Semilinear High-Dimensional Model for Normalization of Microarray Data: A Theoretical Analysis and Partial Consistency

Jianqing FAN, Heng PENG, and Tao HUANG

Normalization of microarray data is essential for removing experimental biases and revealing meaningful biological results. Motivated by a problem of normalizing microarray data, a semilinear in-slide model (SLIM) has been proposed. To aggregate information from other arrays, SLIM is generalized to account for across-array information, resulting in an even more dynamic semiparametric regression model. This model can be used to normalize microarray data even when there is no replication within an array. We demonstrate that this semiparametric model has a number of interesting features. The parametric component and the nonparametric component that are of primary interest can be consistently estimated, the former having a parametric rate and the latter having a nonparametric rate, whereas the nuisance parameters cannot be consistently estimated. This is an interesting extension of the partial consistent phenomena, which itself is of theoretical interest. The asymptotic normality for the parametric component and the rate of convergence for the nonparametric component are established. The results are augmented by simulation studies and illustrated by an application to the cDNA microarray analysis of neuroblastoma cells in response to the macrophage migration inhibitory factor.

KEY WORDS: Aggregation; cDNA microarray; In-slide replications; Normalization; Partial consistency; Semiparametric models; SLIM.

1. INTRODUCTION

DNA microarrays monitor the expression of tens of thousands of genes in a single hybridization experiment using oligonucleotide or cDNA probes. The technique has been widely used in many biomedical research and biological studies. A challenge in analyzing microarray data is the systematic biases due to variations in experimental conditions, such as the efficiency of dye incorporation, intensity effect, DNA concentration on arrays, amount of mRNA, variability in reverse transcription, and batch variation, among others. Normalization is required to remove the systematic effects of confounding factors so that meaningful biological results can be obtained.

Several useful normalization techniques aim to remove the systematic biases such as the dye, intensity, and print-tip block effects. The simplest such technique is the global normalization method featured in software packages such as GenePix4.0 and analyzed by Kroll and Wölfl (2002). Such a technique implicitly assumes that there is no print-tip block effect and no intensity effect. Without such an assumption, the method is statistically biased. The "lowess" method of Dudoit et al. (2002) significantly relaxes the foregoing assumption. But it assumes that the average expression levels of up-regulated and down-regulated genes at each intensity level are about the same in each print-tip block. This assumption was further relaxed by Tseng, Oh, Rohlin, Liao, and Wong (2001) to only a subset of more conservative genes based on a rank-invariant selection method. As admitted by Tseng et al. (2001), the method is not expected to be useful when there are far more genes that are up-regulated (or down-regulated). Such situations can occur when cells are treated with some reagents (Grolleau et al. 2002; Fan, Tam, Vande Woude, and Ren 2004). In an attempt to further relax the foregoing biological assumption, Huang, Wang, and Zhang (2003) and Huang and Zhang (2003) introduced a semilinear model to account for the intensity effect and to aggregate information from other arrays to assess the intensity effect. The method is expected to work well when the gene effect is the same across arrays.

In an attempt to further relax the aforementioned statistical and biological assumptions in the cDNA microarray normalization, Fan et al. (2004) developed a new method of estimating the intensity and print-tip block effects by aggregating information from the replications within a cDNA array. cDNA microarray chips are usually constructed by dipping a printer head containing 16 spotting pins into a 96-well plate containing cDNA solutions, printing these 16 spots on the slide, washing the spotting pins, dipping them into different 16 wells and printing again, and so on (see Craig, Black, and Doerge 2003 for details). For the specific designs of cDNA microarrays used by Fan et al. (2004), there are 111 clones that are printed twice on the cDNA chips. The locations of these 222 replications appear random in the 32 blocks (see Sec. 4.2). In other words, replications are achieved not by printing twice the same 16 spots on a printer head, but by constructing the wells in plates themselves. The replications of the clones in the cDNA chips contain much information about systematic biases, such as the print-tip block and intensity effects. In fact, for two identical clones of cDNA in the same slide, apart from the random errors, the expression ratios should be the same. Observed differences of expression ratios tell us a lot of information about the print-tip block and intensity effects. The seemingly random patterns of replications enable one to unveil the print-tip block effect. This cannot be achieved if only the same 16 spots in a well plate are printed twice.

To put the foregoing problem into a statistical framework, let G be the number of genes and let I be the number of replications of gene g within an array. (I should depend on g, because most do not have replications.) Following Dudoit et al. (2002), let R_{gi} and G_{gi} be the red (Cy5) and green (Cy3) intensities of the

Jianqing Fan is Professor (E-mail: *jqfan@princeton.edu*) and Heng Peng is Postdoctoral Fellow (E-mail: *pheng@princeton.edu*), Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Tao Huang is Postdoctoral Associate, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06511 (E-mail: *t.huang@yale.edu*). This work was supported in part by National Science Foundation grant DMS-03-54223 and National Institutes of Health grant R01-HL69720. The authors thank the editor, associate editor, and three anonymous referees for helpful comments that led to significant improvement of the presentation of the article.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214504000001781

gth gene in the ith replication. Let Y_{gi} be the log-intensity ratio of red over green channels of the gth gene in the ith replication, and let X_{gi} be the corresponding average of the log intensities of the red and green channels, that is,

$$Y_{gi} = \log_2 \frac{R_{gi}}{G_{gi}}, \qquad X_{gi} = \frac{1}{2}\log_2(R_{gi}G_{gi}).$$

To model the intensity and print-tip block effects, we consider the following high-dimensional partial linear model for microarray data:

$$Y_{gi} = \alpha_g + \beta_{r_{gi}} + \gamma_{c_{gi}} + m(X_{gi}) + \varepsilon_{gi}, \tag{1}$$

where α_g is the treatment effect associated with the gth gene; r_{gi} and c_{gi} are the row and column of the print-tip block where the gth gene of the ith replication resides; β and γ are the row and column effects with constraints

$$\sum_{i=1}^{r} \beta_i = 0 \qquad \text{and} \qquad \sum_{j=1}^{c} \gamma_j = 0,$$

where r and c are the number of rows and columns of the print-tip block; $m(\cdot)$ with constraint $Em(X_{gi}) = 0$ is a smoothing function of X representing the intensity effect; and ε_{gi} is random error with mean 0 and variance $\sigma^2(X_{gi})$.

In our illustrative example in Section 4.2, there are 19,968 genes in an array, residing in 8×4 blocks with $r = 8$ and $c = 4$. Among those, are 111 genes with two replications. For those genes without replications, because α_g is free, they provide no information about the parameters β and γ and the smooth function $m(\cdot)$. We need to estimate the parameters from the genes with replications. With a slight abuse of notation, for this illustrative example, $G = 111$ and $I = 2$.

For the normalization purpose, our aim is to find a good estimate of the print-tip block and intensity effects. Let $\widehat{\beta}$, $\widehat{\gamma}$, and $\widehat{m}(\cdot)$ be good estimates for model (1). Then the normalization is to compute

$$Y_g^* = Y_g - \widehat{\beta}_{r_g} - \widehat{\gamma}_{c_g} - \widehat{m}(X_g) \tag{2}$$

for all genes. Interpolations and extrapolations are needed to expedite the computation when $m(\cdot)$ is estimated over a set of fine grid points. According to model (1), $Y_g^* \approx \alpha_g + \varepsilon_g$, in which the effects of confounding factors have been removed. Thus, as far as the process of the normalization is concerned, the parameters β and γ and the function $m(\cdot)$ are of primary interest and the parameters $\{\alpha_g\}$ are nuisance parameters. Of course, in the analysis of treatment effect on genes, the parameters $\{\alpha_g\}$ represent biological fold changes and are of primary interest.

Model (1) has a much wider spectrum of applications than at first appearance. First, if there is no replication within an array but there are four (say) replications across arrays, by imaging a super array that contains these four arrays, "within-array" replications are artificially created. In this case I is the number of arrays, and G is the number of genes per array. The basic assumption behind this method is that the treatment effect on the genes remains the same across arrays. This is not an unreasonable assumption when the same experiment is repeated several times. Second, by removing the row and column effects and applying model (1) directly to each block of microarrays, resulting in

$$Y_{gib} = \alpha_g + m_b(X_{gi}) + \varepsilon_{gib}, \tag{3}$$

the model allows nonadditive effect between the intensity and blocks. [The index b can be removed from model (3), and hence this model becomes a submodel of (1).] In this case G is the number of genes within a block. For example, if there are 624 genes in a block and 4 replications of arrays, then $G = 624$ and $I = 4$. Third, the idea can also be adapted to normalize the Affymetrix arrays by imaging "treatment" and "control" arrays as the outputs from green channels and red channels. This will enable us to remove intensity effects in the Affymetrix arrays. Finally, by thinking of rows as blocks and deleting the column effects, model (1) can accommodate nonadditive column and row effects. The additivity in model (1) is to facilitate the applications in which G is relatively small.

The challenge of our problem is that the number of nuisance parameters is large. In fact, for many practical situations, $I = 2$ and G can be large, on the order of hundreds or larger. So our asymptotic results are based on the assumption that $G \to \infty$. This is in contrast with the assumption of Huang and Zhang (2003), where I tends to infinity. The number of nuisance parameters in (1) grows with the sample size. In our illustrative example, half of the parameters are nuisance ones. The question is whether the parameters of primary interest can be consistently estimated in the presence of a large number of nuisance parameters and how much it costs to estimate these parameters. Such a problem is poorly understood, and a thorough investigation is needed.

To provide more insight into the problem, consider writing model (1) in the matrix form as

$$\mathbf{Y}_n = \mathbf{B}_n\boldsymbol{\alpha}_n + \mathbf{Z}_n\boldsymbol{\beta} + \mathbf{M} + \boldsymbol{\epsilon}_n, \qquad n = G \times I, \tag{4}$$

where $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^T$, $\mathbf{B}_n$ is an $n \times G$ design matrix, $\mathbf{Z}_n$ is an $n \times d$ random matrix with d being the sum of the numbers of rows and columns, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r, \gamma_1, \ldots, \gamma_c)$ is the print-tip block effect, $\mathbf{M} = (m(X_1), \ldots, m(X_n))^T$ is the intensity effect, and $\boldsymbol{\epsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^T$. The theory on the partial linear model is usually based on the assumption that G is fixed or at least G/n tends to 0 at certain rate (see Härdle, Liang, and Gao 2000). However, in our application, $\boldsymbol{\alpha}_n$ cannot be consistently estimated, because $G/n = 1/I$ in (4). It is not clear whether the parameters $\boldsymbol{\beta}$ and the function $m(\cdot)$ can be consistently estimated. The answer might not be affirmative for general matrix $\mathbf{B}_n$. For our application, the matrix $\mathbf{B}_n$ is in a specific form, $\mathbf{B}_n = \mathbf{I}_G \otimes \mathbf{1}_I$, where $\otimes$ is the Kronecker product, $\mathbf{I}_G$ is the $G \times G$ identity matrix, and $\mathbf{1}_I$ is a vector of length I with all elements 1. Using such a structure, the answer is affirmative. In the next section we show that $\boldsymbol{\beta}$ can be estimated at the parametric rate $n^{-1/2}$ and $m(\cdot)$ can be estimated at the nonparametric rate $n^{-2/5}$. Further, we derive the asymptotic normality of $\boldsymbol{\beta}$ and describe the exact cost for estimating the nuisance parameters. Our results are interesting extensions of the partial consistent phenomenon studied by Neyman and Scott (1948). They show that when the number of nuisance parameters grow with sample size, one parametric component can be consistently estimated, but the other part cannot be. Our results are of theoretical interest in their own right, in addition to providing insights and methodological advance for the problem of microarray normalization.

The remainder of the article is organized as follows. In Section 2 we derive the profile least squares estimators for the

parameters $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}$ and the function $m(\cdot)$. We further demonstrate that the proposed estimators for $\boldsymbol{\beta}$ and $m(\cdot)$ are consistent and admit optimal rates of convergence. Section 3 extends model (1) to aggregate information from other arrays. Both simulated and real data examples are given in Section 4. Technical proofs and regularity conditions are relegated to the Appendix.

2. PROFILE LEAST SQUARES AND ASYMPTOTIC PROPERTIES

2.1 Profile Least Squares Estimator

In this section we derive the profile least squares estimators for $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}$ and the function $m(\cdot)$. For any given $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}$, (4) can be written as

$$\mathbf{Y}_n - \mathbf{B}_n\boldsymbol{\alpha}_n - \mathbf{Z}_n\boldsymbol{\beta} = \mathbf{M} + \boldsymbol{\epsilon}_n. \tag{5}$$

Thus, adopting a local linear regression technique, the estimator of $\mathbf{M}$ is

$$\widehat{\mathbf{M}} = \mathbf{S}(\mathbf{Y}_n - \mathbf{B}_n\boldsymbol{\alpha}_n - \mathbf{Z}_n\boldsymbol{\beta}),$$

where $\mathbf{S}$ is a smoothing matrix that depends only on the observations $\{X_i, i = 1, \ldots, n\}$. The explicit form of $\mathbf{S}$ is shown in the Appendix for the local linear smoother (Fan 1992). Substituting $\widehat{\mathbf{M}}$ into (5), we have

$$\widetilde{\mathbf{Y}}_n = \widetilde{\mathbf{B}}_n\boldsymbol{\alpha}_n + \widetilde{\mathbf{Z}}_n\boldsymbol{\beta} + \boldsymbol{\epsilon}_n,$$

where $\widetilde{\mathbf{Y}}_n = (\mathbf{I} - \mathbf{S})\mathbf{Y}_n$, $\widetilde{\mathbf{B}}_n = (\mathbf{I} - \mathbf{S})\mathbf{B}_n$, and $\widetilde{\mathbf{Z}}_n = (\mathbf{I} - \mathbf{S})\mathbf{Z}_n$. This is a synthetic (but not a true) linear model. By the least squares method and a slight complex computation (Rao and Toutenburg 1999), we have the following estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_n$:

$$\widehat{\boldsymbol{\beta}} = \big(\widetilde{\mathbf{Z}}_n^T\widetilde{\mathbf{Z}}_n - \widetilde{\mathbf{Z}}_n^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\widetilde{\mathbf{Z}}_n\big)^{-1}\widetilde{\mathbf{Z}}_n^T\big(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{B}}_n}\big)\widetilde{\mathbf{Y}}_n \tag{6}$$

and

$$\widehat{\boldsymbol{\alpha}}_n = (\widetilde{\mathbf{B}}_n^T\widetilde{\mathbf{B}}_n)^-\widetilde{\mathbf{B}}_n^T(\widetilde{\mathbf{Y}}_n - \widetilde{\mathbf{Z}}_n\widehat{\boldsymbol{\beta}}), \tag{7}$$

where $\mathbf{P}_{\widetilde{\mathbf{B}}_n} = \widetilde{\mathbf{B}}_n(\widetilde{\mathbf{B}}_n^T\widetilde{\mathbf{B}}_n)^-\widetilde{\mathbf{B}}_n^T$ is a projection matrix of $\widetilde{\mathbf{B}}_n$.

The foregoing profile least squares estimators can be computed by the following iterative algorithm:

Step 1. Given $\widehat{\boldsymbol{\beta}}$ and $\widehat{m}(\cdot)$ (assume their initial values to be 0's), estimate α_g by

$$\widehat{\alpha}_g = I^{-1}\sum_{i=1}^{I}\big(Y_{gi} - \widehat{\beta}_{r_{gi}} - \widehat{\gamma}_{c_{gi}} - \widehat{m}(X_{gi})\big).$$

Step 2. Given $\widehat{\boldsymbol{\alpha}}_n$ and $\widehat{m}(\cdot)$, estimate $\boldsymbol{\beta}$ by fitting the linear model

$$Y_{gi} - \widehat{\alpha}_g - \widehat{m}(X_{gi}) = \beta_{r_{gi}} + \gamma_{c_{gi}} + \varepsilon_{gi}.$$

Center the estimated coefficients of $\boldsymbol{\beta}$ and γ so that their averages are 0.

Step 3. Given $\widehat{\boldsymbol{\alpha}}_n$ and $\widehat{\boldsymbol{\beta}}$, estimate the function $m(\cdot)$ by smoothing $Y_{gi} - \widehat{\alpha}_g - \widehat{\beta}_{r_{gi}} - \widehat{\gamma}_{c_{gi}}$ on X_{gi}. Center the estimated function to have mean value 0.

Step 4. Continue Steps 1–3 until convergence.

In our implementation, it takes only a couple of iterations for the algorithm to converge. The advantage of this estimation algorithm is that it effectively separates the estimation problem into two parts, so that the nonparametric component can be estimated locally, whereas the parametric components can be estimated globally by using all of the data. The method is in contrast with the spline method of Huang et al. (2003). First, the parameters $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}$ are estimated one by one. This avoids inverting any large matrix and is very useful for the high-dimensional microarray data. Second, $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}$ can be estimated by the weighted least squares in presence of heteroscedasticity [e.g., the noise level depends smoothly on an unknown function $\sigma(X_{gi})$]. Third, the smoothing parameters in Step 3 can be selected by using an existing technique, such as the preasymptotic substitution method of Fan and Gijbels (1995) and the empirical bias method of Ruppert (1997), among others. In our implementation, we use the empirical bias method of Ruppert (1997).

2.2 Asymptotic Properties on Parametric Component

To facilitate the presentation and technical proofs, we assume that

$$\{(r_{gi}, c_{gi}, X_{gi}, \varepsilon_{gi}), i = 1, \ldots, I, g = 1, \ldots, G\}$$

are a random sample from a population. More generally, we assume that the random variables $\{(\mathbf{Z}_j, X_j, \varepsilon_j) : j = 1, \ldots, n\}$ in (4) are a random sample from a population. For model (1), the random variable $\mathbf{Z}_j$ is the indicator variable associated with the column and row effects. These assumptions are made to facilitate theoretical derivations. However, the scope of applications goes beyond these technical assumptions, as demonstrated in our simulation studies. For simplicity, assume for the moment that the model (4) is homoscedastic, namely $\operatorname{var}(\varepsilon_j|\mathbf{Z}_j, X_j) = \sigma^2$. Then we have the following asymptotic property for $\widehat{\boldsymbol{\beta}}$.

Theorem 1. Under the regularity conditions in the Appendix, the profile least squares estimator of $\boldsymbol{\beta}$ is asymptotically normal, that is,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{I}{I-1}\sigma^2\boldsymbol{\Sigma}^{-1}\right),$$

with $\boldsymbol{\Sigma} = \mathrm{E}\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}^T\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}$.

When $\boldsymbol{\alpha}_n$ is known, model (4) reduces to the partial linear model. In this case the asymptotic variance is $\boldsymbol{\Sigma}$ (see, e.g., Speckman 1988; Carroll, Fan, Gijbels, and Wand 1997), which is the semiparametric information bound. Because there are G nuisance parameters, they cost at least G data points to estimate, and the remaining degrees of freedom are $n - G = n(I-1)/I$. The factor $I/(I-1)$ is the price that we have to pay for estimating the nuisance parameters $\boldsymbol{\alpha}_n$. This price factor is the minimum that we can have. To appreciate this, let us consider a very simple model,

$$Y_{gi} = \alpha_g + \beta U_{gi} + \varepsilon_{gi},$$

where $\varepsilon_{gi} \sim \mathrm{N}(0, \sigma^2)$ and U_{gi} is distributed with $\mathrm{E}U_{gi} = 0$ and $\mathrm{E}U_{gi}^2 = 1$. The efficient information bound for estimating $\boldsymbol{\beta}$ is

$n^{-1}\sigma^2 I/(I-1)$, a factor $I/(I-1)$ larger than the case where α_g's are known.

Our theoretical result is derived under the random design of $\mathbf{Z}$; it also holds for fixed design of $\mathbf{Z}$ satisfying certain mathematical conditions. This is indeed shown via simulation studies in Examples 3 and 4 in Section 4.1.

We now consider the situation of heteroscedastic error in the model (4), that is,

$$\boldsymbol{\epsilon}_n = (\sigma(X_1)\varepsilon_1, \ldots, \sigma(X_n)\varepsilon_n)^T$$

with a continuous standard deviation function $\sigma(\cdot)$ on the support of X. Suppose that we ignore the heteroscedasticity in the estimation procedure. We then have the following theorem for the asymptotic property of $\widehat{\boldsymbol{\beta}}$.

Theorem 2. Under the regularity conditions in the Appendix, the profile least squares estimator of $\boldsymbol{\beta}$ is asymptotically normal, that is,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{I^2}{(I-1)^2}\boldsymbol{\Sigma}^{-1}\mathbf{V}\boldsymbol{\Sigma}^{-1}\right),$$

where, with $\boldsymbol{\Sigma}$ given in Theorem 1,

$$\mathbf{V} = \frac{(I-1)^2}{I^2}\mathrm{E}\sigma^2(X)\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}^T\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\} + \frac{I-1}{I^2}\mathrm{E}\sigma^2(X)\cdot\boldsymbol{\Sigma}.$$

Theorem 2 examines the impact of heteroscedasticity on the ordinary least squares estimate. The heteroscedasticity is not explicitly taken into account for the following reasons. First, even if the conditional variance function is known, after standardization, our model structure will be changed—the special structure of $\mathbf{B}_n$ in (3) does not hold any more. In our simulation studies in Example 2, not much improvement is achieved by using weighted least squares. In fact, it is not clear to us whether the weighted least squares must outperform the ordinary least squares for this special class of models. Further research is needed.

2.3 Nonparametric Part

The purpose of this article is to show that the nonparametric function $m(\cdot)$ and parametric parameter $\boldsymbol{\beta}$ can be estimated consistently and efficiently. Theorems 1 and 2 have already shown that $\boldsymbol{\beta}$ can be estimated at root-n rate, which is negligible for nonparametric estimation. To simplify technical derivations without losing the essential ingredient, we assume that $\boldsymbol{\beta}$ is known. Therefore, model (4) can be simplified as

$$\mathbf{Y}_n = \mathbf{B}_n\boldsymbol{\alpha}_n + \mathbf{M} + \boldsymbol{\epsilon}_n, \tag{8}$$

where $\mathbf{B}_n = \mathbf{I}_G \otimes \mathbf{1}_I$ and $n = G \times I$. Instead, putting identifiability on the function m, we impose the identifiability condition $\sum_{i=1}^G \alpha_i = 0$ to facilitate the technical arguments.

