5, we provide an iterative way for computing the variance of a linear combination of $\widehat{\boldsymbol{\beta}}$, which avoids direct inversion and thus is computationally less intensive. However, for the purpose of detecting differentially expressed genes, we are most interested in the variance of individual $\widehat{\boldsymbol{\beta}}_j$. Therefore, we derive an approximation to $Var(\widehat{\boldsymbol{\beta}}_j)$ that is easier to compute. Let $Z_n = \sum_{i=1}^{n} z_i z_i'$. Because $Z_n\widehat{\beta}_j = \sum_{i=1}^{n} z_i(y_{ij} - \widehat{f}_i(x_{ij}))$, we have $Z_n(\widehat{\beta}_j - \beta_j) = \sum_{i=1}^{n} z_i\varepsilon_{ij} + \sum_{i=1}^{n} z_i[f_i(x_{ij}) - \widehat{f}_i(x_{ij})]$. 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 $\Sigma_{\varepsilon,j} = Z_n^{-1}[\sum_{i=1}^n z_i z_i' \sigma_{ij}^2] Z_n^{-1}$. 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 $\widehat{\Sigma}_{\varepsilon,j} = Z_n^{-1}[\sum_{i=1}^n z_i z_i' \hat{\sigma}_{ij}^2] Z_n^{-1}$. For the second term $\Sigma_{f,j}$, we approximate $\widehat{\boldsymbol{f}}_i$ by the ideal normalization curve, that is, $\widehat{f}_i(\boldsymbol{x}_i) = Q_i(\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i) \approx Q_i(\boldsymbol{y}_i - \boldsymbol{\beta} z_i)$. 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 $Var(\widehat{f}_i(x_{ij})) \approx \mathbf{e}_j' Q_i Var(\boldsymbol{\varepsilon}_i) Q_i \mathbf{e}_j$, 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

$$\widehat{\Sigma}_{\beta,j} = \widehat{\Sigma}_{\varepsilon,j} + \widehat{\Sigma}_{f,j}. \tag{20}$$

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

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

As is shown in Section 5, for large J, the distribution of t_j can be approximated by the standard normal distribution under the null $c'\beta_j = 0$. However, to be conservative, we use a t distribution with an appropriate degrees of freedom to approximate the null distribution of t_j when $c'\beta_j = 0$. For example, for a direct comparison design, the degrees of freedom are $n - 1$. For a reference design in a two sample comparison, the variances for the two groups can be estimated separately, and then Welch's correction for the degrees of freedom can be used. Resampling methods such as the permutation method (Dudoit, Yang, Speed and Callow 2002; Reiner, Yekutieli and Benjamini 2003) and the balanced sign test (Fan et al. 2004) can also be used to evaluate the distribution of t_j and the false discovery rate.

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

$$\hat{\sigma}_j^2 = \frac{1}{n-1} \sum_{i=1}^n \hat{\varepsilon}_{ij}^2. \tag{21}$$

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 (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 (12). The resulting spline estimator $\hat{\sigma}_i^2$ can be expressed as

$$\hat{\sigma}_i^2(x) = \boldsymbol{b}_i'(x)(B_i'B_i)^{-1}B_i\widehat{\boldsymbol{\varepsilon}}_i^2, \tag{22}$$

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. Theoretical results In this section, we provide theoretical results concerning the distribution of $\widehat{\boldsymbol{\beta}}$ and the rate of convergence for the normalization of f_i. Our results are derived under subsets of the following four conditions. We assume that the data from different arrays are independent, and impose conditions on the n individual arrays. Our conditions depend on n only through the uniformity requirements across the n arrays, so that all the theorems in this section hold in the case of fixed $n \geq 2$ as the number of genes $J \to \infty$ as well as the case of $(n, J) \to (\infty, \infty)$ with no constraint on the order of n in terms of J. In contrast, the asymptotic results in Huang and Zhang (2004) required $\max_i \text{rank}(Q_i)/n \to 0$, which may not be realistic for certain microarray experiments. The results in this section hold for any basis functions b_{ik} in (9), e.g. spline, Fourier, or wavelet bases, as long as Q_i in (15) are projections from $\mathbb{R}^J$ to $\{f(\boldsymbol{x}_i) : f \in S_i\}$ with $Q_i\mathbf{e} = \mathbf{e}$, where $\mathbf{e} = (1, \ldots, 1)'$. Furthermore, with some modifications in the proof, the results hold when Q_i are replaced by non-negative definite smoothing matrices A_i with their largest eigenvalues bounded by a fixed constant, see for example, Lemmas 1 to 3 in the Appendix.

Condition I: *In (14), $\boldsymbol{x}_i$, $i = 1, \ldots, n$, are independent random vectors, and for each i $\{x_{ij}, j \leq J\}$ are exchangeable random variables. Furthermore, for each $i \leq n$, the space S_i in (9) depends on design variables $\{\boldsymbol{x}_k, z_k, k \leq n\}$ only through the values of $\boldsymbol{x}_i$ and $\{z_k, k \leq n\}$.*

The independence assumption follows from the independence of different arrays which is satisfied in a typical microarray experiment. The exchangeability condition within individual arrays is reasonable if there is no prior knowledge about the total intensity of expression values of the genes under study. It holds when $\{x_{ij}, j \leq J\}$ are conditionally i.i.d. variables given certain (unobservable random) parameters, including within-array i.i.d. $x_{ij} \sim G_i$ as a special case. The exchangeability condition also holds if $\{x_{ij}, j \leq J\}$ are sampled without replacement from a larger collection of variables.

Condition II: *The matrix* $Z_n \equiv \sum_{i=1}^n z_i z_i'$ *is of full rank with* $\max_{i\le n} z_i' Z_n^{-1} z_i \le \kappa^* < 1$.

Condition II is satisfied by common designs such as the reference and direct comparison designs. Since $\sum_{i=1}^n Z_n^{-1} z_i z_i' = I_d$, $\sum_{i=1}^n z_i' Z_n^{-1} z_i = d$. In balanced designs or orthogonal designs with replications, $Z_n \propto I_d$, n is a multiplier of d, and $z_i' Z_n^{-1} z_i = \kappa^* = d/n < 1$ for all $i \le n$. In particular, (6) describes a balanced design with $d = 1$, so that Condition II holds as long as $n \ge 2$.

Condition III: *For the projections* Q_i *in (15),* $K_{J,n}^* \equiv \max_{i\le n} E\{\text{trace}(Q_i) - 1\} = o(J^{1/2})$.

An assumption on the maximum dimensions of the approximation spaces is usually required in nonparametric smoothing. Condition III assumes that the ranks of the projections Q_i be uniformly of the order $o(J^{1/2})$ to avoid over-fitting, and more important, to avoid co-linearity between the approximation spaces for the estimation of $\{f_i(\boldsymbol{x}_i), i \le n\}$ and the design variables for the estimation of $\boldsymbol{\beta}$. Clearly, $E\{\text{trace}(Q_i) - 1\} \le K_i$ for the K_i in (9).

Condition IV: $\rho_{J,n}^* \equiv \max_{i\le n} E\big\|f_i(\boldsymbol{x}_i) - Q_i f_i(\boldsymbol{x}_i)\big\|^2/(J-1) \to 0$.

Condition IV demands that the ranges of the projections Q_i be sufficiently large so that the approximation errors for $f_i(\boldsymbol{x}_i)$ are uniformly $o(1)$ in an average sense. Although this is the weakest possible condition on Q_i for the consistent estimation of $f_i(\boldsymbol{x}_i)$, the combination of Conditions III and IV does require careful selection of spaces S_i in (9) and certain condition on the tail probability of x_{ij}. We provide two specific examples, one below and one in Section 5.2.

Example 1. Let S_i be the collection of splines of degree d^* with equally spaced knots for certain bandwidth (span) h_J. Let I_i be the smallest intervals containing $\{x_{ij}, j \le J\}$. Condition III holds if $\max_{i\le n} E(|I_i| + 1)/h_J = o(J^{1/2})$, where $|I_i|$ is the length of I_i, while Condition IV holds if $h_J \to 0$ and $\{f_i\}$ is uniformly continuous. Now, suppose $P\{|x_{ij} - \mu_i| > t\} \le c_1 \exp(-c_2 t^\kappa)$ for all (i, j) and $t > c_3$ for certain constants μ_i and c_ℓ. Then, $K_{J,n}^* \le M_1\{d^* + (\log J)^{1/\kappa}/h_J\}$ for certain M_1 depending only on κ and the three c_ℓ. If $\{f_i\}$ satisfies a Lipschitz condition with a smoothness index $1/2 < \alpha \le d^* + 1$, then $\rho_{J,n}^* \le M_2 h_J^{2\alpha}$ for certain $M_2 < \infty$. In particular, the orders of the approximation error $\rho_{J,n}^*$ and the estimation error $K_{J,n}^*/J$ for $f_i(\boldsymbol{x}_i)$ reach the balance $((\log J)^{1/\kappa}/J)^{2\alpha/(2\alpha+1)}$ at the bandwidth $h_J \asymp \{(\log J)^{1/\kappa}/J\}^{1/(2\alpha+1)}$. For $\kappa = \infty$, i.e. $P\{|x_{ij} - \mu_i| \le c_3\} = 1$, this example includes the commonly imposed condition $\max_{i\le n}\big(\max_{j\le J} x_{ij} - \min_{j\le J} x_{ij}\big) = O(1)$ in nonparametric smoothing, under which $\rho_{J,n}^* + K_{J,n}^*/J = O(J^{-2\alpha/(2\alpha+1)})$ with $h_J \asymp J^{1/(2\alpha+1)}$ for Lipschitz $\{f_i\}$ with smoothness index α.

5.1. Distribution of $\widehat{\boldsymbol{\beta}}$ We now describe the distribution of $\widehat{\boldsymbol{\beta}}$ in (15) conditionally on all the covariates and provide an upper bound for the conditional bias of $\widehat{\boldsymbol{\beta}}$.