The profile least squares estimator can be regarded as the iterative solution to the following equation (see the algorithm in Sec. 2.1):

$$\begin{bmatrix} \mathbf{I} & \mathbf{P} \\ \mathbf{S} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{B}_n\boldsymbol{\alpha}_n \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{P} \\ \mathbf{S} \end{bmatrix}\mathbf{Y}_n,$$

where

$$\begin{aligned}\mathbf{P} &= \mathbf{B}_n(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T \\ &= (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n)\mathbf{B}_n(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T.\end{aligned}$$

According to the results of Opsomer and Ruppert (1997), the estimate of $\mathbf{M}$ has the explicit form

$$\widehat{\mathbf{M}} = \{\mathbf{I} - (\mathbf{I} - \mathbf{SP})^{-1}(\mathbf{I} - \mathbf{S})\}\mathbf{Y}_n, \tag{9}$$

provided that the inverse matrix exists. Then we have the following asymptotic property for $\widehat{\mathbf{M}}$.

Theorem 3. Under the regularity conditions in the Appendix, the risk of the profile least squares $\widehat{\mathbf{M}}$ is bounded as follows:

$$\begin{aligned}&\mathrm{MSE}\{\widehat{m}(x)|X_1, \ldots, X_n\} \\ &\quad\le \frac{I}{n(\sqrt{I}-1)^2}\sum_{i=1}^n\left\{\frac{\mu_2^2h^4}{4}\{m''(X_i)\}^2 + \frac{\sigma^2\nu_0}{nhf(X_i)}\right\} \\ &\qquad + o_p\left(h^4 + \frac{1}{nh}\right) \\ &\quad= \frac{I}{(\sqrt{I}-1)^2}\left(\frac{\mu_2^2h^4}{4}\mathrm{E}\{m''(X)\}^2 + \frac{\sigma^2\nu_0|\Omega|}{nh}\right) \\ &\qquad + o_p\left(h^4 + \frac{1}{nh}\right),\end{aligned}$$

where

$$\mathrm{MSE}\{\widehat{m}(x)|X_1, \ldots, X_n\} = \frac{1}{n}\sum_{i=1}^n \mathrm{E}\left\{[\widehat{m}(X_i) - m(X_i)]^2|X_1, \ldots, X_n\right\}$$

and Ω is the support of the random variable X.

For the situation of heteroscedastic errors, if we ignore the error structure and apply ordinary least squares, then we have the following theorem.

Theorem 4. Under the regularity conditions in the Appendix, the full-iterative backfitting estimator $\widehat{\mathbf{M}}$ has

$$\begin{aligned}&\mathrm{MSE}\{\widehat{m}(x)|X_1, \ldots, X_n\} \\ &\quad\le \frac{I}{(\sqrt{I}-1)^2}\left(\frac{\mu_2^2}{4}\mathrm{E}\{m''(X)\}^2h^4\right. \\ &\qquad \left. + \left\{\frac{(I-1)\mathrm{E}\sigma^2(X)\cdot|\Omega|}{I^2} + \frac{(I-1)^2}{I^2}\int_\Omega \sigma^2(x)\,dx\right\}\frac{\nu_0}{nh}\right) \\ &\qquad + o_p\left(h^4 + \frac{1}{nh}\right).\end{aligned}$$

It is clear from Theorems 3 and 4 that the nonparametric components achieve the optimal rate of convergence $O(n^{-2/5})$ when the bandwidth h is of order $n^{-1/5}$.

3. AGGREGATION ACROSS ARRAYS

In the preceding section, the intensity effect and the gene effect were estimated using the information within one slide. An advantage of this is that the arrays are allowed to have different gene effects, namely α_g can be slide-dependent. This occurs when samples were drawn from different subjects. In many other situations, the samples may come from the same subject. In this case it is natural to assume that the treatment effects on genes are the same across arrays, and one can aggregate the information from other arrays. This kind of aggregation idea has appeared in work of Huang and Zhang (2003) for a different semiparametric model.

Model (3) is an aggregated model, allowing interactions between blocks and intensities. Dropping the block label b, it becomes

$$Y_{gj} = \alpha_g + m(X_{gj}) + \varepsilon_{gj}, \qquad j = 1, \ldots, J, \tag{10}$$

where J is the number of arrays. This model is the same as (8), and Theorems 3 and 4 give asymptotic performance for the intensity effect (at each block) of m. In model (10), the intensity effects are the same across arrays. The results can be generalized to the array-dependent model

$$Y_{gj} = \alpha_g + m_j(X_{gj}) + \varepsilon_{gj}, \qquad j = 1, \ldots, J. \tag{11}$$

The functions m_j can be estimated at rate $O(n^{-2/5})$.

A generalization of model (1) is the semiparametric model

$$Y_{gij} = \alpha_g + \beta_{j,r_{gi}} + \gamma_{j,c_{gi}} + m_j(X_{gij}) + \varepsilon_{gij}, \tag{12}$$

$g = 1, \ldots, G$, $i = 1, \ldots, I$, and $j = 1, \ldots, J$, with J being the number of arrays, where r_{gi} and c_{gi} are the row and column of the print-tip block where the gth gene of the ith replication resides (this usually does not depend on the array) and $\boldsymbol{\beta}_j = (\beta_{j,1}, \ldots, \beta_{j,r}, \gamma_{j,1}, \ldots, \gamma_{j,c})$ and $m_j(\cdot)$ represent the block effect and intensity effect for each array j. The model (12) is not identifiable when there is no replication within an array ($I = 1$) and is identifiable when there is a replication $I > 1$. Because all arrays share the same amount of gene effect, α_g, the nuisance parameters α_g can be estimated more accurately. The question is how much better the parameters of interest, $\boldsymbol{\beta}_j$ and $m_j(\cdot)$, can be estimated by using the aggregation.

The algorithm in Section 2.1 can be modified as follows:

Step 1. Given $\widehat{\boldsymbol{\beta}}_j$ and $\widehat{m}_j(\cdot)$ (assume their initial values to be 0's), estimate α_g by averaging over

$$Y^*_{gij} = Y_{gij} - \widehat{\beta}_{j,r_{gi}} - \widehat{\gamma}_{j,c_{gi}} - \widehat{m}_j(X_{gij})$$

with respect to i and j.

Step 2. Given $\widehat{\boldsymbol{\alpha}}_n$ and $\widehat{m}_j(\cdot)$, estimate $\boldsymbol{\beta}_j$ using the linear model

$$Y^*_{gij} = Y_{gij} - \widehat{\alpha}_g - \widehat{m}(X_{gij}) = \beta_{j,r_{gi}} + \gamma_{j,c_{gi}} + \varepsilon_{gij}.$$

This is done for each separate j. Center-estimate $\boldsymbol{\beta}$'s and γ's to have mean 0.

Step 3. Given $\widehat{\boldsymbol{\alpha}}_n$ and $\widehat{\boldsymbol{\beta}}_j$, estimate function $m_j(\cdot)$ with

$$Y^*_{gij} = Y_{gij} - \widehat{\beta}_{j,r_{gi}} - \widehat{\gamma}_{j,c_{gi}} = m_j(X_{gij}) + \varepsilon_{gij}.$$

Again, this is done for each separate j. Center $\widehat{m}_j$ so that its average is 0.

Step 4. Continue Steps 1–3 until convergence.

Assume that the data from each array are iid samples with homoscedastic error. Then the problem is similar to (1), but the cost for estimating nuisance parameters is shared by J arrays. One key difference is that the rows r_{gi} and columns c_{gi} do not depend on the array j. This makes the problem under study more difficult. Put model (12) into the matrix form as

$$\mathbf{Y} = \mathbf{B}\alpha + \mathbf{Z}^*\boldsymbol{\beta} + \mathbf{M} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_J^T)^T$, $\mathbf{Y}_j = (Y_{11j}, Y_{12j}, \ldots, Y_{GIj})^T$, $\mathbf{B} = (\mathbf{B}_1, \ldots, \mathbf{B}_J)^T$, $\mathbf{B}_j = \mathbf{I}_G^T \otimes \mathbf{1}^T$, $\mathbf{Z}^* = \operatorname{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_J)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T)^T$, $\mathbf{M} = (\mathbf{M}_1^T, \ldots, \mathbf{M}_J^T)^T$, and $\mathbf{M}_j = \{m_j(X_{gij}), g = 1, 2, \ldots, G, i = 1, \ldots, I\}^T$, with $\mathbf{Z}_1 = \mathbf{Z}_2 = \cdots = \mathbf{Z}_J = \mathbf{Z}$ as the gene position in each array remains the same. Assume that the conditional distribution of $\mathbf{Z}$ given (X_i, X_j) is the same for all $i \neq j$. Following a proof similar to that of Theorem 1, we have the following asymptotic normality for the aggregated estimator. The asymptotic is based on the assumption that $G \to \infty$.

Theorem 5. Under the regularity conditions in the Appendix, the profile least squares estimator of $\boldsymbol{\beta}_j$ is asymptotically normal, that is,

$$\sqrt{GI}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_j),$$

where $\boldsymbol{\Sigma}^* = \mathrm{E}\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X_1)\}^T\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X_2)\}$, $\boldsymbol{\Sigma} = \mathrm{E}\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}^T\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}$ and

$$\begin{aligned}\boldsymbol{\Sigma}_j = {} & \left(\frac{IJ-1}{IJ}\boldsymbol{\Sigma} + \frac{1}{IJ}\boldsymbol{\Sigma}^*\right)^{-1} \\ & + \frac{1}{J}\left(I \cdot \boldsymbol{\Sigma}^{*-1} \cdot \left(\frac{IJ-1}{IJ}\boldsymbol{\Sigma} + \frac{1}{IJ}\boldsymbol{\Sigma}^*\right) - \mathbf{I}\right)^{-1} \\ & \times \left(\frac{IJ-1}{IJ}\boldsymbol{\Sigma} + \frac{1}{IJ}\boldsymbol{\Sigma}^*\right)^{-1}\end{aligned}$$

with X, X_1, and X_2 having the same distribution as that of X_{gij}.

Despite its complicated asymptotic expression, Theorem 5 shows the extent to which the cost of estimating nuisance parameter α_g can be reduced. To simplify our results, we consider the fact that Z_{gij} is independent of X_{gij}. Then $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^*$ and

$$\boldsymbol{\Sigma}_j = \frac{J(I-1)+1}{J(I-1)}\boldsymbol{\Sigma}^{-1}. \tag{13}$$

This shows that model is not identifiable when $I = 1$. Because the sample size for estimating β_j is GI when the α_g's are known, the cost for estimating the α_g's for each array is about

$$GI - \frac{GIJ(I-1)}{J(I-1)+1} = \frac{GI}{J(I-1)+1}$$

data points. For example, if $I = 2$, $J = 6$, and $G = 111$ as in the illustrative example in Section 4.2, then the loss of degrees of freedom due to estimating nuisance parameters decreases from 111 to 31.71. However, the efficiency of the intensity effect m_j cannot be improved very much, because it is estimated separately from each slide. Indeed, from an asymptotic standpoint, $\boldsymbol{\beta}$ can be treated as if they were known for estimating $m_j(\cdot)$, whether or not the information from other arrays are aggregated. Because the errors in estimating $m_j(\cdot)$ dominate

those in estimating the block effects, the accuracy for normalization (2) cannot be improved very much by aggregating information from other arrays. This gives theoretical support to the arraywise normalization method of Fan et al. (2004). That method has an additional advantages of computational expedience and robustness to the assumption that the gene effects remain the same across arrays.

Aggregation does give us one important advantage: It reduces the cost of estimation nuisance parameters per array. With aggregated sample size IJG, which amounts to 1,332 for our illustrative example, we may be willing to relax the additive column and row effect. Namely, we may extend model (12) to the more flexible model

$$Y_{gij} = \alpha_g + \delta_{j,b_{gi}} + m_j(X_{gij}) + \varepsilon_{gij}, \tag{14}$$

where $\{b_{gi}\}$ (ranging from 1 to 32 in our application) is the block where the gene g with repetition i resides and $\{\delta_{j,k}\}$ measures the block effect of the jth array, consisting of $J(cr-1)$ parameters. This can also be considered as a specific mathematical model of (12) by thinking $\beta = \delta$ and $\gamma = 0$. Thus Theorem 5 continues to apply, and their estimation errors are negligible in comparison with those in estimation $m_j(\cdot)$. Model (12) is a specific case of (14) with additive row and column effects. Thus model (14) reduces possible modeling biases. The normalization (2) now becomes

$$Y_g^* = Y_g - \widehat{\delta}_{b_g} - \widehat{m}(X_g), \tag{15}$$

for each slide.

The foregoing results are based on the assumption that $\mathbf{Z}_1 = \cdots = \mathbf{Z}_J$ for microarray applications. In contrast, it is possible to design the case that $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_J$ are independent. This amounts to using model (12) with c_{rgj} and r_{rgj}. The following theorem gives the result on this specific case.

Theorem 6. Under the regularity conditions in the Appendix, the profile least squares estimator of $\boldsymbol{\beta}_j$ is asymptotically normal, that is,

$$\sqrt{G(IJ-1)/J}(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_j^{-1}),$$

where $\boldsymbol{\Sigma}_j = \mathrm{E}\{\mathbf{Z}_j - \mathrm{E}(\mathbf{Z}_j|X_j)\}^T\{\mathbf{Z}_j - \mathrm{E}(\mathbf{Z}_j|X_j)\}$.

This specific model is of theoretical interest, with possible applications to other statistical problems. For this specific model, $\boldsymbol{\Sigma}^* = \mathbf{0}$ in Theorem 5, and Theorem 6 can be deduced from Theorem 5. With this specific design, the cost for estimating the α_g's for each reduces array further to G/J data points. For example, if $I = 2$, $J = 6$, and $G = 111$, then the loss of degrees of freedom due to estimating nuisance parameters decreases from 111 to 18.5. This compares with 31.7 data point with fixed design $\mathbf{Z}_1 = \cdots = \mathbf{Z}_J$ mentioned earlier.

4. SIMULATION AND APPLICATION

In this section we use several simulated examples to augment the partial consistent phenomenon demonstrated in the last two sections. We conclude this section with an application of the proposed method to the normalization of the microarray data arising from the study of neuroblastoma cells in response to the stimulation of the macrophage migration inhibitory factor (MIF), a growth factor.

4.1 Simulations

Our theoretical results are illustrated empirically by four examples. The first two examples study the situation under which the genes with replications are randomly placed on arrays. One of these is a homoscedastic model, and the other is a heteroscedastic model. The last two examples show that our results continue to apply to fixed designs. Because our models and the partial consistent results are motivated by the analysis of microarray data, the validity of the randomness assumption arises for replicated genes. We use Example 3 to demonstrate that our methods continue to work for the replications similar to our illustrative example in Section 4.2. Finally, in Example 4 we apply our method to the case in which no gene has replications within an array and the design of genes is fixed. In all examples, we assume that there are 32 print-tip blocks with 4 rows and 8 columns. The performance of $\widehat{\boldsymbol{\alpha}}_n$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{m}(\cdot)$ is assessed by the mean squared errors (MSEs),

$$\mathrm{MSE}(\widehat{\boldsymbol{\alpha}}_n) = \frac{1}{G}\sum_{g=1}^{G}(\widehat{\alpha}_g - \alpha_g)^2,$$

$$\mathrm{MSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{r+c}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2,$$

and

$$\mathrm{MSE}(\widehat{m}) = \frac{1}{GI}\sum_{g=1}^{G}\sum_{i=1}^{I}\{\widehat{m}(X_{gi}) - m(X_{gi})\}^2.$$

The MSEs are examined by varying G and I. To examine the impact of heteroscedasticity on the efficiency of parameters, we also consider the weighted least squares method, in which Steps 1 and 2 are implemented by using the weighted least squares with the conditional variance function estimated by smoothing the squared residuals on X_{gi} (see Fan and Yao 1998).

Example 1. In this example we choose $G = 100, 200, 400, 800$ and $I = 2, 3, 4$. For each pair of (G, I), we simulate $N = 200$ datasets from model (1). To examine how much the aggregated method in Section 3 can improve estimation of the intensity effect $m(\cdot)$ and print-tip block effect ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$), we simulate data from model (12) with the same print-tip block effect and intensity effect as in model (1). We assume that there are $J = 4$ arrays available for us to aggregate the information. We repeat the simulation $N = 50$ times, each time consisting of $J = 4$ arrays for aggregation. The details of simulation scheme for this example are as follows:

- $\boldsymbol{\alpha}_n$. The expression levels of the genes are generated from the standard double-exponential distribution.
- $\boldsymbol{\beta}$. For the row effects, first generate $\{\beta_i', i = 1, \ldots, 4\}$ from N(0, .5), then set $\beta_i = \beta_i' - \bar{\beta}'$, which will guarantee that $\sum_{i=1}^4 \beta_i = 0$. The column effects are generated in the same way.
- X. The intensity is generated from a mixture distribution. We generate x from probability distribution $.0004 \times (x-6)^3\mathrm{I}(6 < x < 16)$ with probability .7 and from uniform distribution over [6, 16] with probability .3.

Table 1. MSEs of Example 1, Ordinary Least Squares and Weighted Least Squares Estimations, $n = 200$

		Ordinary least squares				Weighted least squares			
	I	$G = 100$	$G = 200$	$G = 400$	$G = 800$	$G = 100$	$G = 200$	$G = 400$	$G = 800$
m	2	.1454	.0752	.0358	.0201	.1890	.0882	.0422	.0242
	3	.0780	.0397	.0234	.0137	.1112	.0505	.0275	.0162
	4	.0515	.0273	.0167	.0100	.0739	.0339	.0204	.0117
β	2	.0668	.0299	.0151	.0069	.0691	.0306	.0151	.0069
	3	.0318	.0148	.0071	.0033	.0330	.0148	.0072	.0033
	4	.0211	.0098	.0050	.0024	.0216	.0099	.0050	.0024
α	2	.6428	.5690	.5290	.5203	.7041	.5919	.5389	.5259
	3	.3930	.3607	.3520	.3411	.4383	.3765	.3582	.3443
	4	.2788	.2676	.2630	.2573	.3103	.2784	.2678	.2596

$m(\cdot)$. Set the function $m(X) = \sqrt{5}(\sin X - .2854)$, whose expectation is 0.

Z. For each given gene, its associated block is assigned at random at one of 32 print-tip blocks.

ε. ε_{gi} is generated from the standard normal distribution.

This is a homoscedastic model. The estimation procedure described in Section 2.1 is used to estimate $\boldsymbol{\alpha}_n$, $\boldsymbol{\beta}$, and $m(\cdot)$. Table 1 presents the MSEs of the ordinary least squares and weighted least squares estimators for $\widehat{\boldsymbol{\alpha}}_n$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{m}(\cdot)$. The table shows that when the number of replications I is fixed, the MSEs of $m(\cdot)$ and $\boldsymbol{\beta}$ decrease as the number of genes increases, which indicates the consistency of the estimators of $m(\cdot)$ and $\boldsymbol{\beta}$. However, the MSEs of $\boldsymbol{\alpha}_n$ are very stable as the number of genes increases, which demonstrates the inconsistency of the estimator for $\boldsymbol{\alpha}_n$. Furthermore, to visualize the MSEs and the convergence rates, these MSEs are also depicted in Figure 1, where $\log(\text{MSE}) - \log(I/(I-1))$ is plotted against $\log(n)$. Note that $n = IG$ and that the factor $I/(I-1)$ is used to correct the intercept term (see Thm. 1). The figure shows that $\boldsymbol{\beta}$ has the parametric rate $n^{-1/2}$ and $m(\cdot)$ has the nonparametric rate $n^{-2/5}$. Table 2 presents the MSEs for $\widehat{\boldsymbol{\alpha}}_n$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{m}(\cdot)$, which are also

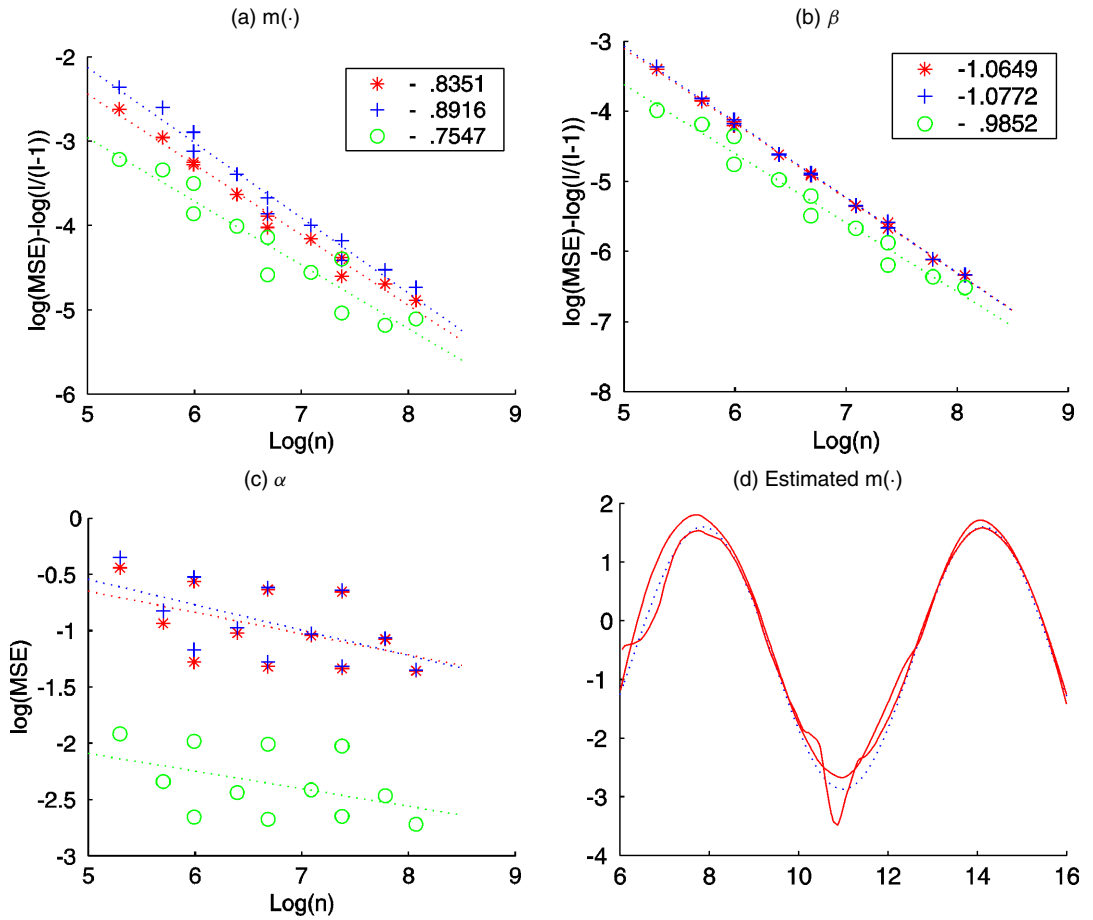

Figure 1. Example 1. (a)–(c) Plots of MSEs of $m(\cdot)$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$: Ordinary least squares ($*$), weighted least squares ($+$), aggregated method ($\circ$). The dotted lines are the regression lines for the MSEs of the three different estimators. The slopes are shown for $m(\cdot)$ and $\boldsymbol{\beta}$. (d) The performance of $\widehat{m}(\cdot)$ when $G = 400$ and $I = 2$. The dotted line is the true function of $m(\cdot)$, and the solid lines are two estimated functions.

Table 2. MSEs of Example 1, Aggregation Across Arrays Estimation, n = 50, J = 4

	I	G = 100	G = 200	G = 400	G = 800
m	2	.0761	.0413	.0216	.0149
	3	.0502	.0269	.0166	.0116
	4	.0373	.0188	.0146	.0093
$\boldsymbol{\beta}$	2	.0433	.0193	.0086	.0043
	3	.0246	.0109	.0054	.0027
	4	.0160	.0077	.0041	.0019
$\boldsymbol{\alpha}$	2	.1553	.1420	.1370	.1346
	3	.0980	.0885	.0906	.0881
	4	.0688	.0668	.0689	.0666

plotted in Figure 1, when the information from four arrays is aggregated. Clearly, $\boldsymbol{\alpha}_n$ is estimated more accurately by aggregating information across arrays with the assumption that the gene effects remain the same across arrays. In contrast, the improvements on the estimation of $\boldsymbol{\beta}$ and $m(\cdot)$ are relatively smaller. According to Theorems 1 and 5, the asymptotic relative efficiency for estimating $\boldsymbol{\beta}$ is $(4I-4)/(4I-1)$. The relative efficiencies are in line with those given in Tables 1 and 2.

To demonstrate the effectiveness of estimated functions and parameters, Figure 1 also shows two randomly selected estimates $\widehat{m}(\cdot)$ for $G = 400$ and $I = 2$. Figure 2 summarizes the boxplots of row $\{\widehat{\beta}_k\}$ and column $\{\widehat{\gamma}_k\}$ effects based on 200 simulations with $G = 400$ and $I = 2$. The biases of these estimates are clearly negligible, and coefficients are estimated with similar accuracy. This is consistent with our design of simulations. For simplicity, we present only the results based on the ordinary least squares method.

Example 2. In this example we also choose $G = 100, 200, 400, 800$ and $I = 2, 3, 4$. For each pair of (G, I), we simulate $n = 200$ datasets from the model (1). To examine the effectiveness of using information across arrays, we simulate $n = 50$ datasets from the model (12), each consisting of $J = 4$ arrays with parameters taken from model (1). The parameters in this example are taken to mimic the real data in the next section. The details of simulation scheme are as follows:

$\boldsymbol{\alpha}_n$. The expression levels of the first 50 genes follow standard double exponential, and the rest are 0's.

$\boldsymbol{\beta}$. The row and column effects are fixed. Set

$$\boldsymbol{\beta}_r = (.2, .15, -.2, -.15)'$$

and

$$\boldsymbol{\beta}_c = (.15, .125, .1, .075, -.15, -.125, -.10, -.075)'.$$

X. Same as in Example 1.

$m(\cdot)$. Set the function $m(X) = 10(.2708 - \sqrt{(16-X)/32})$, whose expectation is 0.

Z. Same as in Example 1.

ε. ε_{gi} is generated from the normal distribution with mean 0 and variance $\sigma^2(X_{gi}) = .15 + .015(12 - X_{gi})^2 \times \mathrm{I}\{X_{gi} < 12\}$.

This heteroscedastic model contains many features similar to those in the real microarray data. Tables 3 and 4 give similar results to those of Tables 1 and 2. They demonstrate that the estimation of $m(\cdot)$ and $\boldsymbol{\beta}$ is consistent, but estimation of $\boldsymbol{\alpha}_n$ is not. Also note that for both homoscedastic and heteroscedastic models, ordinary least squares and weighted least squares yield the similar results. This is somewhat surprising. But we speculate that the degree of heteroscedasticity is not sufficiently large for the ordinary least squares and the weighted least squares to perform analogously. Further, as pointed out after Theorem 2, it is not clear that weighted least squares must outperform ordinary least squares under the current model. The results in Figure 3 are similar to those of Figure 1. Here we present the results only for weighted least squares estimators [in (d)]. Indubitably, the aggregation dramatically improves the estimate of $\boldsymbol{\alpha}_n$. It also improves the accuracy of the estimated intensity effect and print-tip block effect.

Example 3. In this example we assess the impact of randomness assumption to our proposed method by fixing the replicated pairs throughout simulation. We simulate $n = 200$ datasets from model (1) for the pair $(G = 111, I = 2)$. The simulation scheme is the same as in Example 2, except the manner in which these 222 genes are placed onto microarrays. To be closer to reality, we mimic the real data in the next section and fix the print-tip block positions of these 222 genes throughout simulations. The locations of the repeated pairs are identical to those in the real data. Table 5 compares the MSEs of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $m(\cdot)$ under both random and fixed designs. They are comparable, which in turn indicates that the asymptotic results continue to hold for fixed designs (in which the randomness assumption is violated).

Example 4. This example examines the effectiveness of normalization when there are no replicated genes within an array. Data are simulated from model (11) with $\boldsymbol{\alpha}$, X, $m(\cdot)$, and ε generated in the same manner as those in Example 2. Model (11) is fitted for different number of genes G within a block of an array and for different number of available arrays J for aggregation. Table 6 presents the results.

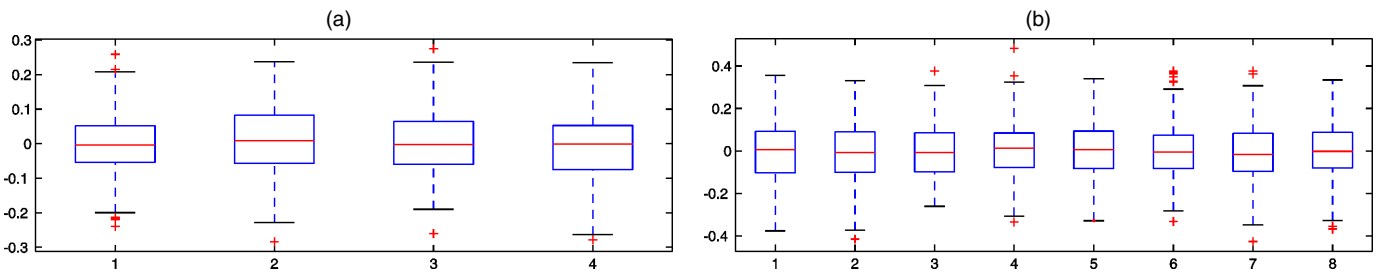

Figure 2. Boxplots of the Estimated (a) Row $\{\widehat{\beta}_k\}$ and (b) Column $\{\widehat{\gamma}_k\}$ Effects When G = 400 and I = 2 for Example 1.

Table 3. MSEs of Example 2, Ordinary Least Square and Weighted Least Square Estimation, $n = 200$

		Ordinary least squares				Weighted least squares			
	I	$G = 100$	$G = 200$	$G = 400$	$G = 800$	$G = 100$	$G = 200$	$G = 400$	$G = 800$
m	2	.0398	.0214	.0099	.0047	.0402	.0216	.0101	.0047
	3	.0245	.0105	.0058	.0027	.0289	.0107	.0078	.0027
	4	.0160	.0073	.0042	.0019	.0171	.0074	.0044	.0020
β	2	.0129	.0061	.0026	.0014	.0127	.0060	.0025	.0013
	3	.0058	.0027	.0013	.0007	.0055	.0026	.0012	.0006
	4	.0038	.0018	.0009	.0004	.0035	.0017	.0008	.0004
α	2	.1261	.1085	.1004	.0960	.1253	.1059	.0960	.0908
	3	.0766	.0677	.0655	.0637	.0783	.0643	.0627	.0584
	4	.0551	.0503	.0487	.0474	.0531	.0469	.0447	.0432

Table 4. MSEs of Example 2, Aggregation Across Arrays Estimation, $n = 50$, $J = 4$

	I	$G = 100$	$G = 200$	$G = 400$	$G = 800$
m	2	.0219	.0102	.0049	.0027
	3	.0134	.0068	.0035	.0018
	4	.0106	.0053	.0025	.0014
β	2	.0073	.0034	.0017	.0008
	3	.0041	.0020	.0010	.0005
	4	.0029	.0015	.0008	.0004
α	2	.0302	.0256	.0247	.0241
	3	.0178	.0170	.0163	.0161
	4	.0142	.0128	.0122	.0118

It is clear when the number of available arrays increases, the MSEs for the m's (averaging MSE over J arrays) decrease, because the average cost per array for estimating the $\boldsymbol{\alpha}$'s decreases. When G increases, the MSEs for m decrease dramatically, demonstrating that m is a consistent estimate. This is not true for the MSEs of the $\boldsymbol{\alpha}$'s, because $\boldsymbol{\alpha}$ cannot be estimated consistently. These are consistent with our theoretical results.

4.2 Application

The dataset used here was collected and analyzed by Fan et al. (2004). The biological aim of the study is to understand how genes are affected by the macrophage MIF in neuroblas-

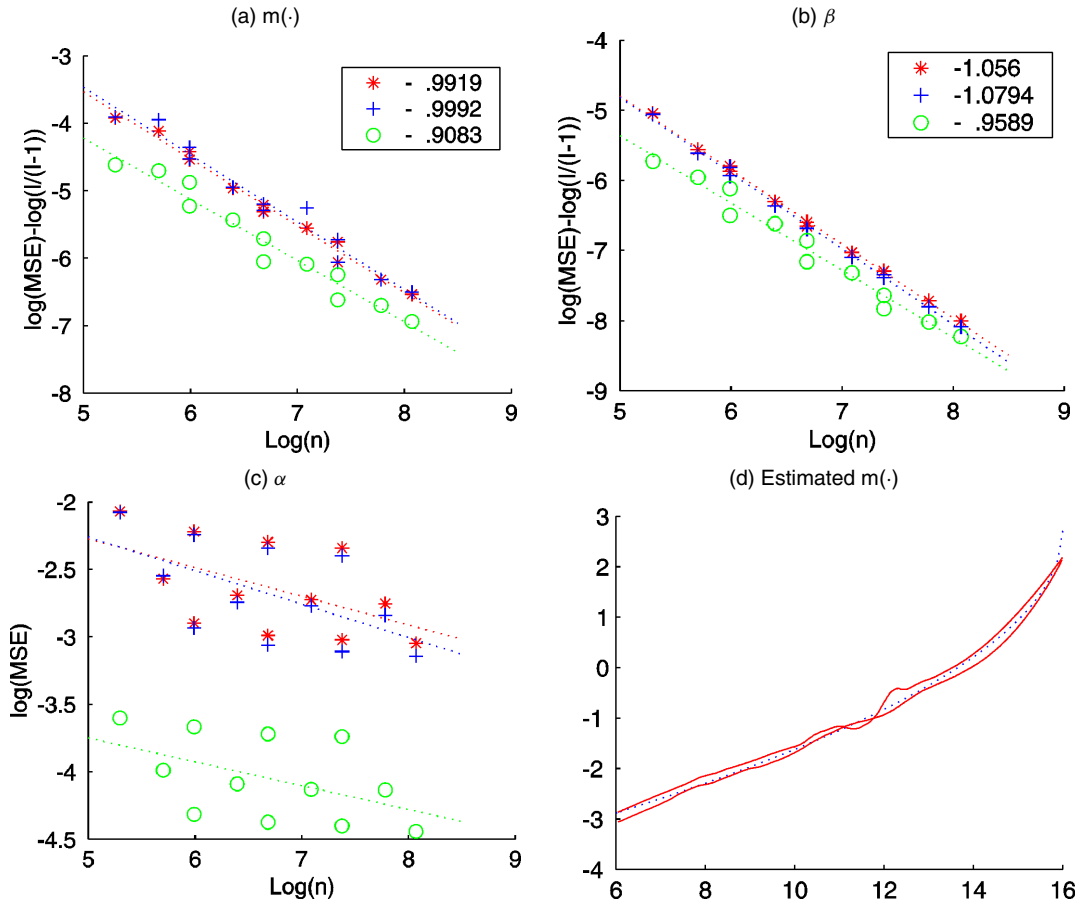

Figure 3. Example 2. (a)–(c) Plots of MSEs of $m(\cdot)$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$: Ordinary least-squares ($*$), weighted least-squares ($+$), aggregated method ($\circ$). The dotted lines are the regression lines for the MSEs of the three different estimators. The slopes are shown for $m(\cdot)$ and $\boldsymbol{\beta}$. (d) The performance of $\widehat{m}(\cdot)$ when $G = 400$ and $I = 2$. The dotted line is the true function of $m(\cdot)$ and the solid lines are two estimated functions.

Table 5. MSEs of Example 3, $G = 111$, $I = 2$, $n = 200$

	Ordinary least squares		Weighted least squares	
	Fixed design	Random design	Fixed design	Random design
m	.0368	.0373	.0382	.0395
β	.0114	.0116	.0110	.0114
α	.1253	.1212	.1237	.1216

Table 6. MSEs of Example 4, $n = 100$

	G	$J = 2$	$J = 4$	$J = 6$	$J = 8$
m	500	.0142	.0092	.0084	.0080
	1,000	.0065	.0043	.0040	.0038
α	500	.0973	.0483	.0323	.0242
	1,000	.0957	.0474	.0317	.0237

toma cells. Neuroblastoma is the second most common pediatric solid cancer and accounts for approximately 15% cancer deaths. MIF plays a central role in control of the host inflammatory and immune response and is linked to fundamental processes that control cell proliferation, cell survival, and tumor progression. It is overexpressed in several human cancers. To gain better understanding of the role of MIF in the development of neuroblastoma, the global gene expression of the neuroblastoma cell line stimulated with MIF is compared with that of those without MIF stimulation using cDNA microarrays. The details of the design and experiments were given by Fan et al. (2004).

The cDNA microarrays used here consist of 19,968 clones of sequence-validated human genes, printed on 8×4 print-tip blocks. Among these, 111 cDNA clones of genes were printed twice on each array. Figure 4(a) shows schematically the 32 print-tip blocks, with a point in the block indicating one of these 111 genes (the dot represents one clone, and the triangle represents its replication). These 222 clones are nearly uniformly distributed on the 32 blocks [see Figs. 4(b) and 4(c)]. Figure 4(d) shows the distribution, among 111 pairs of repeated genes, of the distance between two repeated clones. For example, if one gene is located in block $(3, 2)$ and the other is located in block $(5, 3)$, then its distance is $\sqrt{4+1}$.

In the notation that we have introduced, $c = 8$, $r = 4$, $G = 111$, and $I = 2$. Figures 5(a) and 5(b) shows the values of log ratios and log intensities for the repeated genes for a given array. Because the clones are identical, the differences in expression provide valuable information for estimating the systematic biases. The differences come from the location of the clones and the random errors. Following the work of Tseng et al. (2001), Dudoit et al. (2002), and Huang et al. (2003), we need to remove the systematic biases before carrying out any statistical analysis.

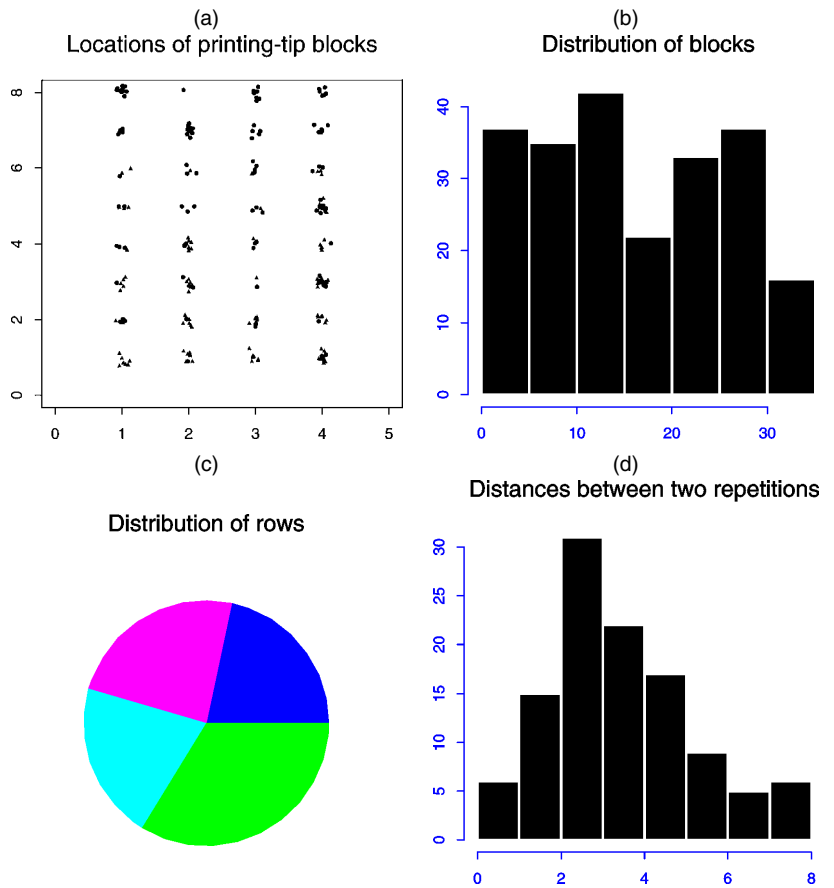

Figure 4. cDNA Microarrays. (a) Schematic representation of the locations with replications; a dot represents one clone, and a triangle represents its replication. (b) The distribution of the blocks where genes with repetitions reside. (c) Pie chart for the rows where genes with repetitions reside. (d) Histogram for the distances of the blocks for two repeated clones.

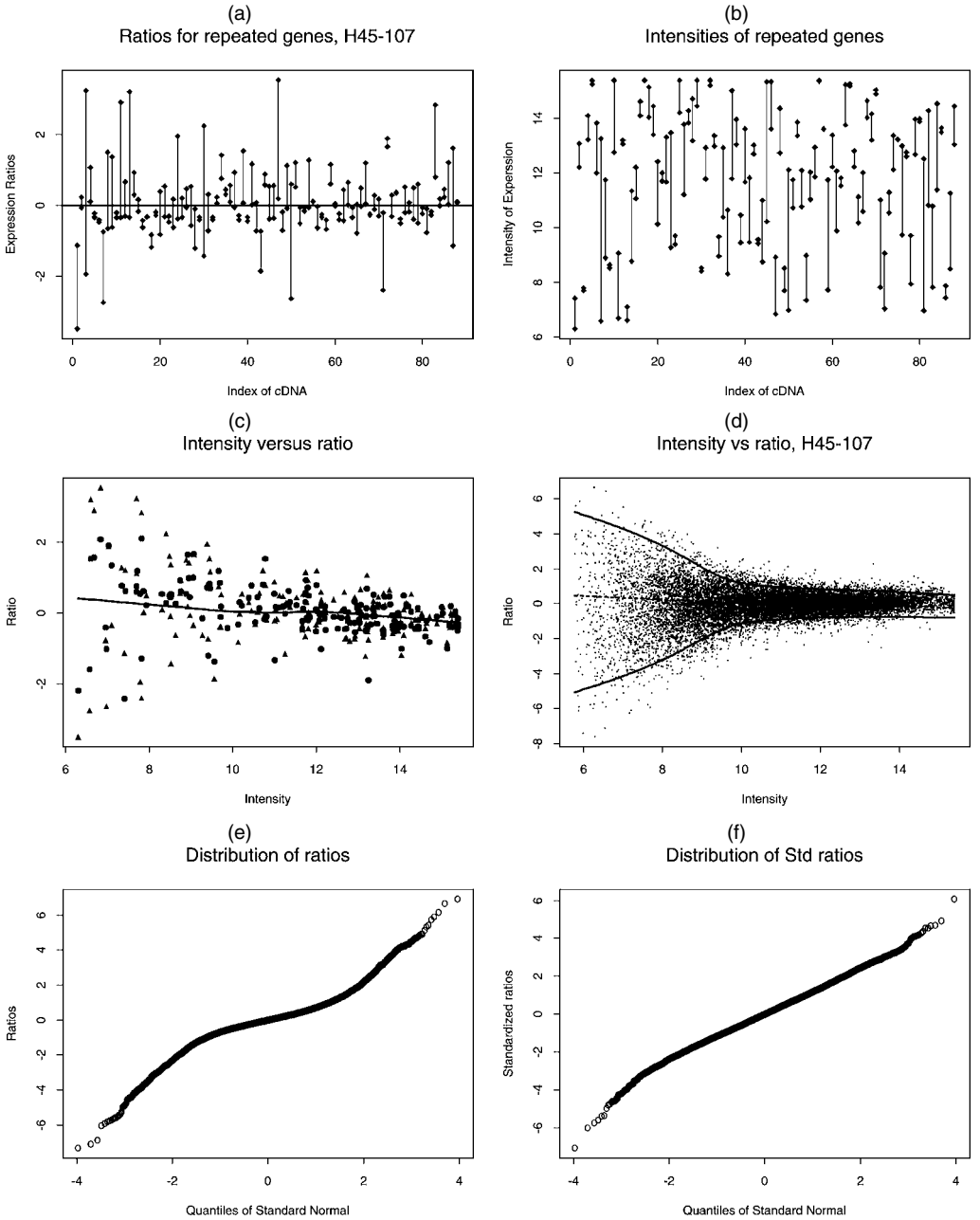

Figure 5. Repeated Genes. (a) and (b) Observed log ratios and log intensities for pairs of repeated clones. (c) Fitted function m along with normalized (squares) and unnormalized (triangles) ratios for a given array plotted against the log intensity. (d) Normalized log ratios (vs. their log intensities), with intensity and print-tip effect removed for the given array. [Thick curves are the standard deviation curves multiplied by 2, and the dashed curve is the estimated function from (c).] (e) and (f) The Q–Q plot for log ratios and standardized ratios.