Let $\widehat{\Lambda}_{J,n}$ be the information operator in (18). Define

$$\widetilde{\boldsymbol{b}}_{J,n} = -\Pi_{J,n}\boldsymbol{\beta} + \widehat{\Lambda}_{J,n}^{-1}\left\{\frac{1}{n}\sum_{i=1}^n (I_J - Q_i) f_i(\boldsymbol{x}_i) z_i'\right\}, \tag{23}$$

where $\Pi_{J,n}$ is the projection to $\big\{\boldsymbol{b} \in \Omega_0^{J\times d} : \widehat{\Lambda}_{J,n}\boldsymbol{b} = 0\big\}$. Define

$$V_{J,n} = \frac{1}{n}\sum_{i=1}^n V_i \otimes z_i z_i', \quad V_i = \big(I_J - Q_i\big)\text{Var}(\boldsymbol{\varepsilon}_i)\big(I_J - Q_i\big). \tag{24}$$

Here $\widehat{\Lambda}_{J,n}^{-1}$, the generalized inverse of $\widehat{\Lambda}_{J,n}$, is uniquely defined as a one-to-one mapping from the range of $\widehat{\Lambda}_{J,n}$ to the space $\big(I_J \otimes I_d - \Pi_{J,n}\big)\Omega_0^{J\times d} = \big\{\boldsymbol{b} \in \Omega_0^{J\times d} : \Pi_{J,n}\boldsymbol{b} = 0\big\}$. For any $J \times b$ matrix $\boldsymbol{b}$, the matrix $B = \widehat{\Lambda}_{J,n}^{-1}\boldsymbol{b}$ can be computed by the following recursion:

$$B^{(k+1)} \leftarrow n(\boldsymbol{b} - \Pi_{J,n}\boldsymbol{b})Z_n^{-1} + \sum_{i=1}^{n} Q_i B^{(k)} z_i z_i' Z_n^{-1} \tag{25}$$

with the initialization $B^{(1)} = n(\boldsymbol{b} - \Pi_{J,n}\boldsymbol{b})Z_n^{-1}$ and $Z_n = \sum_{i=1}^n z_i z_i'$.

Theorem 1. *Let $\widehat{\boldsymbol{\beta}}$, $\widehat{\Lambda}_{J,n}$ and $V_{J,n}$ be as in (15), (18) and (24) respectively. Suppose that given $\{\boldsymbol{x}_i, i \le n\}$, $\boldsymbol{\varepsilon}_i$ are independent normal vectors. Then, conditionally on $\{\boldsymbol{x}_i, i \le n\}$,*

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left(\widetilde{\boldsymbol{b}}_{J,n}, \frac{1}{n}\widehat{\Lambda}_{J,n}^{-1} V_{J,n} \widehat{\Lambda}_{J,n}^{-1}\right) \tag{26}$$

In particular, for all $\boldsymbol{b} \in \Omega_0^{J\times d}$, $\lim_{k\to\infty} B^{(k)} = \widehat{\Lambda}_{J,n}^{-1}\boldsymbol{b}$ with the $B^{(k)}$ in (25), and

$$\sigma_{J,n}^2(\boldsymbol{b}) \equiv \mathrm{Var}\Big[\mathrm{trace}\big(\boldsymbol{b}'\widehat{\boldsymbol{\beta}}\big)\,\Big|\,\{\boldsymbol{x}_i, i \le n\}\Big] = \frac{1}{n^2}\sum_{i=1}^{n} z_i'\big(\widehat{\Lambda}_{J,n}^{-1}\boldsymbol{b}\big)' V_i \big(\widehat{\Lambda}_{J,n}^{-1}\boldsymbol{b}\big) z_i. \tag{27}$$

Our next theorem provides sufficient conditions under which the bias of $\widehat{\boldsymbol{\beta}}$ is of smaller order than its standard error.

Theorem 2. *Suppose Conditions I to IV hold. If $c_{J,n}/\rho_{J,n}^* \to \infty$, then*

$$\sup\left\{E\min\left(1, \frac{\mathrm{trace}^2(\boldsymbol{b}'\widetilde{\boldsymbol{b}}_{J,n})}{\mathrm{trace}(\boldsymbol{b}Z_n^{-1}\boldsymbol{b}')c_{J,n}}\right) : \boldsymbol{b} \in \Omega_0^{J\times d}, \boldsymbol{b} \neq 0\right\} = o(1). \tag{28}$$

In particular, if given $\{\boldsymbol{x}_i, i \le n\}$, $\boldsymbol{\varepsilon}_i$ are independent normal vectors with Var$(\boldsymbol{\varepsilon}_i) \ge \sigma_*^2 I_J$ for certain $\sigma_* > 0$, then

$$\sup_{\boldsymbol{b}\in\Omega_0^{J\times d}, \boldsymbol{b}\neq 0}\left\{\sup_{x\in\mathbb{R}}\left|P\left(\frac{\mathrm{trace}(\boldsymbol{b}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}))}{\sigma_{J,n}(\boldsymbol{b})} \le x\right) - \Phi(x)\right|\right\} = o(1), \tag{29}$$

where Φ is the cumulative distribution function for $N(0,1)$.

This result states that, under Conditions I to IV, appropriate linear combinations of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, such as contrasts, have an approximate normal distribution with mean zero and the approximation is uniform over all linear combinations. Therefore, this result provides theoretical justification for inference procedures based on $\widehat{\boldsymbol{\beta}}$, such as those described in Section 4. Without the normality condition, (29) is expected to hold under the Lindeberg condition as $(n, J) \to (\infty, \infty)$, even in the case $n = o(J)$ [for example $n = O(\log J)$]. We assume the normality here so that (29) holds for fixed n as well as large n.

5.2. Convergence rates of estimated normalization curves $\widehat{f}_i$ Normalization is not only important in detecting differentially expressed genes, it is also a basic first step for other high

level analysis, including classification and cluster analysis. Thus, it is of interest in itself to study the behavior of the estimated normalization curves. Here we study the convergence rates of $\widehat{f}_i$.

Since $\widehat{f}_i(\boldsymbol{x}_i) = Q_i(\boldsymbol{y}_i - \widehat{\boldsymbol{\beta}} z_i)$, it follows from (14) that

$$\widehat{f}_i(\boldsymbol{x}_i) = Q_i\big\{f_i(\boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i\big\} - Q_i\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) z_i. \tag{30}$$

Therefore, the convergence rates of $\|\widehat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|$ are bounded by the sums of the rates of $\|Q_i\big\{f_i(\boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i\big\} - f_i(\boldsymbol{x}_i)\|$ for the "ideal fits" $Q_i(\boldsymbol{y}_i - \boldsymbol{\beta} z_i)$ and the rates of $\|Q_i\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) z_i\|$.

Theorem 3. *Suppose Conditions I to IV hold and* $\mathrm{Var}(\boldsymbol{\varepsilon}_i) \le (\sigma^*)^2 I_J$ *for certain* $0 < \sigma^* < \infty$. *Then, for certain* $\epsilon_{J,M}$ *with* $\lim_{M\to\infty} \lim_{J\to\infty} \epsilon_{J,M} \to 0$,

$$\max_{i \le n} P\Big\{ \|\widehat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|^2/J > M\big(\rho^*_{J,n} + (\sigma^*)^2 K^*_{J,n}/J\big) \Big\} \le \epsilon_{J,M}$$

In particular, if $K^*_{J,n} = O(1) J^{1/(2\alpha+1)}$ *and* $\rho^*_{J,n} = O(1) J^{2\gamma/(2\alpha+1)}$ *for certain* $0 < \gamma \le \alpha$, *then* $\|\widehat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|^2/J = O_P(J^{-2\gamma/(2\alpha+1)})$, *where the* O_P *is uniform in* $i \le n$.

In the case of $\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \sigma^2 I_J$, $\max_{i\le n} E\|Q_i(\boldsymbol{y}_i - \boldsymbol{\beta} z_i) - f_i(\boldsymbol{x}_i)\|^2/J \ge \max\big\{\rho^*_{J,n}, \sigma^2 K^*_{J,n}/J\big\}$ is the convergence rate for the ideal fits $Q_i(\boldsymbol{y}_i - \boldsymbol{\beta} z_i)$ for $f_i(\boldsymbol{x}_i)$. Theorem 3 asserts that $\widehat{f}_i(\boldsymbol{x}_i)$ have the same convergence rates as the ideal fits. Thus, $\widehat{f}_i(\boldsymbol{x}_i)$ achieve or nearly achieve the optimal rate of convergence for normalization under various settings. If $\{f_i\}$ satisfies a Lipschitz condition with a smoothness index $\alpha > 1/2$ and S_i are chosen as in Example 1 with $h_J \asymp J^{1/(2\alpha+1)}$, then the optimal minimax convergence rate $\|\widehat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|^2/J = O_P(1) J^{-2\alpha/(2\alpha+1)}$ for normalization is achieved (within a logarithmic factor) for $\kappa = \infty$ ($\kappa < \infty$) when $P\{|x_{ij} - \mu_i| > t\} \le c_1 \exp(-c_2 t^\kappa)$, $t > c_3$, as in Example 1. Similar results are provided in Example 2 below for Sobolev classes.