We applied the model (1) to estimate the print-tip block and intensity effects for each array. For the array given in Figures 5(a) and 5(b), the estimated function m is depicted in Figure 5(c), in which unnormalized and normalized log ratios are plotted against their associated log intensities. The estimated values of β and γ are not reported here. The estimates were obtained by the ordinary least squares method. With the estimated block and intensity effects, the systematic biases in 19,968 genes can be removed via (2). The normalized results are presented in Figure 5(d). It is clear that the low intensity is associated with high variation in log ratios. The degree of heteroscadesticity (i.e., the conditional standard deviation) is estimated by using the method of Fan and Yao (1998). The estimated standard deviation function, $\hat{\sigma}(X_g)$, is also plotted in Figure 5(d). Let $\hat{\mu}(\cdot)$ be the regression function of the normalized log ratios $\{Y_g^*\}$ on its associated log intensities $\{X_g\}$. The quantile–quantile (Q–Q plots for checking normality for the normalized log ratios $\{Y_g^*\}$ and the standardized log ratios $\{(Y_g^* - \hat{\mu}(X_g))/\hat{\sigma}(X_g)\}$ are depicted in Figures 5(e) and 5(f). As shown in Figure 5(f), after standardization, the data become more normally distributed, indicating that the degree of heteroscedasticity has been properly assessed. Further analysis of this dataset was done by Fan et al. (2004).

5. DISCUSSION

Motivated by the problem of normalizing cDNA microarray data, two semiparametric models are proposed. An interesting feature of the models is that the number of nuisance parameters is proportional to the sample size. These nuisance parameters cannot be estimated consistently. However, the parameters of main interest for the normalization problem can be estimated consistently. The cost for estimating the nuisance parameters is pinned down; each nuisance parameter costs us basically one data point. This is the minimum price that we have to pay, as we have demonstrated.

Our proposed model has a wide spectrum of applications. In addition to be applicable to various situations with within-array replications, it can even be applied to the case without within-array replications. It can also be used at a block level, and this avoids the additive assumption on the intensity and block effects.

Aggregation is a powerful approach to improving the accuracy of estimated parameters. As we have demonstrated, it reduces the price for estimating one nuisance parameter per data point to per $1/J$ data point per array. As a result, the block effects are estimated more accurately. In the normalization process, the main source of estimation errors comes from the nonparametric component—the intensity effect. Thus, as far as the asymptotic is concerned, aggregation helps only partially in the accuracy of normalization. However, as we have shown in the simulation, the aggregation does help in estimating both block and intensity effects for finite samples.

Aggregation gives us much more data points for removing systematic biases. It allows us to impose some more flexible models to access the block and intensity effects. Because of increased sample size, we have more flexibility in proposing different kind of semiparametric models for normalization and analysis of data.

Within-array replications are powerful for removing systematic biases and level of measurement errors. It is not difficult to print several hundreds of repeated clones in a cDNA chip. Thus our requirement that G be large should be easy to fulfill. With increased sample sizes and our analysis technique, the systematic biases due to the experimental variations can be better removed.

APPENDIX: PROOFS

The following technical conditions are imposed. They are not weakest possible conditions, but they are imposed to facilitate the technical proofs:

1. The function $m(\cdot)$ has a bounded second derivative.
2. $\mathbf{\Sigma}$ is nonsingular, and $\mathrm{E}(\mathbf{Z}|X=x)$ is Lipschitz continuous in x.
3. Each component of $\mathbf{Z}$ is bounded.
4. The random variable X has a bounded support Ω. Its density function $f(\cdot)$ is Lipschitz continuous and bounded away from 0 on Ω.
5. The function $K(\cdot)$ is a symmetric density function with compact support.
6. $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$.

The following notation is used in the proofs of the lemmas and theorems. Let $\mu_i = \int u^i K(u)\,du$, $\nu_i = \int u^i K^2(u)\,du$, and $c_n = \{\log(1/h)/(nh)\}^{1/2} + h^2$. Also let

$$\mathbf{D}_x = \begin{pmatrix} 1 & \frac{X_1 - X}{h} \\ \vdots & \vdots \\ 1 & \frac{X_n - X}{h} \end{pmatrix} \quad \text{and}$$

$$\mathbf{S} = \begin{pmatrix} [1\ \ 0]\{\mathbf{D}_x^T \mathbf{W}_x \mathbf{D}_x\}^{-1} \mathbf{D}_{x_1}^T \mathbf{W}_{x_1} \\ \vdots \\ [1\ \ 0]\{\mathbf{D}_{x_n}^T \mathbf{W}_{x_n} \mathbf{D}_{x_n}\}^{-1} \mathbf{D}_{x_n}^T \mathbf{W}_{x_n} \end{pmatrix},$$

where $\mathbf{W}_x = \mathrm{diag}\{K_h(X_1 - X), \ldots, K_h(X_n - X)\}$, $K(\cdot)$ is a kernel function, h is a bandwidth, and $K_h(\cdot) = K(\cdot/h)/h$. Set $\Phi(X) = \mathrm{E}(\mathbf{Z}^T|X)$.

Lemma A.1. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid random vectors, where the Y_i's are scalar random variables. Further assume that $\mathrm{E}|Y_i|^4 < \infty$ and $\sup_x \int |y|^4 f(x, y)\,dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support that satisfies Lipschitz's condition. Then, under condition 6,

$$\sup_X \left| \frac{1}{n} \sum_{i=1}^{n} \left[K_h(X_i - X) Y_i - \mathrm{E}\{K_h(X_i - X) Y_i\} \right] \right| = O_p\left(\left\{ \frac{\log(1/h)}{nh} \right\}^{1/2} \right).$$

Proof. This follows immediately from the result of Mack and Silverman (1982).

Lemma A.2. Under conditions 1–6, we have

$$n^{-1} \widetilde{\mathbf{Z}}_n^T (\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{B}}}) \widetilde{\mathbf{Z}}_n \xrightarrow{P} \frac{I-1}{I} \mathbf{\Sigma},$$

where $\mathbf{\Sigma} = \mathrm{E}\{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}^T \{\mathbf{Z} - \mathrm{E}(\mathbf{Z}|X)\}$.

Proof. Note that

$$\mathbf{D}_x^T \mathbf{W}_x \mathbf{D}_x = \begin{pmatrix} \sum_{i=1}^n K_h(X_i - X) & \sum_{i=1}^n \frac{X_i - X}{h} K_h(X_i - X) \\ \sum_{i=1}^n \frac{X_i - X}{h} K_h(X_i - X) & \sum_{i=1}^n (\frac{X_i - X}{h})^2 K_h(X_i - X) \end{pmatrix}.$$

Each element of the foregoing matrix is in the form of a kernel regression. By Lemma A.1,

$$n^{-1} \mathbf{D}_x^T \mathbf{W}_x \mathbf{D}_x = f(X) \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} + O_p(c_n). \tag{A.1}$$

Using the same argument, we have

$$n^{-1} \mathbf{D}_x^T \mathbf{W}_x \mathbf{Z}_n = f(X) \Phi(X) (1, 0)^T + O_p(c_n).$$

Combining the last two results yields that, uniformly in $X \in \Omega$,

$$[1, 0]\{\mathbf{D}_x^T \mathbf{W}_x \mathbf{D}_x\}^{-1} \mathbf{D}_x^T \mathbf{W}_x \mathbf{Z}_n = \Phi(X) + O_p(c_n).$$

Then we have

$$\mathbf{S}\mathbf{Z}_n = \begin{pmatrix} \Phi(X_1) \\ \vdots \\ \Phi(X_n) \end{pmatrix} + O_p(c_n)$$

and

$$\widetilde{\mathbf{Z}}_n = \begin{pmatrix} \mathbf{Z}_1^T - \Phi(X_1) \\ \vdots \\ \mathbf{Z}_n^T - \Phi(X_n) \end{pmatrix} + O_p(c_n) \hat{=} A + O_p(c_n). \tag{A.2}$$

By the law of large numbers, we have

$$n^{-1}\widetilde{\mathbf{Z}}_n^T\widetilde{\mathbf{Z}}_n = n^{-1}\sum_{i=1}^n\{\mathbf{Z}_i-\Phi(X_i)^T\}\{\mathbf{Z}_i^T-\Phi(X_i)\}+O_p(c_n)$$
$$\xrightarrow{P}\mathrm{E}\big(\mathbf{Z}-E(\mathbf{Z}|X)(\mathbf{Z}-E(\mathbf{Z}|X)\big)^T=\mathbf{\Sigma}.$$

Hence, to prove the lemma, we only consider the limit of $n^{-1}\widetilde{\mathbf{Z}}_n^T\times\mathbf{P}_{\widetilde{\mathbf{B}}_n}\widetilde{\mathbf{Z}}_n$. It is easy to show that

$$n^{-1}\widetilde{\mathbf{Z}}_n^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\widetilde{\mathbf{Z}}_n=n^{-1}\mathbf{A}^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\mathbf{A}+O_p(c_n).$$

Let $(\mathbf{P}_{\widetilde{\mathbf{B}}_n})_{ij}\hat{=}p_{ij}$ and $(\mathbf{A})_{ij}\hat{=}a_{ij}=Z_{ij}-\mathrm{E}(Z_{ij}|X_i)$, where Z_{ij} is the jth component of random vector $\mathbf{Z}_i$, which represents the ith observation of $\mathbf{Z}$. Then the (i,j) component of $\frac{1}{n}\mathbf{A}^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\mathbf{A}$ is

$$\begin{aligned}\left(\frac{1}{n}\mathbf{A}^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\mathbf{A}\right)_{ij}&=\frac{1}{n}\sum_{k=1}^n\sum_{l=1}^n a_{ki}p_{kl}a_{lj}\\&=\frac{1}{n}\sum_{k=1}^n a_{ki}p_{kk}a_{kj}+\frac{1}{n}\sum_{k\neq l}^n a_{ki}p_{kl}a_{lj}\\&\hat{=}I_1+I_2.\end{aligned}$$

For the term I_2, we have

$$\mathrm{E}I_2^2=\frac{1}{n^2}\mathrm{E}\left\{\sum_{k_1\neq l_1}^n\sum_{k_2\neq l_2}^n a_{k_1i}p_{k_1l_1}a_{l_1j}a_{k_2i}p_{k_2l_2}a_{l_2j}\right\}.$$

Note that $(\mathbf{Z}_1,X_1),\ldots,(\mathbf{Z}_n,X_n)$ are iid, and that p_{ij} depends only on $\{X_1,\ldots,X_n\}$ for any pair (i,j). Because $\mathrm{E}(a_{k_1j}|X_{k_1})=0$, we have

$$\begin{aligned}&\mathrm{E}\big\{a_{k_1i}p_{k_1l_1}a_{l_1j}a_{k_2i}p_{k_2l_2}a_{l_2j}\big\}\\&=\mathrm{E}\big\{p_{k_1l_1}p_{k_2l_2}\mathrm{E}\big(a_{k_1i}a_{l_1j}a_{k_2i}a_{l_2j}|X_{k_1},X_{k_2},X_{l_1},X_{l_2}\big)\big\}\\&=\mathrm{E}\big\{p_{k_1l_1}p_{k_2l_2}\mathrm{E}\big(a_{l_1j}a_{k_2i}a_{l_2j}|X_{k_2},X_{l_1},X_{l_2}\big)\mathrm{E}\big(a_{k_1i}|X_{k_1}\big)\big\}\\&=0,\end{aligned}$$

when $k_1\neq k_2$ and $k_1\neq l_2$. Using the same argument and $p_{kl}=p_{lk}$, we have

$$\mathrm{E}I_2^2=\frac{1}{n^2}\sum_{k\neq l}^n\mathrm{E}\{a_{ki}p_{kl}a_{lj}\}^2+\frac{1}{n^2}\sum_{k\neq l}^n\mathrm{E}\{p_{kl}^2a_{ki}a_{kj}a_{lj}a_{li}\}.$$

Because the a_{ij}'s are uniformly bounded by condition 3,

$$\mathrm{E}I_2^2\le\frac{2C}{n^2}\sum_{k\neq l}^n p_{kl}^2\le\frac{2C}{n^2}\operatorname{tr}\big(\mathbf{P}_{\widetilde{\mathbf{B}}_n}^2\big)\le\frac{2C}{n},$$

where C is a constant. Hence

$$I_2=o_p(1).\tag{A.3}$$

Note that I_1 can be decomposed as

$$I_1=\frac{1}{n}\sum_{k=1}^n p_{kk}(a_{ki}a_{kj}-\mathrm{E}a_{ki}a_{kj})+\frac{1}{n}\sum_{k=1}^n p_{kk}\mathrm{E}a_{ki}a_{kj}\hat{=}J_1+J_2.$$

Because $\operatorname{tr}(\mathbf{S})=O_p(1/h)$ and $\operatorname{tr}(\mathbf{S}\mathbf{S}^T)=O_p(1/h)$, it is easy to know that $\operatorname{tr}(\mathbf{P}_{\widetilde{\mathbf{B}}_n})=\frac{n}{I}+O_p(1/h)$. Hence,

$$J_2=\frac{1}{I}\Sigma_{ij}+O_p\left(\frac{1}{nh}\right),\tag{A.4}$$

where Σ_{ij} is the (i,j)th element of $\mathbf{\Sigma}$. Furthermore, if we can show that

$$1\ge p_{kk}\ge\frac{1}{I}+O_p\left(\frac{1}{nh}\right),\tag{A.5}$$

then, by

$$\operatorname{tr}\big(\mathbf{P}_{\widetilde{\mathbf{B}}_n}\big)=\sum_{k=1}^n p_{kk}=\frac{n}{I}+O_p\left(\frac{1}{h}\right),\tag{A.6}$$

it is easy to show that

$$\frac{1}{n}\sum_{k=1}^n(p_{kk}-I^{-1})^2=O_p\left(\frac{1}{nh^2}\right).$$

By the law of large numbers, J_1 is bounded as

$$\begin{aligned}J_1&=\frac{1}{n}\sum_{k=1}^n(p_{kk}-I^{-1})(a_{ki}a_{kj}-\mathrm{E}a_{ki}a_{kj})\\&\quad+\frac{1}{nI}\sum_{k=1}^n(a_{ki}a_{kj}-\mathrm{E}a_{ki}a_{kj})\\&\le\frac{1}{n}\left\{\sum_{k=1}^n(p_{kk}-I^{-1})^2\right\}^{1/2}\left\{\sum_{k=1}^n(a_{ki}a_{kj}-\mathrm{E}a_{ki}a_{kj})^2\right\}^{1/2}\\&\quad+o_p(1)\\&=o_p(1).\end{aligned}\tag{A.7}$$

By (A.4) and (A.7), we have

$$I_1=\frac{1}{I}\Sigma_{ij}+o_p(1).\tag{A.8}$$

By (A.3) and (A.8), Lemma A.2 holds, that is,

$$\frac{1}{n}\widetilde{\mathbf{Z}}_n^T(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}})\widetilde{\mathbf{Z}}_n\xrightarrow{P}\frac{I-1}{I}\mathbf{\Sigma}.$$

We now need to establish (A.5). Consider the projection matrix

$$\mathbf{P}_{\widetilde{\mathbf{B}}_{n1}}=(\mathbf{I}-\mathbf{S})\mathbf{B}_{n1}\big(\mathbf{B}_{n1}^T(\mathbf{I}-\mathbf{S}^T)(\mathbf{I}-\mathbf{S})\mathbf{B}_{n1}\big)^{-}\mathbf{B}_{n1}^T(\mathbf{I}-\mathbf{S}^T),$$

where $\mathbf{B}_{n1}$ is the first column vector of $\mathbf{B}_n$. It is easy to show that

$$\mathbf{P}_{\widetilde{\mathbf{B}}_n}\mathbf{P}_{\widetilde{\mathbf{B}}_{n1}}=\mathbf{P}_{\widetilde{\mathbf{B}}_{n1}}\mathbf{P}_{\widetilde{\mathbf{B}}_n}\qquad\text{and}$$
$$\mathbf{P}_{\widetilde{\mathbf{B}}_n}-\mathbf{P}_{\widetilde{\mathbf{B}}_{n1}}=\big(\mathbf{P}_{\widetilde{\mathbf{B}}_n}-\mathbf{P}_{\widetilde{\mathbf{B}}_{n1}}\big)^2\ge0.$$

By the definition of $\mathbf{S}$, it is easy to show that

$$\mathbf{S}=(\mathbf{S}_1,\ldots,\mathbf{S}_n)^T[\mathbf{I}+\operatorname{diag}\{O_p(c_n)\}],$$

where

$$\mathbf{S}_i=\left(\frac{K_h(X_1-X_i)}{nf(X_i)},\ldots,\frac{K_h(X_n-X_i)}{nf(X_i)}\right)^T.$$

Thus we have

$$\begin{aligned}\mathbf{B}_{n1}^T(\mathbf{I}-\mathbf{S}^T)(\mathbf{I}-\mathbf{S})\mathbf{B}_{n1}&=\sum_{i=1}^n\left(\sum_{j=1}^I\frac{K_h(X_j-X_i)}{nf(X_i)}\right)^2\{1+O_p(c_n)\}\\&\quad+I-2\sum_{i=1}^I\sum_{j=1}^I\frac{K_h(X_j-X_i)}{nf(X_i)}.\end{aligned}$$

Because

$$\sum_{i=1}^I\left(\sum_{j=1}^I\frac{K_h(X_j-X_i)}{nf(X_i)}\right)\{1+O_p(c_n)\}=O_p\left(\frac{1}{nh}\right)$$

and

$$\sum_{i=1}^n\left(\sum_{j=1}^I\frac{K_h(X_j-X_i)}{nf(X_i)}\right)^2\{1+O_p(c_n)\}=O_p\left(\frac{1}{nh}\right),$$

we have

$$\mathbf{B}_{n1}^T(\mathbf{I}-\mathbf{S}^T)(\mathbf{I}-\mathbf{S})\mathbf{B}_{n1} = I\left\{1+O_p\left(\frac{1}{nh}\right)\right\}.$$

Hence, for $i=1,\ldots,I$, we obtain

$$\begin{aligned}\left(\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)_{ii} &\geq \left(\mathbf{P}_{\widetilde{\mathbf{B}}_{n1}}\right)_{ii}\\ &= \frac{1}{I}\left\{1+O_p\left(\frac{1}{nh}\right)\right\}\left\{1-\sum_{j=1}^{I}\frac{K_h(X_j-X_i)}{nf(X_i)}\{1+O_p(c_n)\}\right\}^2\\ &= \frac{1}{I}+O_p\left(\frac{1}{nh}\right), \qquad i=1,\ldots,I.\end{aligned}$$

By a similar argument, we can show that

$$\left(\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)_{ii} \geq \frac{1}{I}+O_p\left(\frac{1}{nh}\right), \qquad i=I+1,\ldots,n.$$

This completes the proof of Lemma A.2.

Lemma A.3. Under conditions 1–6, we have

$$n^{-1}\widetilde{\mathbf{Z}}_n^T(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}})(\mathbf{I}-\mathbf{S})\mathbf{M} = O_p(c_n^2).$$

Proof. By a similar proof to that of Lemma A.2, and by Lemma A.1, it is easy to see that

$$(\mathbf{I}-\mathbf{S})\mathbf{M} = O_p(c_n).$$

Defining $\mathbf{P} = (P_1,\ldots,P_n)^T = (\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n})(\mathbf{I}-\mathbf{S})\mathbf{M}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}P_i^2 \leq n^{-1}\|(\mathbf{I}-\mathbf{S})\mathbf{M}\|^2 = O_p(c_n^2).$$

By (A.2) and the Cauchy–Schwartz inequality,

$$\begin{aligned}\frac{1}{n}\widetilde{\mathbf{Z}}_n\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})\mathbf{M} &= \frac{1}{n}\sum_{i=1}^{n}P_i\{\mathbf{Z}_i-\Phi(X_i)+O_p(c_n)\}\\ &= n^{-1}\sum_{i=1}^{n}P_i\{\mathbf{Z}_i-\Phi(X_i)\}+O_p(c_n^2).\end{aligned}$$

We now deal with the first term. Note that P_i depends only on X random variables and $\mathrm{E}(\mathbf{Z}_i|X_i)=\Phi(X_i)$. Hence, for some constant C,

$$\mathrm{E}\left[n^{-1}\sum_{i=1}^{n}P_i\{\mathbf{Z}_i-\Phi(X_i)\}\right]^2 \leq Cn^{-2}\sum_{i=1}^{n}P_i^2 = O(n^{-1}c_n^2).$$

This leads to

$$\frac{1}{n}\widetilde{\mathbf{Z}}_n\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})\mathbf{M} = O_p(c_n^2).$$

Hence, Lemma A.3 holds.

Lemma A.4. Under conditions 1–6, we have

$$n^{-1/2}\widetilde{\mathbf{Z}}_n^T(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}})(\mathbf{I}-\mathbf{S})\boldsymbol{\epsilon}_n \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{I-1}{I}\sigma^2\boldsymbol{\Sigma}\right).$$

Proof. By Lemma A.1, we have $\mathbf{S}\boldsymbol{\epsilon}_n = O_p(c_n)$. By similar arguments as in Lemma A.3 and under condition 6, we have

$$\frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_n\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)\mathbf{S}\boldsymbol{\epsilon}_n = O_p(\sqrt{n}c_n^2) = o_p(1).$$

Therefore, by Slutsky's theorem, the central limit theorem (see van der Vaart 1998), and Lemma A.2, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_n^T\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})\boldsymbol{\epsilon}_n\\ &\quad= \frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_n^T\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)\boldsymbol{\epsilon}_n+o_p(1)\\ &\quad\xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{I-1}{I}\sigma^2\boldsymbol{\Sigma}\right).\end{aligned}$$

Proof of Theorem 1

By (6) and Lemma A.2, we have

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \sqrt{n}\left(\widetilde{\mathbf{Z}}_n^T\widetilde{\mathbf{Z}}_n-\widetilde{\mathbf{Z}}_n^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\widetilde{\mathbf{Z}}_n\right)^{-1}\widetilde{\mathbf{Z}}_n^T\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})(\mathbf{M}+\boldsymbol{\epsilon}_n).$$

By Lemmas A.2 and A.3, we have

$$\begin{aligned}&\sqrt{n}\left(\widetilde{\mathbf{Z}}_n^T\widetilde{\mathbf{Z}}_n-\widetilde{\mathbf{Z}}_n^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\widetilde{\mathbf{Z}}_n\right)^{-1}\widetilde{\mathbf{Z}}_n^T\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})\mathbf{M}\\ &\qquad= O_p(\sqrt{n}c_n^2) = o_p(1).\end{aligned}$$

Hence,

$$\begin{aligned}&\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\\ &\quad= \sqrt{n}\left(\widetilde{\mathbf{Z}}_n^T\widetilde{\mathbf{Z}}_n-\widetilde{\mathbf{Z}}_n^T\mathbf{P}_{\widetilde{\mathbf{B}}_n}\widetilde{\mathbf{Z}}_n\right)^{-1}\widetilde{\mathbf{Z}}_n^T\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})\boldsymbol{\epsilon}_n+o_p(1).\end{aligned}$$

By Lemma A.4, we have

$$\frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_n^T\left(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\right)(\mathbf{I}-\mathbf{S})\boldsymbol{\epsilon} \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{I-1}{I}\sigma^2\boldsymbol{\Sigma}\right).$$

Therefore, using Lemma A.2, we conclude that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \frac{I}{I-1}\sigma^2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \frac{I}{I-1}\sigma^2\boldsymbol{\Sigma}^{-1}\right).$$

This completes the proof of Theorem 1.