Example 2. Let $h_J > 0$ and S_i be the collection of splines of degrees $d_i \le d^*$ with knots $s_{ik} = \inf\{x : x \ge s_{i,k-1} + h_J |I_i| \text{ or } \#(j : s_{i,k-1} < x_{ij} \le x) \ge h_J J\}$, $k \ge 1$, where $I_i \ni x_{ij}$ are as in Example 1 and $s_{i0} = \min_j x_{ij}$. Clearly, $K^*_{J,n} \le d^* + 2/h_J$ so that Condition III holds for $1/h_J = o(J^{1/2})$. For intervals I, let $\mathcal{F}_{\alpha,M}(I)$ be the Sobolev space of functions f in I, e.g. satisfying $\int_I (f^{(\alpha)})^2 \le M^2/I^{2\alpha-1}$ for integers α. If $f_i \in \mathcal{F}_{\alpha_i, M_i}(I_i)$ with $\alpha_i \le d_i + 1$ and (α_i, M_i) dependent on I_i, then $\rho^*_{J,n} \le M^* \max_{i\le n} E M_i^2 h_J^{2\alpha_i}$. In particular, for $h_J \asymp J^{-1/(2\alpha+1)}$ with $\alpha > 1/2$ and in the case where $\alpha_i = \gamma$ and $\sup_i E M_i^2 < \infty$, $\rho_{J,n} = O(1) J^{-2\gamma/(2\alpha+1)}$ and $J^{-1} K^*_{J,n} \asymp J^{-2\alpha/(2\alpha+1)}$, so that by Theorem 3 the optimal convergence rate of $\|\widehat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|/J = O_P(1) J^{-2\alpha/(2\alpha+1)}$ for normalization is achieved when $\{f_i\}$ actually has smoothness index α.

5.3. Invertibility of the information operator It follows from (16) and (30) that the invertibility of the information operator $\widehat{\Lambda}_{J,n}$ in (18) is crucial in our investigation of the SLSE (10). Let $I_{J,0} = I_J - J^{-1}\mathbf{e}\mathbf{e}'$ with $\mathbf{e} = (1, \ldots, 1)' \in \mathbb{R}^J$, so that $I_{J,0} \otimes I_d$ is the identity operator in the parameter space $\Omega_0^{J\times d}$ for $\boldsymbol{\beta}$. We shall study (18) by comparing it with

$$\widetilde{\Lambda}_{J,n} \equiv n^{-1} I_{J,0} \otimes Z_n, \tag{31}$$

which is the information operator for the estimation of $\boldsymbol{\beta}$ in the two-way linear model

$$y_{ij} = z_i'\beta_j + \mu_i + \varepsilon_{ij}, \qquad \mu_i \in \mathbb{R}\,.$$

For nonnegative-definite matrices or linear operators A, we denote by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ their largest and smallest eigenvalues.

Proposition 1. Let $\widehat{\Lambda}_{J,n}$, $\widetilde{\Lambda}_{J,n}$ and $\Pi_{J,n}$ be as in (18), (31) and (23) respectively.
(i) *Let* $\widehat{K}_i = \text{trace}(Q_i) - 1$. *Suppose Condition I holds and* Z_n *is of rank* d. *Then,*

$$0 \le \lambda_{\max}\Big(I_{J,0} \otimes I_d - \widetilde{\Lambda}_{J,n}^{-1/2}\widehat{\Lambda}_{J,n}\widetilde{\Lambda}_{J,n}^{-1/2}\Big) - \lambda^* \le \left(\frac{n-1}{n}\right)^{1/2}\zeta^*, \tag{32}$$

where $\lambda^* = \max_{i\le n} z_i'Z_n^{-1}z_i$ *and* ζ^* *is a nonnegative variable satisfying*

$$E\Big(\zeta^*\Big)^2 = \sum_{1\le i\ne k\le n} \frac{E\widehat{K}_iE\widehat{K}_k}{J-1}\big(z_i'Z_n^{-1}z_k\big)^2 \le \frac{d(n-d)}{n(J-1)}\max_{i\le n}\big(E\widehat{K}_i\big)^2. \tag{33}$$

(ii) *Suppose Conditions I to III hold. Then,* $\lambda_{\max}\big(\widehat{\Lambda}_{J,n}^{-1}\big)\lambda_{\min}(Z_n/n) = O_P(1)$ *and*

$$\lambda_{\max}\Big(\widetilde{\Lambda}_{J,n}^{1/2}\widehat{\Lambda}_{J,n}^{-1}\widetilde{\Lambda}_{J,n}^{1/2}\Big) = O_P(1), \quad P\Big\{\Pi_{J,n} = 0\Big\} \to 1.$$

5.4. Relationship to a semilinear model with infinitely many finite-dimensional parameters Fan, Peng and Huang (2004), hereafter referred to as FPH (2004), studies the SLIM (Fan et al. 2004), a semilinear high-dimensional model for the normalization within a single microarray in the presence of replications of genes, and an aggregation of the SLIM across arrays

$$y_{ij} = \beta_{g(j)} + v_{ij}'\gamma_i + f_i(x_{ij}) + \varepsilon_{ij}, \quad j \le J, i \le n, \tag{34}$$

where γ_i are vectors of a relatively low dimensionality for block effects, v_{ij} indicate blocks within arrays, and $g(j)$ is a many-to-one mapping representing replications of genes within arrays, e.g. the 7-th gene is allocated to $\{j : g(j) = 7\}$ within arrays by design. This general model includes as special cases the SLIM with $n = 1$ and the simple TW-SLM (6) with $g(j) = j$ and $\gamma_i = 0$. Note that here we used β to denote gene effect as in (6) or (7), whereras FPH (2004) uses β to denote block effects and α to denote gene effects.

FPH (2004) provides theoretical properties of the profile least squares method in SLIM with balanced replications, e.g. $n = 1$ and $\#\{j : g(j) = \ell\} = k^* > 1$ for $\ell = 1, 2, \ldots, J/k^*$ in (34). For $n > 1$ and balanced replications in the aggregated SLIM (34), FPH (2004) provides an elegant analysis for the estimation of block effects γ_i with semiparametric information bounds, but their comments (FPH, 2004, below Theorem 5) indicate that the block effects γ_i are not identifiable without replications, i.e. $k^* = 1$ as in model (6). Moreover, the simulation results in FPH (2004, Example 4) suggest that the profile lease squares estimator of $\{f_i\}$ is asymptotically consistent in (34) in the absence of replications and block effects, i.e. in model (6).

It is clear that the aggregated SLIM (34) and the general TW-SLM (7) are closely related as they both include (6) as a special case and the applicabiity of these models are much wider than explicitly stated as arrays could be viewed as blocks and vice versa. Since the focus of FPH (2004) is the block effects in the case of aggregated SLIM and our focus is the normalization $\widehat{f_i}$ and resulting estimator $\widehat{\beta}$ for the gene effects in the absence of block effects and replications within arrays, the approaches and results of the two papers well complement each other in many ways. For the statistical theory concerning normalization of microarrays, the simple TW-SLM (6) and the SLIM

$$y_{kg} = \beta_g + f(x_{kg}) + \varepsilon_{kg} \tag{35}$$

of Fan et al. (2004) for a single array (with no block effects and balanced replication of genes) provide the most direct comparison between the two papers, where $k = 1, \ldots, k^*$ is the index for replications and $g = 1, \ldots, G$ is the index for genes, i.e. $j = (k, g)$ in (34). Mapping (k, k^*, g, G) into (i, n, j, J), we observe that (35) is identical to (6) when $f_i = f$ for all i. The analyzes here and in FPH (2004) reveal that the two models are theoretically quite different as the information operators for the estimation of gene effects and block effects are different.

6. An example and simulation studies

6.1. 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 et al. (2002). 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. (2002), we use printing-tip dependent normalization. The TW-SLM model used here is

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

where $i = 1, \ldots, 16, k = 1, \ldots, 16$, and $j = 1, \ldots, 399$. Here i indexes arrays, k indexes printing-tip blocks, and j index genes in a block. 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 $\beta_{kj} = (\beta_{kj1}, \beta_{kj2})$. The contrast $\beta_{kj1} - \beta_{kj2}$ measures the expression difference for the jth gene in the kth block between the two groups.

To compare the proposed method with the existing ones, we also analyzed the data using the *lowess* normalization method as in Dudoit et al. (2002), and a lowess-like method where, instead of using local regression, splines are used in estimating the normalization curves described in (13) at the end of Section 3. We refer to this method as the *spline* (normalization) method below.

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 line is the normalization curve based on the TW-SLM model, and the dashed 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. (2002), 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. The normalization curves estimated using the *spline* method with exactly the same basis functions used in the TW-SLM closely resemble those estimated using the *lowess* method. Because they are indistinguishable by eye-ball examination, these curves are not included in the plots.

Figure 2 displays the volcano plots of $-\log 10$ p-values versus the mean differences of log-expression values between the knock-out and control groups. In the first (left panel) volcano plot, both the normalization and estimation of $\boldsymbol{\beta}$ are based on the TW-SLM. We estimated the variances for $\widehat{\beta}_{kj1}$ and $\widehat{\beta}_{kj2}$ separately. These variances are estimated based on (22) that assumes that the residual variances depend smoothly on the total log-intensities. We then used Welch's correction for the degrees of freedom in calculating the p-values. The second (middle panel) plot is based on the *lowess* normalization method and use the two-sample t-statistics as in Dudoit et al. (2002), but the p-values are obtained based on Welch's correction for the degrees of freedom. The third (right panel) plot is based on the *spline* normalization method and uses the same two-sample t-statistics as in the *lowess* method. The 8 solid circles in the *lowess* volcano plot are the significant genes that were identified by Dudoit et al. (2002). These 8 genes are also plotted as solid circles in the TW-SLM and *spline* volcano plots, and are significant based on the TW-SLM and *spline* methods, as can be seen from the volcano plots. Comparing the three volcano plots, we see that: (i) the $-\log 10$ p-values based on the TW-SLM method tend to be higher than those based on the *lowess* and *spline* methods, as discussed at the end of Section 2.1; (ii) the p-values based on the *lowess* and *spline* methods are comparable.

Because we use exactly the same smoothing procedure in the TW-SLM and *spline* methods, and because the results between the *lowess* and *spline* methods are very similar, we conclude that the differences between the TW-SLM and *lowess* volcano plots are mostly due to the different normalization methods and two different approaches for estimating the variances. We first examine

the differences between the TW-SLM normalization values and the *lowess* as well as the *spline* normalization values. We plot the three pairwise scatter plots of estimated mean expression differences based on the TW-SLM, *lowess*, and *spline* normalization methods, see Figure 3. In each scatter plot, the solid line is the fitted linear regression line. For the TW-SLM versus *lowess* comparison (left panel), the fitted regression line is

$$y = 0.00029 + 1.090x. \tag{36}$$

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. For the TW-SLM versus *spline* comparison (middle panel), the fitted regression line and the standard errors are virtually identical to (36) and its associated standard errors. For the *spline* versus *lowess* comparison (right panel), the fitted regression line is

$$y = 0.00027 + 1.00257x. \tag{37}$$

The standard error of the intercept is 0.00025, and the standard of the slope is 0.0015. Therefore, the mean expression differences based on the *lowess* and *spline* normalization methods are essentially the same, as can also be seen from the scatter plot in the right panel in Figure 3.

Figure 4 shows the histograms of the standard errors obtained based on intensity-dependent smoothing defined in (22) using the residuals from the TW-SLM normalization (top panel), and the standard errors calculated for individual genes using the *lowess* and *spline* methods (middle and bottom panels). 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.

6.2. Simulation studies We use simulation to compare the TW-SLM, *lowess*, and *spline* 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 $\beta_j \sim N(0, \tau^2)$. The percentage of such genes is $1 - \alpha$. For up-regulated genes, we simulate $\beta_j \sim N(\mu, \tau_U^2)$ where $\mu > 0$. For down-regulated genes, we simulate $\beta_j \sim N(-\mu, \tau_D^2)$. We use $\tau = 0.6, \mu = 2, \tau_U = \tau_D = 1$.

(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 the bias, variance, and mean square error of estimated expression values relative to the generating values. In each of the four cases, we consider two levels of the percentage of differentially expressed genes: $\alpha = 0.01$ and 0.06.

Tables 1 to 4 present 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 than the *lowess* method. The *spline* method has higher MSEs than both the TW-SLM and *lowess* methods. 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 than both the *lowess* and *spline* methods. In Table 3 for simulation Model 3, there is non-linear intensity dependent dye bias, but there is symmetry between the up- and down-regulated genes. The TW-SLM has comparable but slightly smaller MSEs than the *lowess* method. The *spline* method has higher MSEs than both the TW-SLM and *lowess* methods. In 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 among the TW-SLM, *lowss*, and *spline* methods. However, the TW-SLM method generally has smaller biases.

7. Discussion The TW-SLM puts normalization and significance 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. 2004) 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 6, 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* and *spline* normalization methods. Thus the proposed TW-SRM for microarray data is a powerful alternative to the existing normalization and analysis methods.

We studied the distributional properties of the SLSE $\widehat{\boldsymbol{\beta}}$ under some appropriate conditions given in Section 5. We also studied the rates of convergence of the estimated normalization curve $\widehat{f_i}$ when the number of genes goes to infinity. This is a reasonable framework for asymptotics in the TW-SLM because J is usually large and n small. The results we obtained provide theoretical justification for the estimation of normalization curves and inference for the gene effects under the TW-SLM. We note that the existing methods and results for semiparametric models (Bickel et al. 1993) do not apply directly to the TW-SLM.

If $J = 1$ and $f_1 = \cdots = f_n \equiv f$, then the TW-SLM simplifies to the standard semiparametric regression model (Wahba 1984; Engel et al. 1986). However, the TW-SLM is qualitatively different from this 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 et al. 1993) in which both n and J play such dual roles of sample size and number of parameters.

There are many other interesting and challenging theoretical and computational questions arising from the TW-SLM that are beyond the scope of the present paper, for example, questions involving 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.

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., Speed, T. P. and Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistical Sinica, 12: 111-139.

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

9. 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, 1135-1140.

10. Fan, J., Peng, H. and Huang, T. (2004). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. Preprint. Dept of Operations Research and Financial Engineering, Princeton University.

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

12. Huang, J. and Zhang, C.-H. (2003) Asymptotic analysis of a two-way semiparametric regression model for microarray data. *Technical Report* No. 2003-06, Rutgers University.

13. 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.

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. and Simon, R. (2003). Evaluation of normalization methods for microarray data. BMC Bioinformatics, 4: 33-45.

17. R Development Core Team (2003). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: www.R-project.org.

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

19. Reiner, A., Yekutieli D. and Benjamini Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics, 19: 368-375.

20. 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.

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

22. 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

23. 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.

24. Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319-329. Institute of Statistical Mathematics, Tokyo.

25. 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.

26. 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.

8. Appendix We provide the proofs of Theorem 1, Proposition 1, and then Theorems 2 and 3.

Proof of Theorem 1. Since $\widehat{\boldsymbol{\beta}}$ is the solution of (14) and (15) with the shortest Hilbert-Schmidt norm, $\Pi_{J,n}\widehat{\boldsymbol{\beta}} = 0$ and $\boldsymbol{\beta} - \Pi_{J,n}\boldsymbol{\beta} = \widehat{\Lambda}_{J,n}^{-1}\sum_{i=1}^n (I_J - Q_i)\boldsymbol{\beta} z_i z_i'$. Thus, (23) is the conditional bias. Since $\widehat{\Lambda}_{J,n}$ is a function of covariates, (26) holds by (14) and (15) if $V_{J,n}/n$ is the conditional covariance operator of $n^{-1}\sum_{i=1}^n \big(I_J - Q_i\big)\boldsymbol{\varepsilon}_i z_i'$. For $J \times d$ matrices $\boldsymbol{b}$,

$$\text{trace}^2\left(\boldsymbol{b}'\Big(I_J - Q_i\Big)\boldsymbol{\varepsilon}_i z_i'\right) = z_i'\boldsymbol{b}'\Big(I_J - Q_i\Big)\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\Big(I_J - Q_i\Big)'\boldsymbol{b} z_i$$

has the conditional expectation $z_i'\boldsymbol{b}'V_i\boldsymbol{b}z_i$, so that (26) holds by the independence of $\boldsymbol{\varepsilon}_i$, i.e.

$$\text{Cov}\left(\frac{1}{n}\sum_{i=1}^n \big(I_J - Q_i\big)\boldsymbol{\varepsilon}_i z_i' \,\middle|\, \{\boldsymbol{x}_i, i \le n\}\right) = \frac{1}{n^2}\sum_{i=1}^n V_i \otimes z_i z_i' = \frac{V_{J,n}}{n}.$$

It remains to prove the convergence of (25) to $B \equiv \widehat{\Lambda}_{J,n}^{-1}\boldsymbol{b}$. Let $\widetilde{z}_i = Z_n^{-1/2}z_i$ so that $Z_n\sum_{i=1}^n \widetilde{z}_i\widetilde{z}_i' = Z_n$. Set $\widehat{A}_i \equiv I_{J,0}Q_iI_{J,0} = Q_iI_{J,0}$ with the $I_{J,0}$ in (31). By (18) and the definition of $\Pi_{J,n}$ in (23),

$$\boldsymbol{b} - \Pi_{J,n}\boldsymbol{b} = \widehat{\Lambda}_{J,n}B = \frac{1}{n}\sum_{i=1}^n (I - Q_i)Bz_iz_i' = \frac{1}{n}\sum_{i=1}^n (I_{J,0} - \widehat{A}_i)BZ_n^{1/2}\widetilde{z}_i\widetilde{z}_i'Z_n^{1/2},$$

so that $\boldsymbol{b}_* \equiv (\boldsymbol{b} - \Pi_{J,n}\boldsymbol{b})Z_n^{-1/2} = B_* - \widehat{A}_{J,n}B_*$ with $B_* \equiv BZ_n^{1/2}/n$ and $\widehat{A}_{J,n} \equiv \sum_{i=1}^n \widehat{A}_i \otimes \widetilde{z}_i\widetilde{z}_i'$. Since $\widehat{A}_i$ are projections and $\sum_{i=1}^n \widetilde{z}_i\widetilde{z}_i' \le I_d$, the operator $\widehat{A}_{J,n}$ is nonnegative-definite with $\lambda_{\max}(\widehat{A}_{J,n}) \le 1$. Moreover, both $\boldsymbol{b}_*$ and B_* are inside the linear space generated by the eigenmatrices of $\widehat{A}_{J,n}$ with eigenvalues in $[0,1)$. Thus, $B_* = \lim_{N\to\infty} B_*^{(N)}$ with $B_*^{(N)} = \sum_{k=0}^N (\widehat{A}_{J,n})^k\boldsymbol{b}_*$. This proves $B^{(k)} \to B$, since (25) is equivalent to

$$\begin{aligned} B_*^{(k+1)} &\equiv B^{(k+1)}Z_n^{1/2}/n = (\boldsymbol{b} - \Pi_{J,n}\boldsymbol{b})Z_n^{-1/2} + \frac{1}{n}\sum_{i=1}^n Q_iB^{(k)}z_iz_i'Z_n^{-1/2} \\ &= \boldsymbol{b}_* + \sum_{i=1}^n \widehat{A}_i\big(B^{(k)}Z_n^{1/2}/n\big)\widetilde{z}_i\widetilde{z}_i' = \boldsymbol{b}_* + \widehat{A}_{J,n}B_*^{(k)}. \end{aligned}$$

Note that $\mathbf{e}'\boldsymbol{b} = \mathbf{e}'\Pi_{J,n}\boldsymbol{b} = 0$ implies $Q_iB^{(k)} = Q_iI_{J,0}B^{(k)}$. The proof of Theorem 1 is complete.

We need three lemmas for the proof of Proposition 1.