Lemma A.5. Under conditions 1–6,

$$\lambda_i\{(\mathbf{I}-\mathbf{SP})(\mathbf{I}-\mathbf{SP})^T\} \geq \frac{(\sqrt{I}-1)^2}{I}+O_p(c_n)$$

holds uniformly for $\{\lambda_i, i=1,\ldots,n\}$, where $\lambda_i(\mathbf{A})$ denotes the eigenvalue of matrix $\mathbf{A}$.

Proof. We first consider the eigenvalues of the matrix $\mathbf{SPS}^T$. Because

$$\begin{aligned}\mathbf{P} &= (\mathbf{I}-\mathbf{1}_n\mathbf{1}_n^T/n)\mathbf{B}_n(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T\\ &= \mathbf{B}_n(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T-\mathbf{1}_n\mathbf{1}_n^T/n\\ &= \frac{1}{I}\mathbf{I}_G\otimes(\mathbf{1}_I\mathbf{1}_I^T)-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\end{aligned}$$

and $\mathbf{S}\mathbf{1}_n=\mathbf{1}_n$, it is easy to show that

$$\begin{aligned}\mathbf{SPS}^T &= \mathbf{SB}_n(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T\mathbf{S}^T-\mathbf{1}_n\mathbf{1}_n^T/n\\ &= \frac{1}{I}\mathbf{S}\left(\mathbf{I}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{S}^T+\frac{1}{I}\mathbf{S}\{\mathbf{I}_G\otimes(\mathbf{1}_I\mathbf{1}_I^T)-\mathbf{I}\}\mathbf{S}^T\\ &\quad-\left(1-\frac{1}{I}\right)\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\\ &\hat{=} K_1+K_2+K_3.\end{aligned}$$

For the term K_2, we have

$$\begin{aligned}(K_2)_{ij} &= \sum_{l=1}^{n}\sum_{k=1}^{n}\mathbf{S}_{il}\frac{1}{I}\{\mathbf{I}_G\otimes(\mathbf{1}_I\mathbf{1}_I^T)-\mathbf{I}\}_{lk}\mathbf{S}_{jk}\\ &= \frac{1}{I}\sum_{s=0}^{G-1}\sum_{l=1}^{I}\sum_{k=1,k\neq l}^{I}\mathbf{S}_{i(sG+l)}\mathbf{S}_{j(sG+k)}\\ &= \frac{1}{I}\sum_{l=1}^{I}\sum_{k=1,k\neq l}^{I}\sum_{s=0}^{G-1}\mathbf{S}_{i(sG+l)}\mathbf{S}_{j(sG+k)}.\end{aligned}$$

By (A.1) and Lemma A.1,

$$\sum_{s=0}^{G-1}\mathbf{S}_{i(sG+l)}\mathbf{S}_{j(sG+k)} = \left\{\sum_{s=0}^{G-1}\frac{K_h(X_{sG+l}-X_i)}{nf(X_i)}\frac{K_h(X_{sG+k}-X_j)}{nf(X_j)}\right\}\{1+O_p(c_n)\} = \frac{G}{n^2}\{1+O_p(c_n)\} = \frac{1}{In}\{1+O_p(c_n)\},$$

which holds uniformly for $i,j=1,\ldots,n$. Hence, we obtain

$$(K_2)_{ij} = \frac{1}{I}\sum_{l=1}^{I}\sum_{k=1,k\neq l}^{I}\frac{1}{In}\{1+O_p(c_n)\} = \frac{I-1}{In}\{1+O_p(c_n)\}$$

or

$$K_2 = \frac{I-1}{I}\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\{1+O_p(c_n)\} = -K_3\{1+O_p(c_n)\}.$$

Therefore,

$$\mathbf{SPS}^T = K_1 + K_3\cdot O_p(c_n).$$

It is obvious that the eigenvalues of K_1 satisfy

$$0\le\lambda_i(K_1)\le\frac{1}{I},\qquad i=1,\ldots,n.$$

Thus the eigenvalues of $\mathbf{SPS}^T$ have

$$0\le\lambda_i(\mathbf{SPS}^T)\le\frac{1}{I}+O_p(c_n). \tag{A.9}$$

On the other hand, for any vector $\mathbf{z}$ satisfying $\|\mathbf{z}\|=1$ and letting $y^2=\mathbf{z}^T\mathbf{SPS}^T\mathbf{z}$, $y\ge0$, we have

$$\mathbf{z}^T(\mathbf{I}-\mathbf{SP})(\mathbf{I}-\mathbf{SP})^T\mathbf{z} = 1-\mathbf{z}^T(\mathbf{SP}+\mathbf{PS}^T)\mathbf{z}+\mathbf{z}^T\mathbf{SPS}^T\mathbf{z} \ge 1-2y+y^2.$$

By (A.9),

$$y^2\le\frac{1}{I}+O_p(c_n).$$

Hence

$$\mathbf{z}^T(\mathbf{I}-\mathbf{SP})(\mathbf{I}-\mathbf{SP})^T\mathbf{z}\ge\left(1-\frac{1}{\sqrt{I}}\right)^2+O_p(c_n).$$

This leads to the conclusion of Lemma A.5.

Proof of Theorem 3

By (9), we have

$$\widehat{\mathbf{M}} = \{\mathbf{I}-(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\}\{\mathbf{B}_n\boldsymbol{\alpha}_n+\mathbf{M}+\boldsymbol{\epsilon}_n\} = (\mathbf{I}-\mathbf{SP})^{-1}\mathbf{S}(\mathbf{I}-\mathbf{P})\mathbf{B}_n\boldsymbol{\alpha}_n+\{\mathbf{I}-(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\}\mathbf{M} + \{\mathbf{I}-(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\}\boldsymbol{\epsilon}_n.$$

Because $\sum_{i=1}^{G}\alpha_i=0$, it is easy to show that

$$(\mathbf{I}-\mathbf{SP})^{-1}\mathbf{S}(\mathbf{I}-\mathbf{P})\mathbf{B}_n\boldsymbol{\alpha}_n=0.$$

Then we have

$$\widehat{\mathbf{M}}-\mathbf{M} = -(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\mathbf{M}+\{\mathbf{I}-(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\}\boldsymbol{\epsilon}_n$$

and

$$\mathrm{MSE}\{\widehat{\mathbf{M}}\} = \frac{1}{n}\|(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\mathbf{M}\|^2 + \frac{1}{n}\mathrm{E}\big\|\{\mathbf{I}-(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\}\boldsymbol{\epsilon}_n\big\|^2 = L_1+L_2.$$

By Fan (1992), we have

$$(\mathbf{I}-\mathbf{S})\mathbf{M} = \frac{\mu_2}{2}\mathbf{M}''h^2+o_p(h^2).$$

Hence, by Lemma A.5,

$$L_1\le\frac{I}{(\sqrt{I}-1)^2}\frac{1}{n}\frac{\mu_2^2h^4}{4}\sum_{i=1}^{n}\{m''(X_i)\}^2+o_p(h^4) = \frac{I}{(\sqrt{I}-1)^2}\frac{\mu_2^2h^4}{4}\mathrm{E}\{m''(X)\}^2+o_p(h^4). \tag{A.10}$$

For L_2, we have that

$$\big\|\{\mathbf{I}-(\mathbf{I}-\mathbf{SP})^{-1}(\mathbf{I}-\mathbf{S})\}\boldsymbol{\epsilon}_n\big\|^2 = \|(\mathbf{I}-\mathbf{SP})^{-1}\mathbf{S}(\mathbf{I}-\mathbf{P})\boldsymbol{\epsilon}_n\|^2 \le\left\{\frac{I}{(\sqrt{I}-1)^2}+O_p(c_n)\right\}\|\mathbf{S}(\mathbf{I}-\mathbf{P})\boldsymbol{\epsilon}_n\|^2.$$

Thus,

$$L_2\le\left\{\frac{I}{(\sqrt{I}-1)^2}+O_p(c_n)\right\}\frac{1}{n}\mathrm{E}\|\mathbf{S}(\mathbf{I}-\mathbf{P})\boldsymbol{\epsilon}_n\|^2 = \left\{\frac{I}{(\sqrt{I}-1)^2}+O_p(c_n)\right\}\frac{\sigma^2}{n}\mathrm{tr}(\mathbf{S}(\mathbf{I}-\mathbf{P})\mathbf{S}^T) \le\left\{\frac{I}{(\sqrt{I}-1)^2}+O_p(c_n)\right\}\frac{\sigma^2}{n}\mathrm{tr}(\mathbf{SS}^T). \tag{A.11}$$

It is not difficult to know that

$$\mathrm{tr}(\mathbf{SS}^T) = \sum_{i=1}^{n}\frac{\nu_0}{nhf(X_i)}+o_p\left(\frac{1}{nh}\right),$$

and by the law of large numbers, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\nu_0}{nhf(X_i)} = \frac{\sigma^2\nu_0|\Omega|}{nh}+o_p\left(\frac{1}{nh}\right).$$

Therefore, by (A.10) and (A.11), Theorem 3 holds.

Proof of Theorem 2

To deal with the situation of heteroscedastic error, following the proof of Theorem 1 and Lemma A.4, we only prove that

$$n^{-1/2}\widetilde{\mathbf{Z}}_n^T\big(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\big)(\mathbf{I}-\mathbf{S})\boldsymbol{\epsilon}_n\xrightarrow{D}\mathcal{N}(\mathbf{0},\mathbf{V}),$$

where

$$\mathbf{V} = \frac{(I-1)^2}{I^2}\mathrm{E}\sigma^2(X)\{\mathbf{Z}-\mathrm{E}(\mathbf{Z}|X)\}^T\{\mathbf{Z}-\mathrm{E}(\mathbf{Z}|X)\}+\frac{I-1}{I^2}\mathrm{E}\sigma^2(X)\cdot\boldsymbol{\Sigma}.$$

Similar to Lemma A.5 and its proof, we may consider the eigenvalues of the matrix $\mathbf{B}_n^T(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\mathbf{B}_n$. Then we know that the nonzero eigenvalues of this matrix are larger than a constant and that the vector $\mathbf{1}_n$ is the eigenvector of this matrix with zero eigenvalues. Hence, by a special form of $\mathbf{B}_n$, we can show that

$$\mathbf{P}_{\widetilde{\mathbf{B}}_n} = \frac{1}{I}\mathbf{I}_G\otimes(\mathbf{1}_I\mathbf{1}_I^T)+O_p\left(\frac{1}{nh}\right),$$

where $O_p(1/nh)$ denote a matrix whose elements are uniformly with order $1/nh$.

Finally, using Lemma A.1 and the same computation steps as in the proof of Lemma A.2, we obtain that

$$\frac{1}{n}\widetilde{\mathbf{Z}}_n^T\big(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\big)\mathbf{W}\big(\mathbf{I}-\mathbf{P}_{\widetilde{\mathbf{B}}_n}\big)\widetilde{\mathbf{Z}}_n\xrightarrow{P}\mathbf{V},$$

where $\mathbf{W}=\mathrm{diag}\{\sigma^2(X_1),\ldots,\sigma^2(X_n)\}$. Then, following the steps in the proof of Lemma A.4, Theorem 2 can be proved.

Proof of Theorem 4

To prove Theorem 4, following the steps of the proof of Theorem 3, we need only consider $\mathrm{tr}(\mathbf{S}(\mathbf{I}-\mathbf{P})\mathbf{W}(\mathbf{I}-\mathbf{P})\mathbf{S}^T)$. However, we know the detail form of the matrices $\mathbf{P}$ and $\mathbf{W}$, by the properties of $\mathbf{S}$ and some computation. Thus Theorem 4 is easily proved.

Proof of Theorem 5

To prove Theorem 5, we can still follow the steps from (5) to (7), adopting a local linear regression technique. The estimate of $\mathbf{M}$ is

$$\widehat{\mathbf{M}} = \mathbf{S}(\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\beta}),$$

where $\mathbf{S} = \mathrm{diag}(\mathbf{S}_1, \ldots, \mathbf{S}_J)$ and $\mathbf{S}_j$ is a smoothing matrix that depends only on the observations $\{X_{gij},\ g = 1, \ldots, G,\ i = 1, \ldots, I\}$. Hence we can also obtain the estimate of $\boldsymbol{\beta}$ as

$$\widehat{\boldsymbol{\beta}} = (\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^T\mathbf{P}_{\widetilde{\mathbf{B}}}\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^T(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{B}}})\widetilde{\mathbf{Y}},$$

where $\widetilde{\mathbf{Y}} = (\mathbf{I}-\mathbf{S})\mathbf{Y}$, $\widetilde{\mathbf{B}} = (\mathbf{I}-\mathbf{S})\mathbf{B}$, $\widetilde{\mathbf{Z}} = (\mathbf{I}-\mathbf{S})\mathbf{Z}$, and $\mathbf{P}_{\widetilde{\mathbf{B}}} = \widetilde{\mathbf{B}}(\widetilde{\mathbf{B}}^T\widetilde{\mathbf{B}})^T\widetilde{\mathbf{B}}^T$.

Noticing that the special structures of $\mathbf{S}$, $\widetilde{\mathbf{B}}$, $\widetilde{\mathbf{Z}}$, and $\mathbf{P}_{\widetilde{\mathbf{B}}}$ are somewhat different from those in the proof of Theorem 1, the rank of $\mathbf{P}_{\widetilde{\mathbf{B}}}$ is of the order $G + o_p(G)$, and the sample size is GIJ. Following the proof of Lemma A.2, we can show that

$$\frac{1}{GI}(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^T\mathbf{P}_{\widetilde{\mathbf{B}}}\widetilde{\mathbf{Z}}) \xrightarrow{P} \boldsymbol{\Sigma}^{**},$$

where

$$\boldsymbol{\Sigma}^{**} = \frac{IJ-1}{IJ}\mathbf{I}_G \otimes \boldsymbol{\Sigma} + \frac{1}{IJ}\mathbf{I}_G \otimes \boldsymbol{\Sigma}^* - \frac{1}{IJ}\mathbf{1}_G\mathbf{1}_G^T \otimes \boldsymbol{\Sigma}^*. \qquad \text{(A.12)}$$

Then, following the proof of Theorem 1 and using matrix inverse computation, Theorem 5 can be obtained.

Proof of Theorem 6

The difference of proof between Theorem 6 and Theorem 5 lies in the limiting matrix of $\frac{1}{GI}(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^T\mathbf{P}_{\widetilde{\mathbf{B}}}\widetilde{\mathbf{Z}})$. By using the condition that $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_J$ are independent, we have $\boldsymbol{\Sigma}^* = \mathbf{0}$. It follows from (A.12) that

$$\frac{1}{GI}(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}} - \widetilde{\mathbf{Z}}^T\mathbf{P}_{\widetilde{\mathbf{B}}}\widetilde{\mathbf{Z}}) \xrightarrow{P} \frac{IJ-1}{IJ}\mathrm{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_J).$$

Theorem 6 follows from a proof similar to that of Theorem 5.

[Received December 2003. Revised September 2004.]

REFERENCES

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Craig, B. A., Black, M. A., and Doerge, R. W. (2003), "Gene Expression Data: The Technology and Statistical Analysis," *Journal of Agricultural, Biological, and Environmental Statistics*, 8, 1–28.

Dudoit, S., Yang, Y. H., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002), "Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation," *Nucleic Acids Research*, 30, e15.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

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," *Proceedings of the National Academy of Science*, 101, 1135–1140.

Fan, J., and Yao, Q. (1998), "Efficient Estimation of Conditional Variance Functions in Stochastic Regression," *Biometrika*, 85, 645–660.

Grolleau, A., Bowman, J., Pradet-Balade, B., Puravs, E., Hanash, S., Garcia-Sanz, J. A., and Beretta, L. (2002), "Global and Specific Translational Control by Rampamycin in T Cells Uncovered by Microarrays and Proteomics," *Journal of Biology and Chemistry*, 277, 22175–22184.

Härdle, W., Liang, H., and Gao, J. (2000), *Partially Linear Models*, Heidelberg: Springer-Verlag.

Huang, J., Wang, D., and Zhang, C. (2003), "A Two-Way Semi-Linear Model for Normalization and Significant Analysis of cDNA Microarray Data," unpublished manuscript.

Huang, J., and Zhang, C. H. (2003), "Asymptotic Analysis of a Two-Way Semiparametric Regression Model for Microarray Data," Technical Report 2003-006, Rutgers University.

Kroll, T. C., and Wölfl, S. (2002), "Ranking: A Closer Look on Globalisation Methods for Normalisation of Gene Expression Arrays," *Nucleic Acids Research*, 30, e50.

Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression and Density Estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, 405–415.

Neyman, J., and Scott, E. (1948), "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

Opsomer, J. D., and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, 25, 186–211.

Rao, C. R., and Toutenburg, H. (1999), *Linear Models: Least Squares and Alternatives* (2nd ed.), New York: Springer-Verlag.

Ruppert, D. (1997), "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of the American Statistical Association*, 92, 1049–1062.

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

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 Variations and Assessment of Gene Effects," *Nucleic Acids Research*, 29, 2549–2557.

Comment

Chiara SABATTI

The normalization of cDNA arrays is an issue of high practical relevance, being an initial necessary step in the analysis of microarray data, which are nowdays used extensively in genetics and biological research. The impact of statistics on this area can be described as both considerably high and disappointingly low. Statisticians (e.g., Tseng, Oh, Rohlin, Liao, and Wong 2001; Yang et al. 2002) first pointed out that the "global normalization" technique proposed by the developers of array technology overlooked a series of measurable effects and in turn suggested more flexible and sensitive strategies whose value has been recognized by practitioners. Indeed, even commercial

Chiara Sabatti is Assistant Professor, Departments of Human Genetics and Statistics, University of California Los Angeles, Los Angeles, CA 90095.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000439

software currently incorporates methodology inspired by such contributions as those by Tseng et al. (2001) and Yang et al. (2002). Fan, Peng, and Huang further enlarge the bag of statistical tools available for normalization. This represents a clear success. Nonetheless, disappointing that evidence from statistics has not motivated—so far—the development of more reliable technology or better understanding of the nature of these biases. For example, the results from different normalization strategies, including the one described in the present article, clearly indicate a strong intensity effect (m) that is highly variable from slide to slide. Although statistical correction is possible, there is very little understanding of the biochemical process that produces such biases. I wish that this statistically detected distortion was taken more seriously and that more energy was invested in the understanding of its basis and development of better technology.

The normalization strategy described in this article exploits the presence on one array of multiple spots measuring the expression value of the same gene. This design allows the authors to do without the assumption of zero average expression change across all (or a crudely identified large subset of) genes with similar intensity. The analysis of the experiments for which this assumption is more problematic will benefit most from the suggested methodology. One such case is represented by custom arrays, where the spotted genes are preselected according to the "suspicion" that they may be affected by the treatment under investigation. For example, consider arrays that are printed only with genes believed to be expressed in a specific tissue ("brain array") and used to study the effects of biological treatments known to impact the same tissue. More striking, Geschwind et al. (2001) pioneered the use of a subtraction technique in a first step to identify genes that appear differentially expressed in two cell lines under study; in a second step, these genes are spotted on an array, and the amounts of their transcripts are further quantified. In this situation, one clearly expects the majority of genes to change expression, and the use of traditional normalization techniques presents difficulties.

Fan, Peng, and Huang show that the iterative procedure they suggest leads to consistent estimates as the number of genes with replicate spots tends to infinity, as long as the replicates are placed "appropriately" on the array. This result is reassuring for statisticians, but also has important experimental design implications. A more detailed analysis of when asymptotic behavior "kicks in" would be very useful for determining how many replicate spots should be printed on an array—currently this number tends to be rather small, with an unsatisfactory spatial distribution. The fact that presently only a few genes are replicated (so that the relevant sample size for their method is small) may have guided the authors' model choice that appears questionable; they use additive row and column effects to model the print-tip means. Besides an economy of parameters, I fail to understand why one should not use a factor with as many levels as print tips, because this seems to correspond more directly with the technological process used in spotting the arrays. Arguably, the same asymptotic results would hold with this more general model, and again, important design suggestions may be derived in terms of how many replicates are needed to estimate these effects.

From a methodological standpoint, the focus of the article is on partial consistency in the presence of nuisance parameters. Asymptotic analysis is carried out sending G to infinity, that is, the number of genes in the study (albeit with replicate spots), rather than the number of replicate hybridizations. This appropriately adapts to the nature of array experiments that always survey a large number of genes, with few replicates. Indeed, a key element for successful statistical analysis in this context has been the identification of features shared by a large number of genes that can be estimated, turning the curse of dimensionality into a "blessing" (Donoho 2000): print-tip and dye effects as in the present article; variances of expression values, assuming that genes with similar intensities share a common variance (see, e.g., Baldi and Long 2001); the percentage of genes that do not experience changes in expression (see, e.g., Storey and Tibshirani 2003; Storey, Taylor, and Siegmund 2004); and the distribution of the expression values for nonchanger genes (Efron, Tibshirani, Storey, and Tusher 2001; Sabatti, Karsten, and Geschwind 2002). Often the statistical methodologies used to take advantage of the large number of genes fall under the umbrella of empirical Bayes methods (Li and Wong 2001; Newton, Kendziorski, Richmond, Blattner, and Tsui 2001; Efron, Tibshirani, Storey, and Tusher 2001; and many others). The results of Fan, Peng, and Huang are in terms of partial consistency in presence of nuisance parameters, bringing to the attention of the statistical community a different paradigm that can be used successfully with the same goal.