Lemma 1. *Let $\mathbf{v}_i$, $i = 1, \ldots, n$, be vectors in $\mathbb{R}^J$. Then,*

$$\lambda_{\max}\left(\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i'\right) = \sup\left\{\left\|\sum_{i=1}^n c_i\mathbf{v}_i\right\|^2 : \sum_{i=1}^n c_i^2 = 1\right\}. \tag{38}$$

Moreover, for nonnegative-definite matrices M_i,

$$\lambda_{\max}\left(\sum_{i=1}^n M_i\right) = \sup\left\{\left\|\sum_{i=1}^n c_iM_i^{1/2}\mathbf{u}_i\right\|^2 : \sum_{i=1}^n c_i^2 = 1, \|\mathbf{u}_i\| = 1\right\}. \tag{39}$$

Proof. Replacing $\mathbf{v}_i$ by $(\mathbf{v}_i', 0, \ldots, 0)'$ and perturbing $\mathbf{v}_i$ with infinitesimal vectors if necessary, we assume without loss of generality that $\{\mathbf{v}_i, i \le n\}$ are linearly independent. Let $\mathbf{u}$ be an eigenvector of $\sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i'$ with eigenvalue λ. Since $\mathbf{u}$ is in the range of the matrix, $\mathbf{u} = \sum_{i=1}^n c_i\mathbf{v}_i$ for certain real numbers c_i, so that

$$\sum_{i=1}^n (\lambda c_i)\mathbf{v}_i = \lambda\mathbf{u} = \sum_{i=1}^n \mathbf{v}_i\mathbf{v}_i' \sum_{k=1}^n c_k\mathbf{v}_k = \sum_{i=1}^n \Big(\sum_{k=1}^n \mathbf{v}_i'\mathbf{v}_k c_k\Big)\mathbf{v}_i.$$

Thus, due to the linear independence of $\{\mathbf{v}_i, i \le n\}$, $(c_1, \ldots, c_n)'$ is an eigenvector of the matrix $\big(\mathbf{v}_i'\mathbf{v}_k\big)_{n\times n}$ with the same eigenvalue λ. This immediately implies (38).

For (39), let $M_i = \sum_{j=1}^J \lambda_{ij}\mathbf{u}_{ij}\mathbf{u}_{ij}'$ be the eigenvalue decompositions of M_i. With $\mathbf{v}_{ij} = \lambda_{ij}^{1/2}\mathbf{u}_{ij}$, we find $M_i = \sum_j \mathbf{v}_{ij}\mathbf{v}_{ij}'$ and $M_i^{1/2} = \sum_j \mathbf{v}_{ij}\mathbf{u}_{ij}'$. By (38),

$$\lambda_{\max}\Big(\sum_i M_i\Big) = \lambda_{\max}\Big(\sum_{ij} \mathbf{v}_{ij}\mathbf{v}_{ij}'\Big) = \sup\Big\{\Big\|\sum_{ij} c_{ij}\mathbf{v}_{ij}\Big\|^2 : \sum_{ij} c_{ij}^2 \le 1\Big\}. \tag{40}$$

Now, set $c_i = \Big(\sum_j c_{ij}^2\Big)^{1/2}$, $\mathbf{u}_i = \sum_j c_{ij}\mathbf{u}_{ij}/c_i$ for $c_i > 0$, and $\mathbf{u}_i$ as any unit vector for $c_i = 0$. We have $c_{ij} = c_i\mathbf{u}_{ij}'\mathbf{u}_i$ and

$$\sum_j c_{ij}\mathbf{v}_{ij} = \sum_j \lambda_{ij}^{1/2} c_{ij}\mathbf{u}_{ij} = \sum_j \lambda_{ij}^{1/2}\mathbf{u}_{ij}c_i\mathbf{u}_{ij}'\mathbf{u}_i = c_i M_i^{1/2}\mathbf{u}_i.$$

This and (40) imply (39). The proof of Lemma 1 is complete.

Lemma 2. *Let M_i be nonnegative-definite matrices. Then,*

$$0 \le \lambda_{\max}\Big(\sum_{i=1}^n M_i\Big) - \max_{i\le n}\lambda_{\max}(M_i) \le \Big(\frac{n-1}{n}\Big)^{1/2}\zeta^*, \tag{41}$$

where $\zeta^* \equiv \big\{\sum_{1\le i\ne k\le n} \mathrm{trace}\big(M_iM_k\big)\big\}^{1/2}$.

Proof. For all matrices A and B and vectors $\mathbf{u}$ and $\mathbf{v}$ with $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$,

$$\Big(\mathbf{u}'AB\mathbf{v}\Big)^2 \le \lambda_{\max}\big(B'A'AB\big) \le \mathrm{trace}\big(B'A'AB\big) = \mathrm{trace}\big(A'ABB'\big). \tag{42}$$

Let $c_i \in \mathbb{R}$ with $\sum_{i=1}^n c_i^2 = 1$ and $\mathbf{u}_i \in \mathbb{R}^J$ with $\|\mathbf{u}_i\| = 1$. Since $\mathbf{u}_i'M_i\mathbf{u}_i \le \lambda_{\max}(M_i)$,

$$\Big\|\sum_{i=1}^n c_i M_i^{1/2}\mathbf{u}_i\Big\|^2 \le \max_{1\le i\le n}\lambda_{\max}\Big(M_i\Big) + \sum_{1\le i\ne k\le n} c_ic_k\mathbf{u}_i'M_i^{1/2}M_k^{1/2}\mathbf{u}_k. \tag{43}$$

Since $\sum_i c_i^2 = 1$, $\sum_{1\le i\ne k\le n} c_i^2c_k^2 = 1 - \sum_i c_i^4 \le 1 - 1/n$, so that by Cauchy-Schwarz and (42)

$$\Big(\sum_{1\le i\ne k\le n} c_ic_k\mathbf{u}_i'M_i^{1/2}M_k^{1/2}\mathbf{u}_k\Big)^2 \le \frac{n-1}{n}\sum_{1\le i\ne k\le n}\mathrm{trace}\Big(M_iM_k\Big).$$

Inserting this into (43), we prove (41) via Lemma 1. The proof of Lemma 2 is complete.

A random matrix $A \equiv (a_{jk})_{J\times J}$ is exchangeable if for all permutations $\{\ell_1,\ldots,\ell_J\}$ of $\{1,\ldots,J\}$ the vectors $(a_{\ell_j,\ell_k}, j\le J, k\le J)$ and $(a_{jk}, j\le J, k\le J)$ are identically distributed. Under assumption A, the projections Q_i in (15) are independent and exchangeable.

Lemma 3. *Let A be an exchangeable matrix with $A\mathbf{e}=0$, where $\mathbf{e}\equiv(1,\ldots,1)'\in\mathbb{R}^J$. Then,*

$$EA = \frac{E\,\mathrm{trace}(A)}{J-1} I_{J,0}, \quad I_{J,0}\equiv I_J - \frac{\mathbf{e}\mathbf{e}'}{J}. \tag{44}$$

Proof. The exchangeability implies that the matrix EA has identical diagonal elements (say a_0^*) and identical off-diagonal elements (say a_1^*). The value of a_0^* is determined by $a_0^* = E\mathrm{trace}(A)/J$. Since $A\mathbf{e}=0$, $\mathbf{e}'(EA)\mathbf{e}=0$, so that $Ja_0^* + J(J-1)a_1^* = 0$. Thus, $a_1^* = -E\mathrm{trace}(A)/\{J(J-1)\}$ and (44) follows by algebra. The proof of Lemma 3 is complete.

Proof of Proposition 1. Let $\widetilde{z}_i = Z_n^{-1/2} z_i$. It follows from (17) and (31) that for $\boldsymbol{b}\in\Omega_0^{J\times d}$

$$\begin{aligned}\widetilde{\Lambda}_{J,n}^{-1/2}\widehat{\Lambda}_{J,n}\widetilde{\Lambda}_{J,n}^{-1/2}\boldsymbol{b} &= \widetilde{\Lambda}_{J,n}^{-1/2}\widehat{\Lambda}_{J,n}\Big\{\boldsymbol{b}(Z_n/n)^{-1/2}\Big\} \\ &= \widetilde{\Lambda}_{J,n}^{-1/2}\sum_{i=1}^n (I_J - Q_i)\boldsymbol{b}(nZ_n)^{-1/2} z_i z_i' = \sum_{i=1}^n (I_J - Q_i)\boldsymbol{b}\widetilde{z}_i\widetilde{z}_i'.\end{aligned}$$

Since $\sum_{i=1}^n \widetilde{z}_i\widetilde{z}_i' = I_d$, this implies that

$$I_{J,0}\otimes I_d - \widetilde{\Lambda}_{J,n}^{-1/2}\widehat{\Lambda}_{J,n}\widetilde{\Lambda}_{J,n}^{-1/2} = \sum_{i=1}^n (I_{J,0} - I_J + Q_i)\otimes \widetilde{z}_i\widetilde{z}_i' = \sum_{i=1}^n \widehat{A}_i\otimes\widetilde{z}_i\widetilde{z}_i'$$

with $\widehat{A}_i \equiv I_{J,0} - I_J + Q_i = I_{J,0}Q_iI_{J,0}$ due to $Q_i\mathbf{e}=\mathbf{e}$. Since tensor products are linear operators, Lemma 2 applies with $M_i = \widehat{A}_i\otimes\widetilde{z}_i\widetilde{z}_i'$ and $\lambda_{\max}(M_i) = \|\widetilde{z}_i\|^2 = z_i'Z_n^{-1}z_i$. Thus, (32) holds and it remains to prove (33) with $(\zeta^*)^2 = \sum_{i\neq k}\mathrm{trace}(M_iM_k)$.

Since $(A_1\otimes B_1)(A_2\otimes B_2)\boldsymbol{b} = A_1A_2\boldsymbol{b}B_1B_2$ and $\mathrm{trace}(A\otimes B) = \mathrm{trace}(A)\mathrm{trace}(B)$,

$$\mathrm{trace}\big(M_iM_k\big) = \mathrm{trace}\big(\widehat{A}_i\widehat{A}_k\big)(\widetilde{z}_i'\widetilde{z}_k)^2. \tag{45}$$

Since Q_i are projections from $\mathbb{R}^J$ to $\{f(\boldsymbol{x}_i) : f\in S_i\}$, by Condition I, $\widehat{A}_i = I_{J,0}Q_iI_{J,0}$ are independent random matrices with $\mathrm{trace}(\widehat{A}_i) = \mathrm{trace}(Q_i) - 1 = \widehat{K}_i$. Thus, by Lemma 3

$$E\big\{\mathrm{trace}\big(\widehat{A}_i\widehat{A}_k\big)\big\} = \mathrm{trace}\big(E\widehat{A}_iE\widehat{A}_k\big) \le \frac{E\widehat{K}_iE\widehat{K}_k}{J-1}, \quad i\neq k.$$

This and (45) imply the identity in (33). The inequality in (33) follows, since $\sum_{i\neq k}(\widetilde{z}_i'\widetilde{z}_k)^2 = \sum_i\big\{\widetilde{z}_i'\big(\sum_k\widetilde{z}_k\widetilde{z}_k'\big)\widetilde{z}_i - \|\widetilde{z}_i\|^4\big\} = \sum_i\big\{\|\widetilde{z}_i\|^2 - \|\widetilde{z}_i\|^4\big\} \le d - d^2/n$ due to $\sum_i\widetilde{z}_i\widetilde{z}_i' = I_d$. The proof of Proposition 1 is complete, as Part (ii) is an immediate consequence of Part (i).

Lemma 4. *Suppose Condition I holds. Then,*

$$E\left\|\sum_{i=1}^n (I_J - Q_i)f_i(\boldsymbol{x}_i)z_i'Z_n^{-1/2}\right\|_2^2 = \sum_{i=1}^n E\big\|(I_J - Q_i)f_i(\boldsymbol{x}_i)\big\|^2 z_i'Z_n^{-1}z_i'. \tag{46}$$

Proof. Since $\psi_0(x) = 1$ is a member of S_i, $Q_i\mathbf{e} = 0$ and $\mathbf{e}'(I_J - Q_i) = 0$. Since the components of the vector $\mathbf{v}_i \equiv (I_J - Q_i)f_i(\boldsymbol{x}_i)$ are exchangeable variables, the components of $E\mathbf{v}_i$ are all equal, so that $E\mathbf{v}_i = J^{-1}\mathbf{e}\mathbf{e}'E\mathbf{v}_i = 0$. This implies (46), since $\mathbf{v}_i z_i' Z_n^{-1/2}$ are independent matrices and $\|\mathbf{v}_i z_i' Z_n^{-1/2}\|_2^2 = \|\mathbf{v}_i\|^2 z_i' Z_n^{-1} z_i$. The proof is complete.

Proof of Theorem 2. Since $\widetilde{\boldsymbol{b}}_{J,n} \in \Omega_0^{J\times d}$, it suffices to consider $\boldsymbol{b} \in \Omega_0^{J\times d}$. The $J \times d$ matrix $\boldsymbol{b}Z_n^{-1/2} \in \Omega_0^{J\times d}$ can be written as $\boldsymbol{b}Z_n^{-1/2} = \sum_{k=1}^d \mathbf{u}_k a_k'$ with orthonormal $\mathbf{u}_k \in \mathbb{R}^J$ satisfying $\mathbf{u}_k'\mathbf{e} = 0$ and $a_k \in \mathbb{R}^d$ satisfying $\sum_{k=1}^d \|a_k\|^2 = \text{trace}(\boldsymbol{b}Z_n^{-1}\boldsymbol{b}')$. Let $\boldsymbol{b}^* \equiv \sum_{k=1}^d \mathbf{u}_k^* a_k'$ be a random permutation of rows of $\boldsymbol{b}Z_n^{-1/2}$ independent of observations and E^* be the expectation given $\{\boldsymbol{x}_i, i \le n\}$. Under Condition I, the joint distribution of the elements of $\widetilde{\boldsymbol{b}}_{J,n}Z_n^{1/2}$ is invariant under permutations of rows. Thus, due to the concavity of $\min(c, x)$, $c > 0$,

$$\begin{aligned} E\min\Big(c, \text{trace}^2(\boldsymbol{b}'\widetilde{\boldsymbol{b}}_{J,n})\Big) &= E\min\Big(c, \text{trace}^2\big\{(\boldsymbol{b}^*)'\widetilde{\boldsymbol{b}}_{J,n}Z_n^{1/2}\big\}\Big) \\ &\le E\min\Big(c, E^*\text{trace}^2\big\{(\boldsymbol{b}^*)'\widetilde{\boldsymbol{b}}_{J,n}Z_n^{1/2}\big\}\Big). \end{aligned} \tag{47}$$

Since $\mathbf{u}_k^*(\mathbf{u}_\ell^*)'$ are exchangeable $J \times J$ matrices with $\mathbf{u}_k^*(\mathbf{u}_\ell^*)'\mathbf{e} = 0$, by Lemma 3, $E^*\mathbf{u}_k^*(\mathbf{u}_\ell^*)' = \text{trace}(\mathbf{u}_k\mathbf{u}_\ell')(J-1)^{-1}I_{J,0}$. Since $\text{trace}(\mathbf{u}_k\mathbf{u}_\ell') = \mathbf{u}_\ell'\mathbf{u}_k = I\{k = \ell\}$, this implies

$$\begin{aligned} E^*\text{trace}^2\Big((\boldsymbol{b}^*)'\widetilde{\boldsymbol{b}}_{J,n}Z_n^{1/2}\Big) &= E^*\Big(\sum_{k=1}^d (\mathbf{u}_k^*)'\widetilde{\boldsymbol{b}}_{J,n}Z_n^{1/2}a_k\Big)^2 = \sum_{k=1}^d a_k'Z_n^{1/2}\widetilde{\boldsymbol{b}}_{J,n}'\frac{I_{J,0}}{J-1}\widetilde{\boldsymbol{b}}_{J,n}Z_n^{1/2}a_k \\ &\le \sum_{k=1}^d \|a_k\|^2\frac{\text{trace}(\widetilde{\boldsymbol{b}}_{J,n}Z_n\widetilde{\boldsymbol{b}}_{J,n}')}{J-1} = \frac{\text{trace}(\boldsymbol{b}Z_n^{-1}\boldsymbol{b}')}{(J-1)/n}\Big\|\widetilde{\Lambda}_{J,n}^{1/2}\widetilde{\boldsymbol{b}}_{J,n}\Big\|_2^2, \end{aligned} \tag{48}$$

with the $\widetilde{\Lambda}_{J,n}$ in (31), due to $\text{trace}(\widetilde{\boldsymbol{b}}_{J,n}Z_n\widetilde{\boldsymbol{b}}_{J,n}')/n = \big\|\widetilde{\Lambda}_{J,n}^{1/2}\widetilde{\boldsymbol{b}}_{J,n}\big\|_2^2$. In the event of $\Pi_{J,n} = 0$

$$\begin{aligned} \Big\|\widetilde{\Lambda}_{J,n}^{1/2}\widetilde{\boldsymbol{b}}_{J,n}\Big\|_2^2 &= \Big\|\widetilde{\Lambda}_{J,n}^{1/2}\widehat{\Lambda}_{J,n}^{-1}\Big\{\frac{1}{n}\sum_{i=1}^n (I_J - Q_i)f_i(\boldsymbol{x}_i)z_i'\Big\}\Big\|_2^2 \\ &\le \lambda_{\max}^2\Big(\widetilde{\Lambda}_{J,n}^{1/2}\widehat{\Lambda}_{J,n}^{-1}\widetilde{\Lambda}_{J,n}^{1/2}\Big)\Big\|n^{-1/2}\sum_{i=1}^n (I_J - Q_i)f_i(\boldsymbol{x}_i)z_i'Z_n^{-1/2}\Big\|_2^2, \end{aligned}$$

so that by Proposition 1 (ii), Lemma 4 and Condition IV

$$\begin{aligned} \frac{\big\|\widetilde{\Lambda}_{J,n}^{1/2}\widetilde{\boldsymbol{b}}_{J,n}\big\|_2^2}{(J-1)/n} &= O_P(1)(J-1)^{-1}E\Big\|\sum_{i=1}^n (I_J - Q_i)f_i(\boldsymbol{x}_i)z_i'Z_n^{-1/2}\Big\|_2^2 \\ &= O_P(1)\sum_{i=1}^n E\big\|(I_J - Q_i)f_i(\boldsymbol{x}_i)\big\|^2\frac{z_i'Z_n^{-1}z_i}{J-1} = O_P(\rho_{J,n}^*) \end{aligned} \tag{49}$$

since $\sum_{i=1}^n z_i'Z_n^{-1}z_i = d$. We obtain (28) by inserting (48) and (49) into (47) with $c = \text{trace}(\boldsymbol{b}Z_n^{-1}\boldsymbol{b}')c_{J,n}$.

If $\text{Var}(\boldsymbol{\varepsilon}_i) \ge \sigma_*^2 I_J$, then $V_i \ge \sigma_*^2(I_J - Q_i)$ and $V_{J,n} \ge \sigma_*^2\widehat{\Lambda}_{J,n}$, so that in the event of $\Pi_{J,n} = 0$

$$\frac{\sigma_{J,n}^2(\boldsymbol{b})}{\sigma_*^2} \ge n^{-1}\text{trace}(\boldsymbol{b}'\widehat{\Lambda}_{J,n}^{-1}\boldsymbol{b}) \ge \text{trace}(\boldsymbol{b}'\widetilde{\Lambda}_{J,n}^{-1}\boldsymbol{b}) = \text{trace}(\boldsymbol{b}Z_n^{-1}\boldsymbol{b}')$$

due to $n\widehat{\Lambda}_{J,n} = \sum_{i=1}^n (I_J - Q_i) \otimes z_i z_i' \le \sum_{i=1}^n I_{J,0} \otimes z_i z_i' = n\widetilde{\Lambda}_{J,n}$ as nonnegative-definite linear operators. This and (28) imply (29) due to $\rho^*_{J,n} \to 0$ in Condition IV and $P\{\Pi_{J,n} = 0\} \to 1$ in Proposition 1 (ii). The proof of Theorem 2 is complete.