Can we identify other features in microarray data whose dimension does not increase with the number of genes and that are of scientific interest? The deconvolution-type model presented by Liao et al. (2003) may be one such example. The goal is to reconstruct, from gene expression data, the changes in concentration of regulatory proteins. To sketch the biological question, recall that the transcription of genes is controlled in large part by regulatory proteins that switch back and forth between active and inactive states in response to varying cell conditions. A relatively small number of such proteins is responsible for the regulation of the entire genome. Liao et al. (2003) noted the technical difficulty of directly measuring the changes in concentration of these transcription factors and attempt to reconstruct them from the variation in expression levels of their known targeted genes. The authors consider the following model for the log ratio of expression of gene i in experiment t:

$$y_{it} = \sum_{j=1}^{L} a_{ij} p_{jt} + \epsilon_{it}, \qquad t = 1, \ldots, M,\ i = 1, \ldots, G,$$

where a_{ij} quantifies the control strength of transcription factor j on gene i, p_{jt} is the concentration of active form of transcription factor j in experiment t, and ϵ_{it} is an error term. The total number of genes is indicated by G, whereas L and M represent the total number of transcription factors and experiments. In this setting, a large number of a_{ij} are equal to 0, as each gene is regulated by only one or at most few transcription factors. The number and the position of these 0's leads to an identifiability condition for the parameters a and p that is analogous to that reported by Anderson (1984). Liao et al. (2003) described

an iterative estimation procedure, but only sketched the evaluation of the properties of this estimator. In light of the analysis presented by Fan, Peng, and Huang $\{a_{ij}\}$ can be described as a nuisance parameter, with dimension increasing with increasing G. It might be possible to carry out the analogy between these two models at a further, more substantive level and obtain a partial consistency result for p.

Another context in which the methodology described in the article may be applicable is normalization and probe weighting in genotyping arrays. This technology is based on the same principles of gene expression arrays and was introduced at roughly the same time (Pastinen et al. 2000). Its commercialization by such companies as Affymetrix has lately increased its popularity, attracting the attention of statisticians (e.g., Hao, Li, Rosenow, and Wong 2004). One chip is used to determine the alleles of one individual at tens of thousands of genomic locations, where it is known that two variants, different at one nucleotide, are present in the population. cDNA corresponding to each of the two polymorphisms is synthesized on the array. After hybridization with a sample of DNA collected from an individual, the intensity of the spots corresponding to each allele is compared and used to define a genotype. More precisely, in the Affymetrix genotype array, a single SNP is assayed using 40 spots, each spot containing multiple copies of a different probe. Twenty of these represent a match to 1 of the 2 polymorphisms, and 20 are mismatches used to measure cross-hybridization. Half of the match probes (10) are complementary to the first allele (a in what follows), and half are complementary to the second (b allele). The probes complementary to one allele are further divided into two groups, according to which strand of DNA (sense or antisense) they reproduce. Each of the five probes in one of these groups differs because of the position of the polymorphism; one probe has the polymorphic base right in the middle of its length, and the others are obtained by sliding it by one, two, or three base pairs. The mismatch probes differ from corresponding match probes at the central nucleotide. This setting is very similar to that used in gene expression arrays by this same company and opens up the problem of how to best combine the signal coming from all of the 20 probe pairs (match and corresponding mismatch). In the context of gene expression experiments, Li and Wong (2001) and Irizarry et al. (2002) analyzed this question in detail and showed, with a series of data analyses, that the sensitivity of probes varies widely and that different probes should receive different weights in assessing the value of the results. The current Affymetrix algorithm for genotype calling (Liu et al. 2003) performs quality control checks that may result in the exclusion of the signal coming from some probes, but does not carry out a systematic probe selection and weighting. The introduction of such steps may help reduce the no-call rate and facilitate the interpretation of the overall intensity at one marker in terms of DNA copy number (Lin et al. 2004). The problem of defining a summary value starting from multiple probes is not unrelated to that of normalization considered in the article. To clarify this point, consider the model for the difference between matched and mismatched probes proposed by Li and Wong (2001),

$$y_{ijt} = \theta_{it}\phi_j + \epsilon_{ijt}, \qquad t = 1, \ldots, M,\ i = 1, \ldots, G,$$

where θ_{it} is the expression value for gene i in experiment t, ϕ_j is the affinity of probe j, and ϵ_{ijt} is a Gaussian iid error term. To make relevant inference on θ_{it} (which is linked to the genotype of the individual t), we need to estimate ϕ_j. But unlike the situation considered by Fan, Peng, and Huang, the probe effects ϕ_j are gene-specific (and maybe should be denoted by ϕ_{j_i} for clarity). This makes it impossible to take advantage of $G \to \infty$ for their estimation, unless a model based on probe characteristics (e.g., position of the mutation, nucleotide content) is used. However, in the context of genotyping arrays, one can meaningfully assume that the number of experiments T (genotyped individuals) will be large. Indeed, there is considerable interest in using this technology to carry out association mapping studies, which would require genotyping of thousands of individuals. In such cases, $\{\theta_{it}\}$ may be considered nuisance parameters, and it may be possible to describe a consistent estimator for $\{\phi_{j_i}\}$.

ADDITIONAL REFERENCES

Anderson, T. (1984), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley.

Baldi, P., and Long, A. (2001), "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes," *Bioinformatics*, 17, 509–519.

Donoho, D. (2000), "The Curse and Blessing of Dimensionality," lecture delivered at the Conference Math Challenges of the 21st Century, August 2000, Aide-Memoire, available at *http://www-stat.stanford.edu/~donoho/Lectures/AMS2000/AMS2000.html*.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Geschwind, D. H., Ou, J., Easterday, M. C., Dougherty, J. D., Jackson, R. L., Chen, Z., Antoine, H., Terskikh, A., Weissman, I. L., Nelson, S. F., and Kornblum, H. I. (2001), "A Genetic Analysis of Neural Progenitor Differentiation," *Neuron*, 29, 325–339.

Hao, K., Li, C., Rosenow, C., and Wong, W. H. (2004), "Estimation of Genotype Error Rate Using Samples With Pedigree Information: An Application on GeneChip Mapping 10K Array," *Genomics*, 84, 623–630.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2002), "Exploration, Normalization, and Summaries of High-Density Oligonucleotide Array Probe-Level Data," *Biostatistics*, 4, 249–264.

Li, C., and Wong, W. H. (2001), "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection," *Proceedings of the National Academy of Science*, 98, 31–36.

Liao, J., Boscolo, R., Yang, Y., Tran, L., Sabatti, C., and Roychowdhury, V. (2003), "Network Component Analysis: Reconstruction of Regulatory Signals in Biological Systems," *Proceedings of the National Academy of Science*, 100, 15522–15527.

Lin, M., Wei, L.-J., Sellers, W. R., Lieberfarb, M., Wong, W. H., and Li, C. (2004), "dChipSNP: Significance Curve and Clustering of SNP Array-Based Loss-of-Heterozygosity Data," *Bioinformatics*, 20, 1233–1240.

Liu, W.-M., Di, X., Yang, G., Matsuzaki, H., Huang, J., Mei, R., Ryder, T. B., Webster, T. A., Dong, S., Liu, G., Jones, K. W., Kennedy, G. C., and Kulp, D. (2003), "Algorithms for Large-Scale Genotyping Microarrays," *Bioinformatics*, 19, 2397–2403.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Pastinen, T., Raitio, M., Lindroos, K., Tainola, P., Peltonen, L., and Syvnen, A.-C. (2000), "A System for Specific, High-Throughput Genotyping by Allele-Specific Primer Extension on Microarrays," *Genome Research*, 10, 1031–1042.

Sabatti, C., Karsten, S., and Geschwind, D. (2002), "Thresholding Rules for Recovering a Sparse Signal From Microarray Experiments," *Mathematical Biosciences*, 176, 17–34.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Ser. B, 66, 187–205.

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genome-Wide Studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002), "Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation," *Nucleic Acids Research*, 30, e15.

Comment

Bruce A. CRAIG

1. INTRODUCTION

I would like to begin by congratulating Fan, Peng, and Huang on their innovative approach to normalizing spotted cDNA microarray data. Although numerous articles have been written on this topic, the use of duplicate spots in the normalization process has rarely been discussed. Given the apparent within-slide spatial effects (Smyth, Yang, and Speed 2003; Balázsi, Kay, Barabási, and Oltvai 2003), these spots have the potential to be very helpful not only in the normalization process itself, but also in assessing the effectiveness of a normalization procedure. Unfortunately, the focus of the article limits the discussion of this approach and leaves several questions regarding the implementation and effectiveness unanswered. It is on these questions that I want to focus here, in the hope of prompting further investigation into the use of these spots.

2. MULTIPLE VERSUS SINGLE SPOT NORMALIZATION

Much of the popularity of microarray analysis stems from the fact that scientists can view the simultaneous behavior of genes affected by a stimulus at the total genome level. Because duplicate spots take up space on the slide, there is a general reluctance to include them unless there are no other probes of interest (exceptions would be housekeeping or control probes). In addition, having duplicate spots often means that additional genetic material (i.e., labeled cDNA) must be extracted from the cells, because each probe must be in more than one well in the template (see sec. 2). Given these constraints, if duplicate spots are to be recommended, then their benefit needs to be made clear.

For example, there are already several useful normalization methods in the literature that attempt to account for the within-slide biases and do not rely on duplicate spots (Dudoit et al. 2002; Smyth and Speed 2003). Although the authors point out that these alternative approaches rely on some key assumptions that can be relaxed with a duplicate spot analysis, it is not clear how often these key assumptions will be unreasonable. Do the authors propose that duplicate spots should always be used, or that they should be used only in situations when these assumptions will likely be violated? In other words, the relative effectiveness of their approach is unclear. Although the authors' goals were not directed at addressing this issue, their simulation framework (using, e.g., example 2) could be used to compare their approach with, for example, the print-tip loess method.

3. FABRICATION OF SLIDES

I do not agree with the authors about the relative ease of printing several hundred repeated probes on a slide. There are several additional design aspects to consider before fabricating the slides. First, there is the choice of the probes. Because these will be used to estimate the intensity effect, it appears that one would want a set of probes that span a wide range of intensities. Is the selection of such a set an easy task? The composite loess is a similar single-spot normalization approach, in that only a subset of probes are used to estimate the intensity effect (Smyth and Speed 2003). For this approach, a set of control probes (that are not differentially expressed) have been specifically designed to provide a large intensity range. It is not obvious to me that a randomly selected set of probes to duplicate will readily provide this range.

Second, to estimate the print-tip block effects and treatment effect for each probe, it is necessary to spread the duplicate spots across print blocks. Although Fan, Peng, and Huang allude briefly to some of the construction aspects necessary for this to occur, I feel that the importance of this slide layout requires a more detailed discussion. Although their examples assume random placement of duplicates (or placement of duplicates that appears random), this is not readily feasible using the typical automated printing equipment.

3.1 cDNA Printing Process

A spotted cDNA slide contains numerous spots arranged in a row–column format, with each spot comprising cDNA of a particular probe. Printing is commonly performed by a robotic apparatus designed to spot genetic material at specific points on the slide using a printing tool. The printing tool is generally either a 4×4 or 4×12 grid of pins (although other arrangements are possible) that transfers genetic material from the wells of a microtiter plate (commonly a 96 or 384 well plate) to the surface of the slide. The microtiter plate or set of plates is also known as the template.

When using a solid pin printing tool, the printing tool first dips into a collection of template wells, each pin into a separate well, and collects genetic material on each pin tip. The printing tool then moves to the slide and prints this genetic material on the slide. The printing tool is washed, and the process is repeated for a different set of template wells. The sets of template wells and order of printing depends on the particular printing equipment.

Figure 1, from Craig, Black, and Doerge (2003), contains a simple example of the printing process for a small 256 spot microarray using a 4×4 grid of pins. The print tool picks up probe material from 16 adjacent wells on the template (black filled circles) and deposits the probes on the glass slide. This process continues, with the printing performed in such a manner as to allow each pin to print a subarray of spots that lie next to each other (gray spots).

3.2 Generating Duplicate Spots

One method of "replicating" spots that is usually available with the printing equipment is to repeatedly dip the printing tool into the same sets of template wells. Each additional dip is known as an offset (i.e., two dips is known as a double offset).

Bruce A. Craig is Associate Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907 (E-mail: *bacraig@stat.purdue.edu*).

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000448

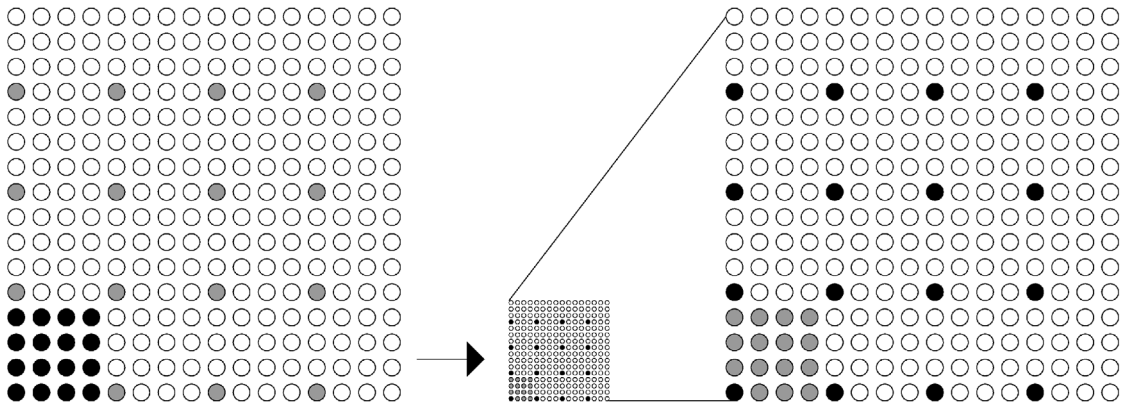

Figure 1. Example of a Simple 256-Spot Microarray. Panel (a) is the template, and (b) and (c) are the microarray drawn to scale and expanded respectively to show the correspondence between the template wells and the spots on the microarray.

If each pin tip is only associated with one print block, the duplicate spots will be printed in the same block, and it would not be possible to account for the spatial effects without making some additional assumptions.

With this in mind, Craig, Vitek, Black, Tanurdzic, and Doerge (2002) discussed designing the template because there is a one-to-one correspondence (provided that there are no offsets) between it and the slide. To use the proposed normalization approach, one must have duplicate wells for the selected probes. Although one could design the template on a well-by-well basis, rarely would one be willing to spend the time to do this. Typically, the template is created using a multi-tip pipette so that one can fill several template wells at once. Craig et al. (2002) described such an approach for a small-probe example involving an eight-tip pipette.

As mentioned earlier, this approach likely requires the generation of additional genetic material to fill multiple wells. Alternatively, one might be able to use one well provided that the template could be rotated (say 90 degrees) after the first set of probes have been printed. Because the microtiter plates are not square, this may not be feasible with the robotic printing equipment.

4. SUMMARY

The authors have proposed a very flexible approach to normalization of cDNA microarray data. Although they discuss numerous variations of this approach, I find the idea of using duplicate spots very intriguing. I have been a proponent of duplicate spots for some time, but have had a hard time convincing myself and scientists of their importance in the two-dye cDNA experiment. It would appear that within-slide replicates can be quite valuable, because they allow one to have better control of the normalization approach. There is always the concern with the single-spot normalization techniques of overfitting and removing more than just noise. I hope that these comments prompt further investigation of this approach. Again, I want to thank the authors for a very thought-provoking article.

ADDITIONAL REFERENCES

Balázsi, G., Kay, K. A., Barabási, A., and Oltvai, Z. (2003), "Spurious Spatial Periodicity of Co-Expression in Microarray Data Due to Printing Design," *Nucleic Acids Research*, 31, 4425–4433.

Craig, B. A., Vitek, O., Black, M. A., Tanurdzic, M., and Doerge, R. W. (2002), "Designing Microarray," in *Proceedings of the 2001 Kansas State University Conference on Applied Statistics in Agriculture*, ed. G. A. Milliken, Manhattan, KS: Department of Statistics, pp. 159–182.

Smyth, G. K., and Speed, T. (2003), "Normalization of cDNA Microarray Data," *Methods*, 31, 265–273.

Smyth, G. K., Yang, Y. H., and Speed, T. (2003), "Statistical Issues in Microarray Data Analysis," in *Functional Genomics: Methods and Protocols*, eds. M. J. Brownstein and A. B. Khodursky, *Methods in Molecular Biology*, Vol. 224, Totowa, NJ: Humana Press, pp. 111–136.

Comment

Jian HUANG and Cun-Hui ZHANG

Normalization is a critical component in microarray data analysis. Its purpose is to remove systematic biases in the observed expression values and to establish baseline intensity ratios across the whole dynamic range. Many researchers have considered this problem (see, e.g., Chen, Dougherty, and Bittner 1997; Kerr, Martin, and Churchill 2000; Yang et al. 2002; Yang, Dudoit, Luu, and Speed 2001; Tseng, Oh, Rohlin,

Jian Huang is Professor, Department of Statistics and Actuarial Science and Program in Public Health Genetics, University of Iowa, Iowa, IA 52242 (E-mail: *jian@stat.uiowa.edu*). Cun-Hui Zhang is Professor, Department of Statistics, Rutgers University, Piscataway, NJ 08855 (E-mail: *cunhui@stat.rutgers.edu*). Huang is supported in part by National Institutes of Health grant HL72288-01 and an Iowa Informatics Initiative grant. Zhang is supported in part by National Science Foundation grants DMS-02-03086 and DMS-04-05202.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000457

Liao, and Wong 2001; Park et al. 2003). In particular, Fan, Tam, Vande Woude, and Ren (2004) proposed a semilinear in-slide model (SLIM) method that makes use of replications of a subset of genes in an array. In the present interesting and stimulating article, Fan, Peng, and Huang 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. A focus of the article is the efficient estimation and calculation of semiparametric information for block effects in the case of fixed numbers of replications and arrays where the gene effects cannot be estimated consistently. This elegant result is a significant contribution to the semiparametric estimation theory, because the existing theory deals mainly with the case where the "nuisance parameters" can be consistently estimated.

We have proposed a two-way semilinear model (TW–SLM) for normalization and analysis of cDNA microarray data (Huang, Kuo, Koroleva, Zhang, and Soares 2003; Huang, Wang, and Zhang 2003; Huang and Zhang 2003). There are three main features of the TW–SLM that are different from the existing methods such as global and lowess normalization. First, normalization for each array in the TW–SLM is based on pooled information from all of the arrays. Second, the TW–SLM normalization curves and the gene effect parameters are estimated simultaneously in a single regression model. Each TW–SLM normalization curve does not attempt to fit the data from an individual array; rather, it fits the data after gene effects are adjusted for. This is in contrast to the lowess method, which estimates the normalization curves without adjusting for gene effects, which may cause the differentially expressed genes to be incorrectly "normalized" and result in a loss of power for detecting differentially expressed genes, because such genes tend to pull the normalization curve toward themselves. Third, in the framework of the TW–SLM, the uncertainty due to normalization is taken into account in the estimation of the standard errors of gene effects. The models proposed by Fan et al. (2004) and Fang, Peng, and Huang and the TW–SLM deal with the same problem with philosophically similar approaches, but our studies focus on different aspects of the problem and present orthogonal theoretical results. Thus we especially appreciate the opportunity to comment on this article. Here we give a brief description of the TW–SLM and some of its extensions, and discuss their relationship to the SLIM and its aggregations.

1. THE TWO–WAY SEMILINEAR MODEL

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. 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,

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

Let $\mathbf{z}_i \in \mathbb{R}^d$ be a covariate vector associated with the ith array. The TW–SLM is

$$y_{ij} = f_i(x_{ij}) + \mathbf{z}_i'\boldsymbol{\beta}_j + \varepsilon_{ij}, \qquad i = 1, \ldots, n, j = 1, \ldots, J, \tag{2}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^d$ is the effect associated with the jth gene, f_i is the intensity-dependent normalization curve for the ith array, and ε_{ij} is the residual term with mean 0 and variance σ_{ij}^2. For (2) to be identifiable, we restrict $\sum_{j=1}^{J} \boldsymbol{\beta}_j = \mathbf{0}$.

We call (2) TW–SLM because it contains the two-way ANOVA model as a special case with $f_i(x_{ij}) = \alpha_i$ and $\mathbf{z}_i = \mathbf{1}$. Our approach naturally leads to the general TW–SLM,

$$y_{ij} = f_i(x_{ij}) + \mathbf{z}_{ij}'\boldsymbol{\beta}_j + \varepsilon_{ij}, \tag{3}$$

which could be used to incorporate additional prior knowledge into the TW–SLM; see Section 3. The identifiability condition $\sum_j \boldsymbol{\beta}_j = \mathbf{0}$ is no longer necessary in (3) unless $\mathbf{z}_{ij} = \mathbf{z}_i$ as in (2).

The covariate vectors $\mathbf{z}_i$ in (2) 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, $\mathbf{z}_i = \mathbf{1}$, $i = 1, \ldots, n$. For an indirect comparison design using a common reference, we can introduce a two-dimensional covariate vector, $\mathbf{z}_i = (z_{i1}, z_{i2})'$. Let $\mathbf{z}_i = (1, 0)'$ if the ith array is of the type 1 sample versus the reference, and $\mathbf{z}_i = (0, 1)'$ if the ith array is of the type 2 sample versus the reference. Now $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2})'$ is a two-dimensional vector and $\beta_{j1} - \beta_{j2}$ represents the difference in the expression levels of gene j after normalization. The covariate vector $\mathbf{z}_i$ can also include other factors that contribute to the variations of the observed expression values.

2. MULTIWAY SEMILINEAR MODELS

Just as TW–SLM is a semilinear extension of two-way ANOVA, for datasets with a higher-dimensional structure, multiway ANOVA can be extended to multiway semilinear models (MW–SLM) in the same manner by including nonparametric and linear functions of covariates as the main and interactive terms/effects in the model. This approach is important in designing experiments and in understanding and interpretating the contribution of different effects and identifiability conditions, because it provides a direct, clear match between MW–SLM and ANOVA models. We describe our approach through the following examples motivated by real datasets.