Proof of Theorem 3. Since Q_i is a projection, $Q_i = \sum_{j=0}^{\widehat{K}_i} \mathbf{u}_j \mathbf{u}_j'$ for certain orthonormal $\mathbf{u}_j \in \mathbb{R}^J$, $\mathbf{u}_0 = J^{-1/2}\mathbf{e}$. Let E^* be the conditional expectation given $\{\boldsymbol{x}_i, i \le n\}$. By Theorem 1,

$$E^*\Big\|Q_i\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)z_i\Big\|^2 = \sum_{j=1}^{\widehat{K}_i} E^*\Big\|\mathbf{u}_j'\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)z_i\Big\|^2 = \Big\|Q_i\widetilde{\boldsymbol{b}}_{J,n}z_i\Big\|^2 + \frac{1}{n}\sum_{j=1}^{\widehat{K}_i}\Big\|V_{J,n}^{1/2}\widehat{\Lambda}_{J,n}^{-1}(\mathbf{u}_j z_i')\Big\|_2^2. \quad (50)$$

Since $\mathrm{Var}(\boldsymbol{\varepsilon}_i) \le (\sigma^*)^2 I_J$, $V_{J,n} \le (\sigma^*)^2\widehat{\Lambda}_{J,n}$, so that by (31) and Proposition 1 (ii)

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{\widehat{K}_i}\Big\|V_{J,n}^{1/2}\widehat{\Lambda}_{J,n}^{-1}(\mathbf{u}_j z_i')\Big\|_2^2 &\le \frac{(\sigma^*)^2}{n}\sum_{j=1}^{\widehat{K}_i}\Big\|\widehat{\Lambda}_{J,n}^{-1/2}(\mathbf{u}_j z_i')\Big\|_2^2 = (\sigma^*)^2\sum_{j=1}^{\widehat{K}_i}\Big\|\widehat{\Lambda}_{J,n}^{-1/2}\widetilde{\Lambda}_{J,n}^{1/2}(\mathbf{u}_j z_i' Z_n^{-1/2})\Big\|_2^2 \\ &= O_P(1)(\sigma^*)^2\sum_{j=1}^{\widehat{K}_i}\Big\|(\mathbf{u}_j z_i' Z_n^{-1/2})\Big\|_2^2 = O_P(1)(\sigma^*)^2\widehat{K}_i z_i' Z_n^{-1} z_i.\end{aligned}$$

It follows from (31) and (49) that in the event of $\Pi_{J,n} = 0$,

$$\Big\|\widetilde{\boldsymbol{b}}_{J,n}z_i\Big\|^2 = n\Big\|\Big(\widetilde{\Lambda}_{J,n}^{1/2}\widetilde{\boldsymbol{b}}_{J,n}\Big)Z_n^{-1/2}z_i\Big\|^2 \le O_P(J)\rho^*_{J,n}\Big\|Z_n^{-1/2}z_i\Big\|^2.$$

Inserting the above two inequalities into (50), we find that uniformly in $i \le n$

$$J^{-1}E^*\Big\|Q_i\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big)z_i\Big\|^2 = O_P(1)z_i' Z_n^{-1} z_i\Big(\rho^*_{J,n} + (\sigma^*)^2 E\widehat{K}_i/J\Big), \quad (51)$$

since $P\{\Pi_{J,n} = 0\} \to 1$, with $z_i' Z_n^{-1} z_i \le \kappa^* \le 1$. It follows from the definition of $K^*_{J,n}$ and $\rho^*_{J,n}$ in Conditions III and IV that $\max_{i\le n} E\|Q_i\{f_i(\boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i\} - f_i(\boldsymbol{x}_i)\|^2 \le O(\rho^*_{J,n} + (\sigma^*)^2 K^*_{J,n}/J)$. Thus, by (30) and (51), $\|\widehat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|^2/J = O_P(\rho^*_{J,n} + (\sigma^*)^2 K^*_{J,n}/J)$ uniformly in $i \le n$. The proof of Theorem 3 is complete.

Table 1: Simulation results for Model 1. $10,000\times$ Summary of MSE. The true normalization curve is the horizontal line at 0. The expression levels of up- and down-regulated genes are symmetric: $\alpha_1 = \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$.

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	3.74	51.59	75.08	88.88	106.20	4980.00
	Lowess	3.38	50.72	72.77	87.89	105.10	7546.00
	Splines	7.08	58.93	85.35	98.25	121.10	4703.00
$\alpha = 0.06$	TW-SRM	6.53	50.03	74.74	93.92	107.30	5120.00
	Lowess	9.09	50.89	73.93	91.87	106.10	6230.00
	Splines	8.95	61.34	89.03	105.60	126.10	6480.00

Table 2: Simulation results for Model 2. $10,000\times$ Summary of MSE. The true normalization curve is the horizontal line at 0. But the percentages of up- and down-regulated genes are different:$\alpha_1 = 3\alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$.

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	5.36	58.04	71.01	83.17	102.50	1416.00
	Lowess	8.86	67.69	95.80	107.40	131.00	1747.00
	Splines	8.91	65.53	94.40	110.40	135.10	1704.00
$\alpha = 0.06$	TW-SRM	6.66	47.85	68.55	78.49	97.50	1850.40
	Lowess	6.45	59.54	87.08	99.00	123.90	1945.10
	Splines	6.74	59.23	86.58	98.67	123.30	1813.10

Table 3: Simulation results for Model 3. $10,000\times$ Summary of MSE. There are non-linear and intensity dependent dye biases. The expression levels of up- and down-regulated genes are symmetric: $\alpha_1 = \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$.

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	5.56	46.15	66.72	7.23	93.91	1898.00
	Lowess	6.71	51.07	74.23	88.79	107.50	3353.00
	Splines	5.90	53.83	76.91	88.64	108.60	1750.00
$\alpha = 0.06$	TW-SRM	6.64	57.26	85.79	102.80	126.40	2290.00
	Lowess	7.39	57.19	85.47	107.70	128.10	2570.00
	Splines	9.37	69.26	102.80	122.80	148.50	2230.00

Table 4: Simulation results for Model 4. $10,000\times$ Summary of MSE. There is non-linear and intensity dependent dye bias. The percentages of up- and down-regulated genes are different: $\alpha_1 = 3\alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$.

		min	1th Quartile	median	mean	3rd Quartile	max
$\alpha = 0.01$	TW-SRM	5.89	51.26	74.53	85.89	107.20	2810.00
	Lowess	9.29	68.30	101.60	118.60	140.00	4088.00
	Splines	9.68	67.85	98.82	119.80	141.00	2465.00
$\alpha = 0.06$	TW-SRM	4.96	54.12	79.92	98.79	122.70	2130.00
	Lowess	6.49	71.54	113.90	130.90	169.50	2474.00
	Splines	5.77	65.46	107.57	128.40	171.60	1898.00

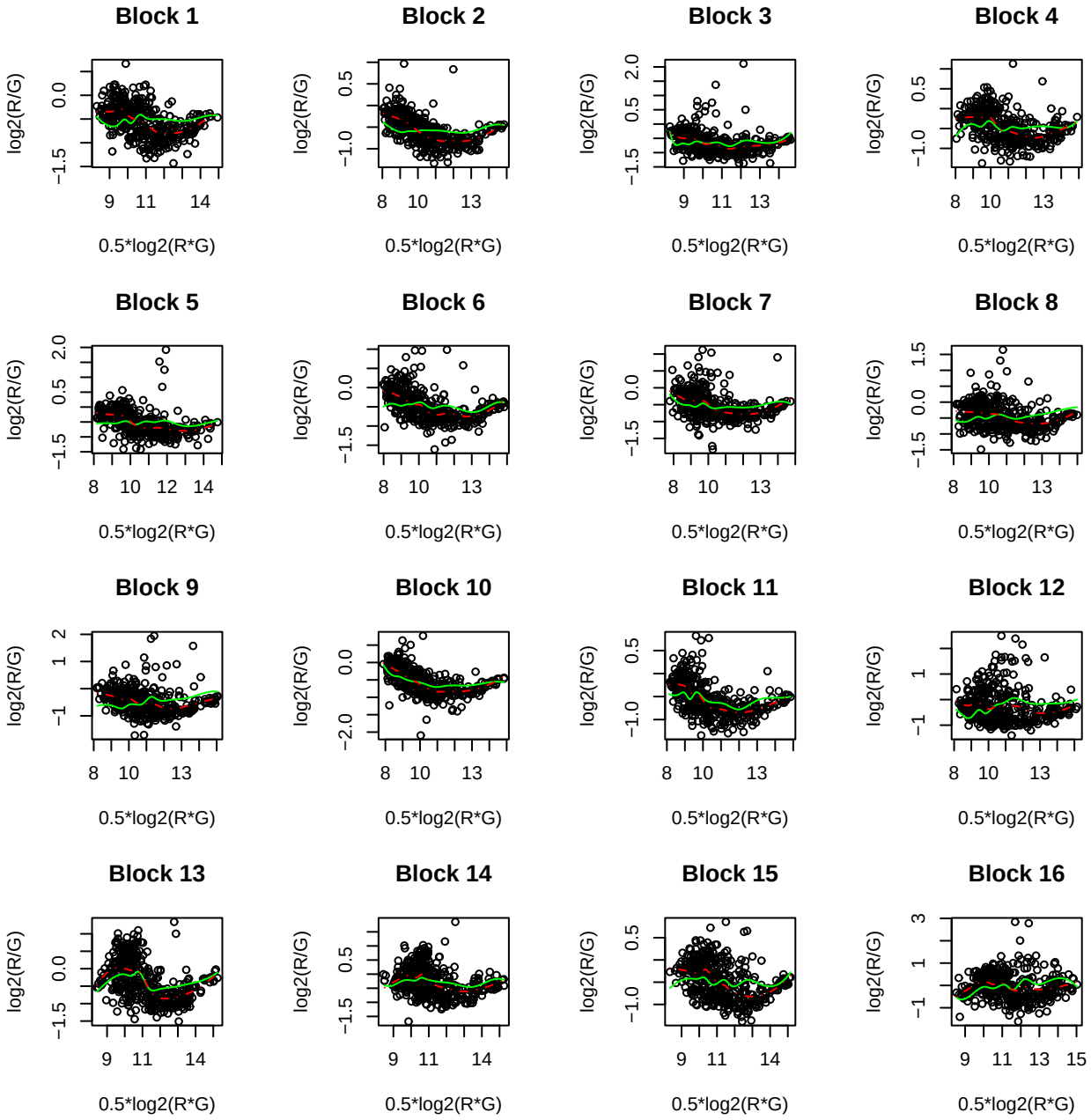

Figure 1: Apo AI data: Comparison of normalization curves in the 16 blocks of the array from one knock-out mouse in the treatment group. Solid (green) line: normalization curve based on TW-SLM; Dashed (red) line: normalization curve based on *lowess*

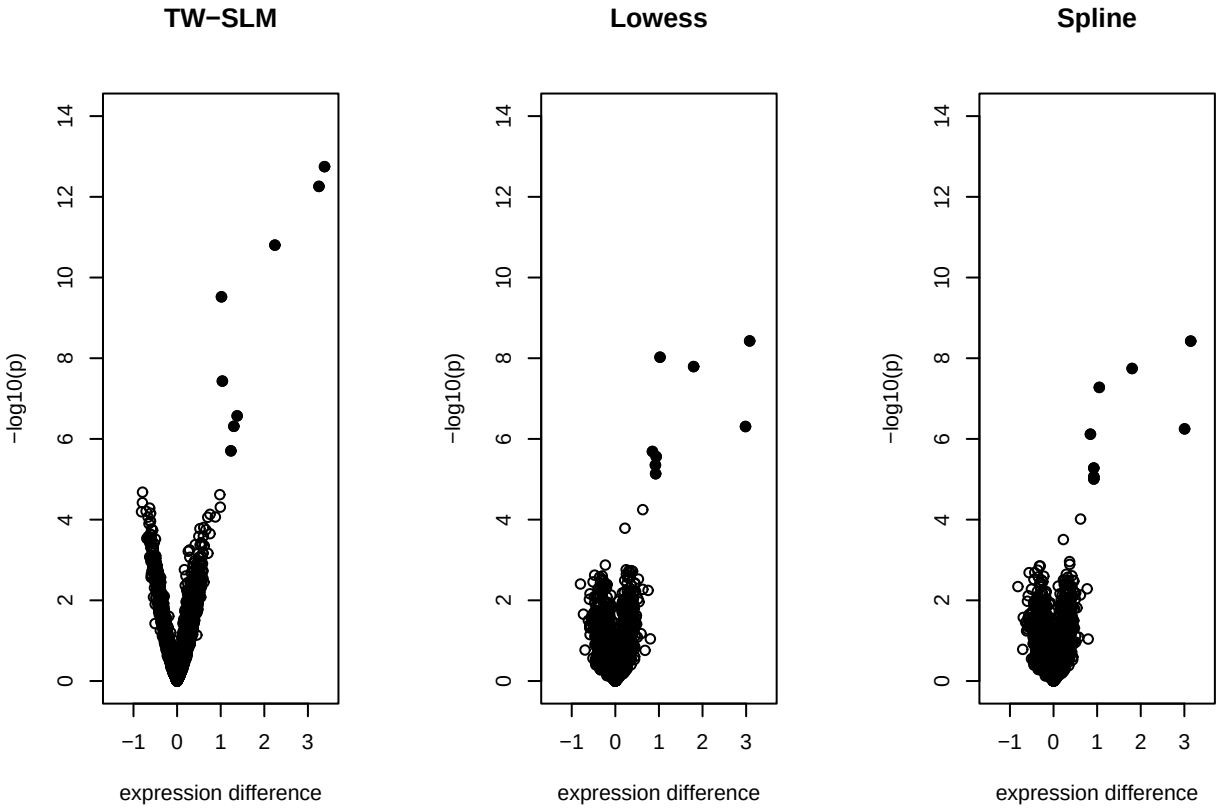

Figure 2: Volcano plots: Scatter plot of $-\log_{10}(\text{p-value})$ versus estimated mean expression value. The left panel shows the volcano plot based on the TW-SLM; the middle panel shows the plot based on the *Lowess* method; the right panel shows the result based on the *Spline* method.

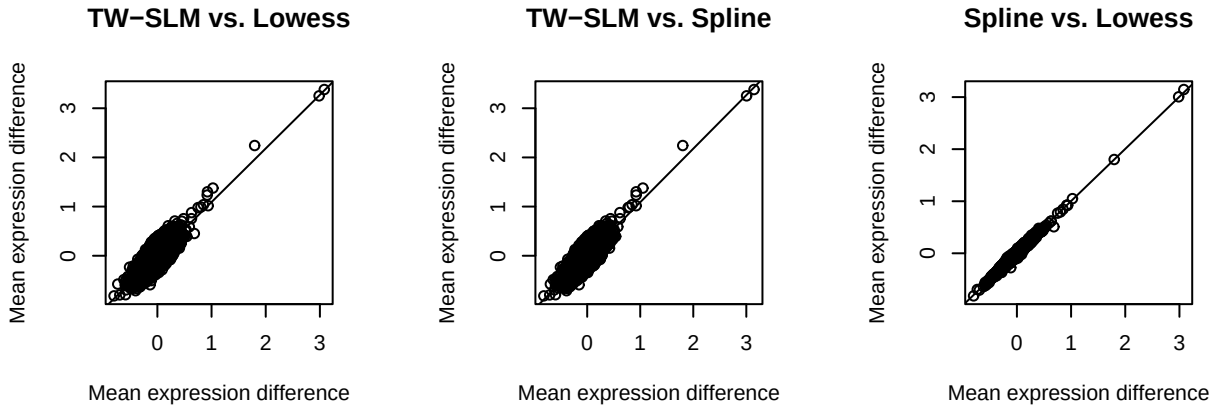

Figure 3: Comparison of normalized expression values. Left panel: the scatter plot of normalized mean expression differences based on TW-SLM versus those based on*Lowess*. Middle panel: the scatter plot of normalized mean expression differences based on TW-SLM versus those based on *Spline*. Right panel: the scatter plot of normalized mean expression differences based on *Spline* versus those based on *Lowess*.

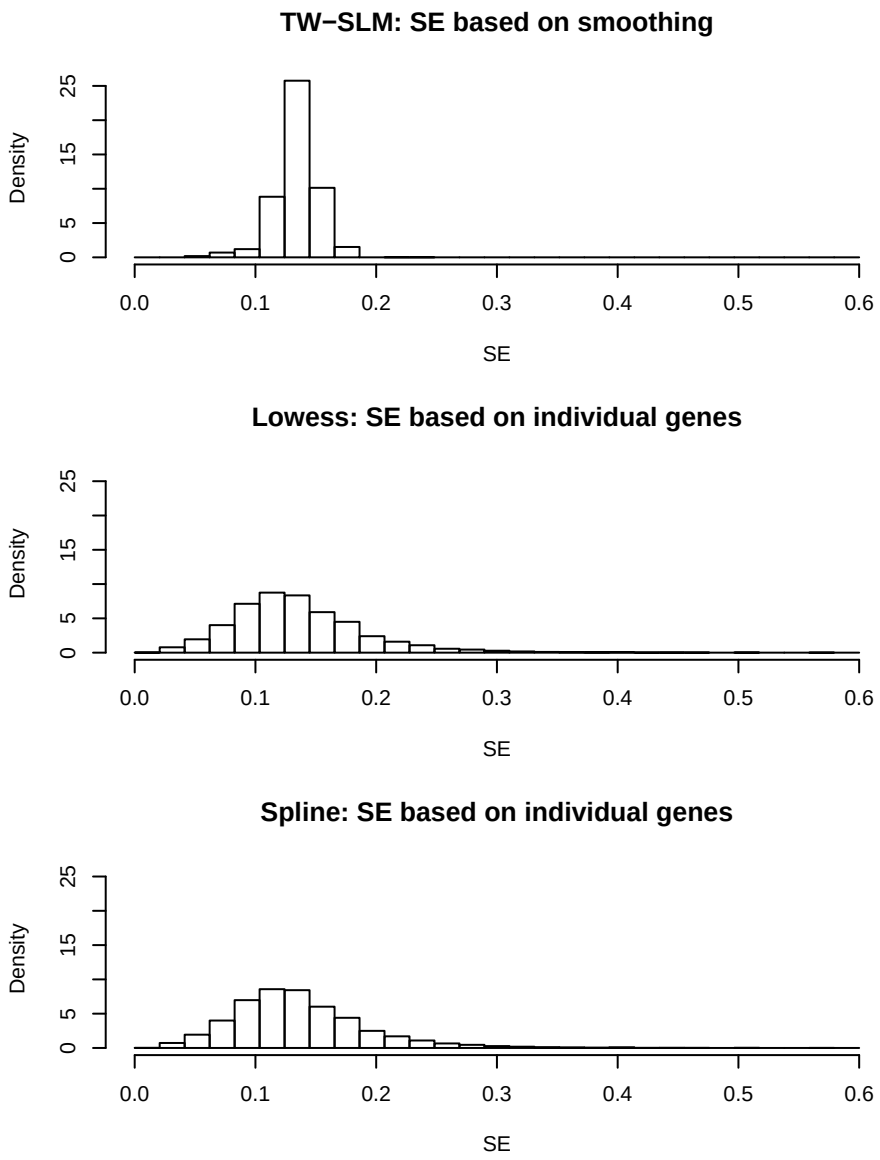

Figure 4: Comparison of variance estimation methods. Top panel: the histogram of SE estimated based on smoothing as described in Section 4.2. Middle panel: SE estimated based on individual genes using the *Lowess* method. Bottom panel: SE estimated based on individual genes using the *Spline* method.