In model (2), it is only made explicit that normalization curves, f_i, are array-dependent. It is straightforward to construct a 3W–SLM to normalize data at the printing-pin block level,

$$y_{ikj} = f_{ik}(x_{ikj}) + \mathbf{z}_i'\boldsymbol{\beta}_{kj} + \varepsilon_{ikj}, \tag{4}$$

with the identifiability condition $\sum_j \boldsymbol{\beta}_{kj} = \mathbf{0}$, where y_{ikj} and x_{ikj} are the log-intensity ratio and log-intensity product of gene j in the kth block of array i. Model (4) includes nonparametric components for the block and array effects and their interaction and linear components for the gene effects and their interaction with the block effects. It was used by Huang et al. (2003) to analyze the Apo A1 data (Callow, Dudoit, Gong, Speed, and Rubin 2000), as an application of the TW–SLM (for each fixed k) at the block level. The interaction between gene and block effects is present in (4) because we assume that different sets of genes are printed in different blocks. If a replication of the same (or entire) set of genes is printed in each block, then we may assume no interaction between gene and block effects ($\boldsymbol{\beta}_{kj} = \boldsymbol{\beta}_j$) in (4) and reduce it to the TW–SLM with (i, k) as a single index, treating a block in an array in (4) as an array in (2).

As an alternative to (4), we may also use constants to model the interaction between array and block effects as in ANOVA, resulting in the model

$$y_{ikj} = f_i(x_{ikj}) + \gamma_{ik} + \mathbf{z}_i'\boldsymbol{\beta}_{kj} + \varepsilon_{ikj}, \tag{5}$$

with identifiability conditions $\sum_i \gamma_{ik} = \sum_k \gamma_{ik} = 0$ and $\sum_{kj} \boldsymbol{\beta}_{kj} = \mathbf{0}$. This can be viewed as an extension of the three-way ANOVA model $Ey_{ikj} = \mu + \alpha_{i\cdot\cdot} + \gamma_{ik\cdot} + \beta_{\cdot k\cdot} + \beta_{\cdot kj} + \beta_{\cdot\cdot j}$ without the $\{i, j\}$ or three-way interaction, via $\mu + \alpha_{i\cdot\cdot} \Rightarrow f_i$ and $\beta_{\cdot k\cdot} + \beta_{\cdot kj} + \beta_{\cdot\cdot j} \Rightarrow \boldsymbol{\beta}_{kj}$. Note that the main block effects are represented by f_{ik} in (4) and by $\boldsymbol{\beta}_{kj}$ in (5).

Our approach easily accommodates designs where genes are printed multiple times in each array. Such a design is helpful for improving the precision and for assessing the quality of an array using the coefficient of variation (Tseng et al. 2001). Suppose that there is a matrix of printing-pin blocks in each array and that a replication of the same (or entire) set of genes is printed in each column of blocks in the matrix in each array. As in (5), a 4W–SLM can be written as

$$y_{icrj} = f_i(x_{icrj}) + \gamma_{icr} + \mathbf{z}_i'\boldsymbol{\beta}_{rj} + \varepsilon_{icrj} \tag{6}$$

for observations with the jth gene in the block at cth column and rth row of the matrix in the ith array, with identifiability conditions $\sum_i \gamma_{icr} = \sum_r \gamma_{icr} = 0$ and $\sum_{rj} \boldsymbol{\beta}_{rj} = \mathbf{0}$, with or without the three-way interaction or the interaction between the column and row effects in γ_{icr}. Note that the matrix of blocks does not have to match the physical columns and rows of printing-pin blocks. In model (6), the only nonparametric component is the array effects, and the block effects are modeled as in ANOVA. If the block effects also depend on the log-intensity product x_{icrj}, then we can combine the f_i and γ_{icr} in (6) as $f_{icr}(x_{icrj})$, resulting in the TW–SLM (for each fixed r) at the row level, equivalent to (4). If the replication of genes is not balanced, then we may use an MW–SLM derived from an ANOVA model with incomplete/unbalanced design or the modeling methodologies described in Section 3.

From the foregoing examples, it is clear that in an MW–SLM, the combination of main and interactive effects represented by a term is determined by the labeling of the parameter (not that of the covariates) of the term, as well as by the presence or absence of associated identifiability conditions. Furthermore, because the center of a nonparametric component [e.g., $\sum_j f_i(x_{ij})$ in a TW–SLM] is harder to interpret than the center of a parametric component, identifiability conditions are usually imposed on parametric components. As a result, a nonparametric component representing an interactive effect usually represents all of the associated main effects as well, and many MW–SLMs are equivalent to an implementation of the TW–SLM with a suitable partition of data, as in (4).

3. CONTROL GENES AND INCORPORATION OF PRIOR KNOWLEDGE IN THE MW–SLM

We describe three methods for incorporating prior knowledge in an MW–SLM: augmenting models, coding covariates, and imposing linear constraints. An important application of these methods is incorporating control genes in normalization.

In many customized microarray experiments, it is useful to include a set of control genes (e.g., spiked genes, housekeeping genes, and specially selected DNA sequences) with equal concentrations in the Cy5 and Cy3 channels. An important reason for using control genes is to calibrate scanning parameters; for example, intensity levels from the control genes can be used for tuning the laser power in each scanning channel to balance the Cy5 and Cy3 intensities. Specially constructed controls can also be used to aid in normalization (Yang et al. 2002). However, control genes do not necessarily show an observed 1:1 ratio because of experimental variations. In addition, because the number of control genes printed in a slide is often small, control genes may not cover the whole dynamic range of intensity levels, or the coverage may be too sparse. Therefore, it is in general not adequate to use just control genes as the basis for normalization.

Let y_{ik}^c and x_{ik}^c be the log-intensity ratio and the product of the kth control gene in the ith array, $i = 1, \ldots, n, k = 1, \ldots, K$. Then we can augment the TW–SLM (2) as

$$y_{ik}^c = f_i(x_{ik}^c) + \varepsilon_{ik}^c, \qquad y_{ij} = f_i(x_{ij}) + \mathbf{z}_i'\boldsymbol{\beta}_j + \varepsilon_{ij}. \tag{7}$$

The first equation is for the control genes, whose corresponding $\boldsymbol{\beta}_k^c$ are $\mathbf{0}$. Note that a common f_i is used in (7) for each array. Data from both control genes and genes under study contribute to the estimation of normalization curves as well as gene effects. With the inclusion of control genes, the identifiability condition $\sum_j \boldsymbol{\beta}_j = \mathbf{0}$ in (2) should be removed here, because it is neither necessary nor appropriate for (7).

We may also use the general TW–SLM (3) to model control genes by simply setting $\mathbf{z}_{ij} = \mathbf{0}$ if a control gene is printed at the jth spot in the ith array and $\mathbf{z}_{ij} = \mathbf{z}_i$ otherwise, where $\mathbf{z}_i$ is the design variable for the ith array as in (2).

A more general (but not necessarily simpler) method of incorporating prior knowledge is to impose constraints in addition to or as alternatives to the identifiability conditions in an MW–SLM. For example, we set $\boldsymbol{\beta}_j = \mathbf{0}$ if j corresponds to a control gene, and $\boldsymbol{\beta}_{j_1} = \cdots = \boldsymbol{\beta}_{j_r}$ if there are r replications of an experimental gene at spots $\{j_1, \ldots, j_r\}$ in each array.

4. LOCATION AND SCALE NORMALIZATION

The models that we have described earlier are for location normalization. It is often necessary to perform scale normalization to make arrays comparable in scale. The standard approach is to perform scale normalization after the location normalization, as discussed by Yang et al. (2001), so that normalization is completed in two separate steps. We can extend the MW–SLM to incorporate the scale normalization by introducing a vector of array-specific scale parameters $(\tau_1, \ldots, \tau_n)$, as in

$$\frac{y_{ij} - f_i(x_{ij})}{\tau_i} = \mathbf{z}_i'\boldsymbol{\beta}_j + \varepsilon_{ij}, \qquad i = 1, \ldots, n,\ j = 1, \ldots, J,$$

for the TW–SLM, where $\tau_1 \equiv 1$ and the τ_i's are restricted to be strictly positive. A more general model would allow τ_i to also depend on the total intensity levels.

5. INCORPORATING DATA QUALITY MEASUREMENTS AND ROBUST ESTIMATION

In the current literature, analysis of cDNA microarray data usually uses only a summary measure, such as the mean or median, of pixel intensities within a spot. We can use a weighted estimation criterion to allow for incorporation of quality measurements of a spot into the analysis. Let w_{ij} be the reciprocal to

the standard deviations associated with the log-intensity ratios. Such w_{ij} can be computed using a simple delta-method argument from the standard deviations of pixel intensities, which are usually available from scanner output files, for example, the GPA files from Axon's GenePix scanner. We can use $\mathbf{w} \equiv (w_{ij})$ to downweight spots with less uniform pixels, but such spots are usually of lower quality because of scratches, dust, and uneven hybridization. We can also use a weight matrix to filter out damaged spots in an array. The weighted estimation criterion function for the TW–SLM is

$$M_{\mathbf{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J} w_{ij}^2 m\big(y_{ij} - f_i(x_{ij}) - \mathbf{z}_i' \boldsymbol{\beta}_j\big).$$

Huang et al. (2003) and Huang and Zhang (2003) have studied the theoretical properties of the estimators of $\boldsymbol{\beta}_j$ and f_i when $m(t) = t^2$ and $w_{ij} \equiv 1$. It is also of interest to consider robust estimation approaches, for example, $m(t) = |t|$ or $m(t) = \rho(t)$, where ρ is Huber's ρ function. Ma et al. (2005) studied computation and inference when $m(t) = |t|$ in a general location and scale normalization model, and (Wang, Huang, Xie, Manzella, and Soares 2005) considered the case when $m(t) = \rho(t)$. However, theoretical properties of the estimators from such robust criterion functions have not been studied. If control genes are included with equal concentrations in two channels, then the M-estimation criterion function becomes

$$M_{\mathbf{w}}^c(\boldsymbol{\beta}, \mathbf{f}) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_{ik}^2 m\big(y_{ik}^c - f_i(x_{ik}^c)\big) + \sum_{j=1}^{J} w_{ij}^2 m\big(y_{ij} - f_i(x_{ij}) - \mathbf{z}_i' \boldsymbol{\beta}_j\big) \right\}.$$

6. RELATIONSHIP BETWEEN THE MW–SLM AND THE SEMILINEAR IN–SLIDE MODEL

SLIM (Fan et al. 2004, 2005) concerns a single array with block structure and replication of genes. In SLIM,

$$y_{kj} = f(x_{kj}) + \mathbf{v}_{kj}' \boldsymbol{\gamma} + \beta_j + \varepsilon_{kj} \tag{8}$$

for the kth replication of the jth gene, where $\boldsymbol{\gamma}$ is a vector of a relatively low dimensionality for block effects and $\mathbf{v}_{kj}$ are block indicators. From our point of view, SLIM is an extension of the model $Ey_{kj} = \mu + \beta_j$, with the mean effect μ replaced by the semiparametric $f(x_{kj}) + \mathbf{v}_{kj}' \boldsymbol{\gamma}$, especially when the $\mathbf{v}_{kj}$'s are treated as iid vectors as in the theoretical results of Fan, Peng, and Huang. In the deterministic case where each replication set of genes is printed in a separate collection of blocks, a 3W–SLM $Ey_{krj} = f(x_{krj}) + \gamma_{kr} + \beta_{rj}$ is an alternative as in (6) with a fixed array i, where r indicates blocks within the replication of the same set of genes. Note that we use $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to denote gene and block effects, whereas Fan et al. (2004), and Fan, Peng, and Huang used $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

A main difference between SLIM and TW–SLM is that SLIM has a single nonparametric component and TW–SLM has more than one and possibly unboundedly many nonparametric components. This remains the case when SLIM is applied to a "super-slide" composed of many arrays as its "super-blocks," resulting in $Ey_{ikj} = f(x_{ikj}) + \mathbf{v}_{ikj}' \boldsymbol{\gamma} + \beta_j$. The aggregated SLIM,

$$y_{ikj} = f_i(x_{ikj}) + \mathbf{v}_{ikj}' \boldsymbol{\gamma}_i + \beta_j + \varepsilon_{ikj}, \tag{9}$$

for the kth replication of the jth gene in the ith array is more closely related to TW–SLM, because (6) is an alternative to (9) when blocks are nested within replications and (9) is identical to (2) for the two-sample direct comparison design ($\mathbf{z}_i = \mathbf{1}$) when there is no replication or block effect.

Let $\widetilde{\mathbf{x}}_{ikj} = (x_{ikj}, v_{ikj})$ and $\widetilde{f}_i(\widetilde{\mathbf{x}}_{ikj}) = f_i(x_{ikj}) + \mathbf{v}_{ikj}' \boldsymbol{\gamma}_i$ in (9). Suppose that $\widetilde{\mathbf{x}}_{ikj}$ are iid as in Fan, Peng, and Huang article. The theoretical results of Huang et al. (2003) and Huang and Zhang (2003) are directly applicable for the estimation of $\widetilde{f}_i$ and β_j in (9) with a single replication, and their proofs work for multiple replications with multiple arrays by including the covariate vectors for the block effects in the basis space for the approximation of $\widetilde{f}_i$, but that does not provide separate estimates for $f_i(x_{ikj})$ and $\mathbf{v}_{ikj}' \boldsymbol{\gamma}_i$ within $\widetilde{f}_i$. Instead, they took a nonparametric approach in modeling the block effects as in (4). Because the focus of the article is the block effects in the case of (9) and the focus of our work is the normalization and resulting estimator $\widehat{\beta}_j$ for the gene effects in the absence of replications within arrays, the approaches and results of these studies complement each other well in many ways.

For the statistical theory concerning normalization of microarrays, the simple TW–SLM $Ey_{ij} = f_i(x_{ij}) + \beta_j$, with $\mathbf{z}_i = \mathbf{1}$ in (2), and the SLIM $Ey_{ij} = \widetilde{f}(\widetilde{x}_{ij}) + \beta_j$, with $k \sim i$ and $\widetilde{f}_i(\widetilde{x}_{ij}) = f(x_{ij}) + \mathbf{v}_{ij}' \boldsymbol{\gamma}$ in (8), provide the most direct comparison. A crucial element in the analysis of such models is the information operator for the estimation of β_j. Assume that linear estimators of f_i and $\widetilde{f}$ are used with random smoothing matrices $\mathbf{A}_i$ and $\mathbf{S}$, depending on covariates x_{ij} and $\widetilde{x}_{ij}$. For the simple TW–SLM, the information operator is approximately $\mathbf{I} - n^{-1} \sum_{i=1}^{n} \mathbf{A}_i$, whereas for SLIM, it is approximately $\mathbf{I} - P_0 \mathbf{S}$, where $\mathbf{I}$ is the identity matrix and P_0 is a deterministic projection. Different methods have been used to analyze these information operators, especially in the key step on their invertibility as random matrices in the parameter spaces characterized by the respective identifiability conditions.

ADDITIONAL REFERENCES

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*, 10, 2022–2029.

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.

Huang, J., Kuo, H.-C., Koroleva, I., Zhang, C.-H., and Soares, M. B. (2003), "A Semi-Linear Model for Normalization and Analysis of cDNA Microarray Data," Technical Report 321, University of Iowa, Dept. of Statistics.

Kerr, M. K., and Churchill, G. A. (2001), "Experimental Design for Gene Expression Microarrays," *Biostatistics*, 2, 183–201.

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.

Ma, S. G., Kosorok, M. R., Huang, J., Xie, H. H., Manzella, L., and Soares, M. B. (2005), "Robust Semiparametric Microarray Normalization and Significance Analysis," preprint, University of Wisconsin, Dept. of Biostatistics and Medical Informatics.

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.

Wang, D. L., Huang, J., Xie, H. H., Manzella, L., and Soares, M. B. (2005), "A Robust Two-Way Semilinear Model for Normalization of cDNA Microarray Data," *BMC Bioinformatics* 2005, 6, 14.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002), "Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation," *Nucleic Acids Research*, 30, e15.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001), "Normalization for cDNA Microarray Data," in *Microarrays: Optical Technologies and Informatics*, eds. M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty, San Jose, CA: SPIE, Society of Optical Engineering, pp. 141–152.

Comment

Heping ZHANG

I congratulate the authors for this excellent and timely work. Publications in microarray data analysis to date have concentrated on methodological issues, including the critical step of normalization as the authors have addressed in this article, among many others. There are few articles that deal with the theoretical aspect of the statistical approaches, making the present article a welcome and novel contribution to the field. Similar models to the authors' model (1),

$$Y_{gi} = \alpha_g + \beta_{\mathbf{r}_{gi}} + \gamma_{\mathbf{c}_{gi}} + m(X_{gi}) + \varepsilon_{gi},$$

have been proposed and used in the microarray data analysis. Here α_g is the treatment effect associated with the gth gene, $\mathbf{r}_{gi}$ and $\mathbf{c}_{gi}$ are the row and column of print-tip block where the gth gene of the ith replication resides, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the row and column effects, $m(\cdot)$ is a smoothing function of $\mathbf{X}$ representing the intensity effect, and ε is the random error.

What makes this article unique and important is the circumstance under which the theory is established. That is, there are many more parameters of interest than the number of samples in terms of the arrays. The authors have successfully built the asymptotic theory based on the classic work of Neymann and Scott (1948). I wish to discuss a few practical and conceptual issues related to the forgoing model and the theory.

First, the authors point out that the α's are the nuisance parameters during the phase of removing the print-tip and intensity effects. This is insignificant from a theoretical standpoint. But because the α's are of ultimate scientific interest for most of the microarray studies, it is critical for statisticians to appropriately communicate the properties of the α estimators with scientists. The authors cite the work of Huang and Zhang (2003), who examined the asymptotic theory when the replication number I tends to infinity. The question is: whether the α estimators are still inconsistent when I is larger and the number of genes is even greater. If the α estimators are inconsistent, does this mean that the normalization leads to a biased correction of the data? If so, then it would be really important to understand the magnitude of the bias. For example, when we compare two groups of samples through the α estimators, can we design microarray experiments to minimize the bias? These are not necessarily questions for the authors, but I would certainly welcome their insights.

My second comment relates to the smoothing function $m(\mathbf{X})$ that represents the intensity effect, where $\mathbf{X}$ is the average log-intensity of the red and green channels. I am more concerned with the consequence than with the necessity of this correction. Both the response variable $\mathbf{Y}$ in model (1) and $\mathbf{X}$ are direct transformations of the intensities of the red and green channels. Why would $m(\mathbf{X})$ not result in an overcorrection of the signal? If this type of correction is helpful, is the average log-intensity of the red and green channels an ideal way to define $\mathbf{X}$? It is certainly a natural choice, but the question is: Does it improve the α estimators? Unlike in standard smoothing, where the independent variable is a fixed variable, such as time, here the independent variable is a random variable composed of the same two variables as the dependent variable. Intuitively, the choice of the independent variable influences the properties of the estimates for all parameters in model (1).

My third comment is about the implementation in the parameter estimation for model (1) as it also relates to my own experience (Zhang 1997). Naturally, the authors propose beginning the estimation process with given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $m(\cdot)$, particularly by setting them to 0. There are two technical questions. First, from numerical experiments, the algorithm "converges" very fast, but in theory, an explanation would be great. Second, it is useful to know whether the starting point changes the final estimates; in other words, are there local optimizers in this particular problem?

Finally, I enjoyed the beautiful presentation for the data analysis, but I did notice something simpler. Figure 5(c) displays the estimated $m(\cdot)$. For practical purposes, the "curve" is pretty much a straight line. Would the authors recommend refitting the data after a nonparametric smoothing reveals a simpler model? If a linear relationship is indeed considered, then, returning to my second comment, model (1) would fit a difference of two variables against the sum of the same two variables. I am still uncertain as to whether one would really want to make this adjustment.

In conclusion, the authors should be congratulated for their masterpiece of work and for shedding light on the theoretical aspects of this overwhelmingly data-analytic–driven area. Obviously many open questions remain, and this important work will stimulate much more to come.

ADDITIONAL REFERENCE

Zhang, H. P. (1997), "Multivariate Adaptive Splines for Longitudinal Data," *Journal of Computational and Graphic Statistics*, 6, 74–91.

Heping Zhang is Professor of Biostatistics, Yale University School of Medicine, New Haven, CT 06520, and Visiting Professor, Jiangxi Normal University, Nanchang, China (E-mail: *Heping.Zhang@yale.edu*).

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000475

Comment

Michael R. KOSOROK and Shuangge MA

We congratulate Fan, Peng, and Huang (FPH hereafter) on their interesting, innovative, and important contribution on microarray normalization. In addition to proposing new methodology that addresses several important open problems, FPH also present new asymptotic theory that both validates their approach and provides insight into the normalization process. Such theoretical work has been sparse in the microarray literature, probably because of the nonstandard way in which the number of parameters is large relative to the number of observations. We appreciate the opportunity to comment on this article.

Normalization is the process of removing systematic background noises of gene expression measurements in microarray experiments. Typically, this is done slide-by-slide before conducting a significance analysis of individual gene effects. However, some authors advocate combining normalization and significance analysis to account for the variability of estimators used in the normalization process (Huang, Wang, and Zhang 2003; Huang and Zhang 2003). We revisit this issue later on in this comment. For now, we note that techniques for normalization and significance analysis share a number of similarities for both cDNA and oligonucleotide microarrays. Hence some of the concepts in FPH on cDNA arrays are also applicable to oligonucleotide arrays. However, there are sufficient structural differences between the two kinds of arrays that significant work is needed before these concepts can be applied to the oligonucleotide setting. Thus we restrict our comments primarily to cDNA arrays.

We first briefly outline the key contributions of FPH, then discuss a few useful extensions. We briefly discuss a connection to marginal asymptotics for controlling false discovery rates (FDR) in significance analysis, then give a few comments on computational issues before presenting our closing comments.

1. THE MAIN CONTRIBUTIONS

FPHs main methodological contribution is a nonparametric method of normalizing cDNA microarrays that takes into account intensity and print-tip block effects without limiting the proportion of up-regulated or down-regulated genes. This represents a significant improvement in flexibility over the methods of both Dudoit et al. (2002) and Tseng, Oh, Rohlin, Liao, and Wong (2001). Moreover, the new methods permit the variability of the residual error to depend nonparametrically on the log-intensity of the red and green channels. The fact that estimation is applied in-slide makes it possible to obtain slide-specific gene effects. The proven partial consistency of the procedure makes it possible to proceed directly to significance analysis, after normalization, without requiring adjustments for uncertainty in the in-slide parameter estimates. In addition, the flexibility of the underlying model allows one to combine information across arrays when such combining may be helpful for estimating certain parameters.

The essential theoretical contribution is that, under realistic assumptions, the proposed estimation procedure works when the number of replicated genes G goes to infinity, even though the number of replicates I of replicated genes is fixed. This occurs in the presence of both location and scale nonparametric intensity effects. This is in contrast with the asymptotics of Huang and Zhang (2003), in which the number of replicates I is assumed to go to infinity. Moreover, whereas the total number of print-tip block effects d is assumed fixed in FPH, a different effect is permitted for each gene, resulting in the total number of parameters going to infinity. Thus FPHs asymptotic theory is somewhat nonstandard and represents a significant contribution to the scarce research on large-sample theory for microarrays (see also van der Laan and Bryan 2001; Huang and Zhang 2003). We discuss asymptotic issues in greater detail later.

Not only does the proposed methodology address an important practical problem in a useful way, but the validity of the procedure is verified both specifically and in general. The specific level of verification is accomplished in two ways. First, the usefulness of the method is clearly demonstrated in the neuroblastoma cell example. Second, the validity of the procedure is verified under several important—but very specific—settings in the simulations studies. The general level of verification is accomplished through a careful large-sample theoretical analysis that validates the procedure for countless realistic settings beyond the specified numerical scenarios. We again congratulate the authors on this important and well-rounded contribution.

2. SOME USEFUL EXTENSIONS

We now point out a few possibly useful extensions of FPHs results. The first extension is a modification of the model to address array-specific "warping" effects; the remaining extensions involve asymptotic issues.

2.1 Array-Specific Warping

One concern with proceeding directly to significance analysis after the proposed normalization is that there may be array-specific warping of the gene effects. One way to model this is with array-specific monotone transformations that apply to all genes after normalization. This is essentially what happens with quantile normalization for oligonucleotide arrays (Bolstad, Irizarry, Åstrand, and Speed 2003). Unfortunately, it is unclear how to apply this meaningfully to the present context. Perhaps

Michael R. Kosorok is Professor, Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, K6/428 Clinical Science Center, Madison, WI 53792 (E-mail: *kosorok@biostat.wisc.edu*). His research is supported in part by grant CA75142 from the National Cancer Institute. Shuangge Ma is Postdoctoral Fellow, Department of Biostatistics, University of Washington, Seattle, WA 98115 (E-mail: *shuangge@u.washington.edu*).

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000466

a more fruitful approach is to use array-specific location-scale transformations. Such transformations are easy to interpret, because the array-specific additive models proposed by FPH are invariant to whether the transformation is applied before or after adding the residual errors.

To fix ideas, apply FPHs model (1) to $j = 1, \ldots, J$ arrays to obtain

$$Y_{gij} = \alpha_{gj} + \beta_{r_{gij}j} + \gamma_{c_{gij}j} + m_j(X_{gij}) + \epsilon_{gij}, \tag{1}$$

where we also assume that

$$\alpha_{gij} = \tau_j(\nu_{gj} + \eta_j), \tag{2}$$

under the constraints $\sum_{j=1}^{J} \eta_j = 0$, $J^{-1}\sum_{j=1}^{J} \tau_j = 1$, $\tau_j > 0$, for $1 \le j \le J$, and the sums of the row, column, and intensity effects are 0 for each array. In (2), ν_{gj} is the unwarped gene effect for array j, whereas τ_j and η_j are the array-specific multiplicative and additive warps. Following FPH, we assume that ϵ_{gij} is random error with mean 0 and standard deviation $\sigma_j(X_{gij})$. Note that model (1)–(2) is essentially equivalent to

$$Y_{gij} = \tau_j\big(\eta_j + \nu_{gj} + \beta_{r_{gij}j} + \gamma_{c_{gij}j} + m_j(X_{gij}) + \epsilon_{gij}\big), \tag{3}$$

under the same constraints.

This modified model allows each array to have its own location and scale adjustment without otherwise modifying FPHs model. One key difference between (3) here and FPHs (1) is that an additional phase of estimation is required before obtaining the normalized gene effects,

$$\nu^*_{gj} = Y_{gj}/\hat{\tau}_j - \hat{\eta}_j - \hat{\beta}_{r_{gij}} - \hat{\gamma}_{c_{gij}} - \hat{m}_j(X_{gj})$$

(for the genes without replication). But this does not appear to add substantially to the computationally difficulty. Moreover, it also appears that the essential asymptotic result—that the normalized gene effects are partially consistent—will still hold.

2.2 Asymptotic Extensions

In the current setup, FPH requires that the number of slides J be fixed. This is reasonable if exact methods such as permutation tests are to be used for significance analysis. For the neuroblastoma cell example, the normalized gene effects appear to be Gaussian, and thus slightly modified t-tests could be used instead of permutation tests. In some cases, as we discuss later, it may be worthwhile to use marginal asymptotics for significance analysis. For such asymptotics, we need to allow J to go to infinity slowly with G. This would require some modification of FPHs asymptotic theory, but it seems feasible that at least uniform consistency of all fixed effects for all arrays can be obtained so that partial consistency will still follow.

It is also worth pointing out that there are restrictions on how many genes can be printed in one block. Thus G/d probably should increase slowly with G. Hence the structure of the slides will probably force $d = r + c$ to go to infinity slowly. This will require a significant modification of FPHs asymptotic theory, because the dimension of β will now be increasing with G (albeit slowly); but it should still be possible to establish uniform consistency of $\hat{\beta}$ despite this. Provided that this uniform consistency of the "fixed" effects occurs at a reasonable rate, valid significance analysis of the normalized gene effects can still be done.

3. MARGINAL ASYMPTOTICS

The main goal in significance analysis is to reliably determine which genes are significantly up- or down-regulated. Typically, this is achieved through computing marginal p values with procedures that hopefully control the FDR, such as that proposed by Benjamini and Hochberg (1995) or the more refined q value approach of Storey (2002). A somewhat related goal is consistent estimation of the mean gene effects. The first goal appears to be feasible for fixed number J of arrays if permutation tests or other exact test procedures are used. But requiring that J be fixed restricts the range of test procedures that can be used and also makes consistent estimation of the mean gene effects impossible.

As mentioned in the previous section, it appears possible to allow J to increase to infinity slowly with G and still obtain normalized gene effects Y^*_{gj} that are uniformly "consistent" for the randomly perturbed array-specific gene effects $\rho_{gj} = \nu_{gj} + \epsilon_{gj}$, where ν_{gj} and ϵ_{gj} are as defined in Section 2.1. It is now reasonable to assume that ρ_{gj} are iid across arrays ($1 \le j \le J$) for each gene $1 \le g \le G$. Thus for any reasonable significance analyses, we can ignore the variability of the estimators used in the normalization phase. Thus, without loss of practical generality, we hereafter assume that the ν_{gj}'s are observed directly. Let F_g be the marginal distribution function of ν_{gj}, and assume that the F_g's are continuous so that there are no ties in the data.

An alternative to permutation tests for testing the null hypotheses $\{H_{0g}: F_g \text{ is symmetric about } 0\}$, $1 \le g \le G$, is the marginal signed-rank test. Let T_g be the marginal rank test for gene g based on J arrays; that is, it is the sum of the ranks of $|\rho_{g1}|, \ldots, |\rho_{gJ}|$ that correspond to the positive ρ_{gj}'s. Now define

$$V_g = \frac{T_g - (J^2 + J)/4}{\sqrt{(3J^3 + 2J^2 + J)/24}}.$$

For each value of J, the exact distribution of V_g is known under the null hypothesis, and thus exact p values for each gene can be obtained. But we can simplify this process by using the well-known result that V_g is asymptotically normal under H_{0g}, as we now argue. Let Φ_J be the exact cumulative distribution function for V_g under H_{0g}, and let Φ be the standard normal cumulative distribution function. It is easy to verify that Φ_J converges uniformly to Φ. Thus, if we let $K \subset \{1, \ldots, G\}$ denote the genes for which H_{0g} holds, then we can now show that there exists a random vector $U_1, \ldots, U_G$ such that U_g is (marginally) uniformly distributed for all $g \in K$ and $\sup_{g \in K} |\Phi_J(V_g) - U_g|$ goes to 0 in probability, as $J \to \infty$, whether or not G also goes to infinity. Under reasonable conditions, $\Phi_J(V_g)$ should be asymptotically nonuniform for all $g \notin K$. This now enables application of the q value methodology under fairly general dependence structures in the gene effects (see thm. 5 of Storey, Tayler, and Siegmund 2004).

It is worth exploring more general inference procedures in this context, including, for example, median-based procedures that are robust to large and asymmetric errors. Uniformly consistent estimation of parameters of F_g is also of interest and also appears to be feasible, as was demonstrated by van der Laan and Bryan (2001), although their assumption that F_g has uniformly bounded support is somewhat unrealistic. We believe that "mar-

ginal asymptotic" issues such as these are potentially very important and should receive more attention in the future. In this context, partial consistency will probably play a very important role, because it effectively enables statisticians to ignore the uncertainty in the normalization process when conducting inference on the marginal gene effects.

4. COMPUTATIONAL ISSUES

As FPH point out, one key advantage of the proposed approach is its computational simplicity. The estimation can be achieved with simple iterations shown in FPHs sections 2 and 3. Moreover, the proposed approach estimates the block effects and the intensity effects with replicated genes within one slide only. Because the number of genes with replicates tends to be small, the proposed approach further simplifies the computations. One missing step in the proposed procedure is the process for selecting the bandwidth h. For density estimation, we have found that twice the interquartile range times $n^{-1/5}$ works quite well (Kosorok 1999), and a similar bandwidth based on the interquartile range of the intensities X_{gi} would probably work in the present setup.

Another issue that we would like to raise is that the comments on the computational cost of the approach of Huang et al. (2003) are not completely fair. Although, in the Huang et al. article the least squares estimators are expressed in a matrix form, the solutions can be obtained by iterations similar to those described by FPH. The gene effects can be estimated one by one according to the special structure of the least squares estimators. In contrast, estimation of the block effects and the nonparametric normalization curves (which are supposed to be splines) do involve inverting matrices. However, the dimensions of the block effects (d of FPH) and the normalization curves (low-dimensional subspaces of Sobolev spaces) are usually much lower than the dimension of the gene effects. It is therefore expected that the computational cost of the Huang et al. (2003) approach will be close to that of the proposed approach.

5. CLOSING WORDS

Overall, FPHs article provides a significant step forward in the statistical analysis of microarrays. Athough semiparametric models have previously been introduced into microarray analyses (see, e.g., Newton, Noueiry, Sarkar, and Ahlquist 2004), rigorous frequentist theory for such approaches has been slow in developing. FPHs asymptotic theory is an important step forward in this direction. We also anticipate that the role of partial consistency for the normalization step will become increasingly important because of the clean manner in which the uncertainty in the normalization process is separated from the significance analysis phase, at least asymptotically. We also believe that marginal asymptotics, in combination with partial consistency, has a potentially crucial role to play in future formal justifications of microarray methodology. As mentioned earlier, formal theoretical justification is critical for ensuring the validity of statistical procedures in broad generality. We again acknowledge the importance of the contributions of this article—both theoretically and methodologically—and are grateful to have had the opportunity to comment on it.

ADDITIONAL REFERENCES

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003), "A Comparison of Normalization Methods for High-Density Oligonucleotide Array Data Based on Variance and Bias," *Bioinformatics*, 19, 185–193.

Kosorok, M. R. (1999), "Two-Sample Quantile Tests Under General Conditions," *Biometrika*, 86, 909–921.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting Differential Gene Expression With a Semiparametric Hierarchical Mixture Method," *Biostatistics*, 5, 155–176.

Storey, J. D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

Storey, J. D., Tayler, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation and Simultaneous Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Ser. B, 66, 187–205.

van der Laan, M. J., and Bryan, J. (2001), "Gene Expression Analysis With the Parametric Bootstrap," *Biostatistics*, 2, 445–461.

Comment

Robert TIBSHIRANI

Fan, Peng, and Huang present an interesting method for normalizing data from cDNA microarrays. The method uses semiparametric models and leads to some challenging mathematical questions concerning their asymptotic performance.

Focusing on the practical aspects, their method seems novel in its adjustment of print-tip effects and its use of replication to borrow strength. The latter seems to crucial to the technique; they need either within-array replicates or replications of the entire array. In the latter case, they need to assume that the "treatment effect" of each genes is the same across different arrays. I wonder about the robustness of such an assumption.

To play devil's advocate, I find plots like the Q–Q-plot in figure 5 to be unconvincing. The more that we model data, the more the residuals tend to look normal. But has our model allowed us to extract more reliable biological information from the data? To answer this question, the method should be applied to data from spiking experiments, in which known quantities of known RNAs are hybrized to an array and the measurement

Robert Tibshirani is Professor, Departments of Health Research and Policy, and Statistics, Stanford University, Stanford, CA 94305 (E-mail: *tibs@stat.stanford.edu*).

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000484

signal can be compared with the true (expected) signal. Moreover, with real data, one can plot the number of genes called significant versus the expected number of false-positive genes, as the cutoff for a given method is varied. If a method A (such as that proposed by the authors) is better than a more standard method B, then the curve for A should lie mostly above that for B. Such an analysis, for example, was done in the "SAM" work of Tusher, Tibshirani, and Chu (2001).

It would be important to carry out these kinds of studies to determine the practical utility of the proposed methods.

ADDITIONAL REFERENCE

Tusher, V., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Science*, 98, 5116–5121.

Rejoinder

Jianqing FAN, Heng PENG, Tao HUANG, and Yi REN

We thank the editor Francisco Samaniego and an associate editor for organizing this stimulating discussion, with a conscientious effort to invite outstanding researchers from diverse backgrounds that make the discussion more thought-provoking. We are also very grateful to all discussants for their insightful and stimulating comments, touching on practical, methodological, and theoretical aspects of microarray designs, experiments, normalization, analysis, and applications, offering some original insights and outlooks. Their contributions are very timely and helpful.

The last couple of years have brought an explosion of statistical techniques for the design and analysis of microarray data. They range from the design of microarray experiments (Kerr and Churchill 2001; Yang and Speed 2002), normalization of microarray data (Tseng, Oh, Rohlin, Liao, and Wong 2001; Dudoit et al. 2002; Fan, Tam, Vande Woude, and Ren 2004; Huang, Wang, and Zhang 2003), the expression indices of Affymetrix oligonucleotide arrays (Li and Wong 2001; Irizarry et al. 2003a), significant analysis of gene expressions (Tseng et al. 2001; Tusher, Tibshirani, and Chu 2001; Lönnstedt and Speed 2002; Fan et al. 2004), classification and clustering (Tibshirani, Hastie, Narasimhan, and Chu 2003; Zhang, Yu, and Singer 2003), and time-course experiments for the expression pathways (Svrakic, Nesic, Dasu, Herndon, and Perez-Polo 2003), among others. (For an overview on the subject, see Sebastiani, Gussoni, Kohane, and Ramoni 2003; Speed 2003; Parmigiani, Garrett, Irizarry, and Zeger 2003.) They revived a surge interest in multiple testing problems (Dudoit, Shaffer, and Boldrick 2003; Storey 2003; Donoho and Jin 2004; Storey, Taylor, and Siegmund 2004; Efron 2004). They exemplify the interactions between statistics and the sciences, tackling problems of high societal impact. All of the discussants call for more statistical understanding of various procedures in use. We agree wholeheartedly with this and contribute the article under discussion in the hope that it will stimulate more statisticians to work on this area. The discipline of statistics should grow stronger when it provides methodologies that address issues of the highest societal importance while at the same time offering foundational understanding of the methodologies that push theory, methods, and applications forward.

1. REPLICATIONS OF cDNA GENES

Normalization is a critical step in removing possible systematic biases in the process of microarray experiments. The process is usually complicated, and the biases are hard to quantify. The ideal situation for assessing the systematic biases is to use within-array replications; all experimental conditions are the same except for the locations of replicated genes. Hence the observed differences of expression for two identical clones in the same array are due to random noises and possible biases. The genesis of our approach is to extract the biases from those duplicated pairs of genes.

We are very grateful to Professor Craig for his careful description on the process of fabrication of slides and to Professor Sabatti for her convincing arguments on the needs of within-slide replications. Sabatti is correct that detected distortion (biases) in cDNA microarrays should be taken more seriously. Greater understanding of the basis of biases should facilitate technological improvements. The degree of distortion can be better understood when two identical tissues are compared using the cDNA microarray experiments. We discuss this issue further in Section 3 of this rejoinder. Sabatti also raised the question of how many replicates are needed at the stage of designing cDNA microarrays. The answer depends on the complexity of models that statisticians would like to use and on the expense of appropriately replicating some of the genes. She is right that a natural model that takes care of print-tip effects would include one extra parameter per print tip due to the technological process used in spotting microarrays. Our asymptotic theory continues to apply, and our asymptotic formulas provide useful guidance for choosing the number of replicated genes. Because the number of print tips is usually large, aggregating information from other arrays is needed and makes it possible

Jianqing Fan is Professor (E-mail: *jqfan@princeton.edu*) and Heng Peng is Postdoctoral Fellow (E-mail: *pheng@princeton.edu*), Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Tao Huang is Postdoctoral Associate, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06511 (E-mail: *t.huang@yale.edu*). Yi Ren is Research Assistant Professor, Department of Surgery, Hong Kong University, Hong Kong, China. This work was supported in part by National Science Foundation grant DMS-03-54223.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214505000000493

to obtain reasonable estimates of the print-tip effect. Validation tests in the next section should also be useful for checking whether systematic biases have been successfully removed. Statistical techniques should also be calibrated with biological experiments to achieve better approximations and understanding.

Craig is right that duplicated spots are very helpful for normalizing expressions of multiple arrays and assessing the effectiveness of a normalization procedure. He expresses some concerns about the costs of duplication unless the benefits outweigh the costs. We agree with such a careful attitude but are far more optimistic on the feasibility of within-array duplications.

First, printing a couple hundred duplicated spots in an array of 20,000 spots does not take up a large percentage of space. For a wide class of biological and biomedical problems, many genes are of little biological interest. Replacing them with duplicated spots enables biologists and statisticians to reduce biases in multiple-array comparisons and to verify certain biological claims. Second, the cost of printing duplicated spots should not be excessive. Once a template is designed, cDNA microarrays can be manufactured, and the same microarrays can be applied to a wide range of biomedical problems. Third, although Craig rightly points out that well-by-well duplication is time-consuming and that choosing the duplicated genes with a wide range of intensity requires some effort, we are far more optimistic. With so many studies using cDNA microarrays, we have already had a good idea of the relative intensity levels of different genes. This issue is eased further if a parametric model is used for modeling the intensity effect. Further, based on our own experience and communication with others, we surmise that randomly placing duplicated genes is feasible. The arrays that we used were designed in 2001 and already contained 111 "randomly" printed duplications. We can easily imagine that this should be easier today, thanks to advances in biotechnology. Finally, with the increased popularity of customized micorarray chips, which print only a couple hundred relevant genes for a specific biomedical problem to increase their specificity and sensitivity, there is plenty of room for duplicated spots. Sabatti gives very eloquent arguments that such customized arrays will benefit the most from our suggested technology.

Craig has made a useful suggestion: Understand the benefit of the SLIM normalization via simulations. We followed his suggestion and compared our SLIM with the loess normalization at each block. That is, we simulated data from model (3) of our article (FPH hereafter), and took the rest of parameters from example 2. To show the biases of the loess normalization method, we considered the treatment effects $\boldsymbol{\alpha}_n$ on 100 genes as realizations from the asymmetric exponential distribution with density

$$f(x) = .7\exp(-x)I(x \geq 0) + .3\exp(x)I(x < 0),$$

and the rest as 0. Table 1 reports the mean squared errors for estimating the intensity and treatment effects. Because there are only 100 genes with an asymmetric treatment effect, the model violation of the loess normalization becomes less severe when G is large. Hence when $G = 200$, the performance of the loess normalization is quite poor, and the performance of the loess normalization gets closer to the SLIM when G is large. This demonstrates the genuine need for the SLIM type of technique even when the model is slightly violated.

Table 1. MSEs for SLIM and Loess Normalization, n = 200

		SLIM normalization			Loess normalization		
	I	$G = 200$	$G = 400$	$G = 800$	$G = 200$	$G = 400$	$G = 800$
m	2	.0190	.0095	.0049	.0784	.0227	.0063
	4	.0079	.0039	.0020	.0584	.0168	.0048
α	2	.1028	.0974	.0955	.1494	.1070	.0958
	4	.0502	.0484	.0474	.0914	.0588	.0492

2. VALIDATION TEST

Craig raised the question of whether the duplicated spots should always be used. Our view is that one should always use them when available. They contain the most valuable information about possible systematic biases in microarray experiments. In addition, duplicated spots allow us to verify the validity of scientific conclusions; if many duplicated spots have very different results, then the validity of the analysis and conclusions becomes questionable. Thus duplicated genes should not be limited to the housekeeping genes. Furthermore, the duplicated spots allow statisticians to validate statistical models; for example, they can be used to test whether the loess normalization has effectively removed the intensity effect and whether the aggregation method is applicable to a specific problem. If the systematic biases have been removed by a normalization approach, then the normalized log-ratios should approximately follow the model

$$Y_{gi} = \alpha_g + \varepsilon_{gi}, \qquad i = 1, 2,\ g = 1, \ldots, G,$$

for replicated genes. Hence the difference $Y_{g2} - Y_{g1}$ should have mean 0 or be symmetrically distributed around 0. One can apply a χ^2-test or a sign test to this kind of problem.

It is unfortunate that many cDNA microarrays are designed without repeated genes and without careful consultation with statisticians. One must create the synthetic duplications or aggregate information from other arrays. These impose extra statistical assumptions that we now discuss.

3. AGGREGATION AND NORMALIZATION

Aggregation is a powerful tool that enables us to pool information from multiple arrays. Through statistical modeling, we can extract information from other arrays. This allows more flexible models and/or better estimates. Using the aggregated SLIM (12) from FPH, and with more careful use of available data, the estimation of the intensity and block effects can be further improved.

To appreciate this, let us recall that nonreplicated genes contributed the data with the following model:

$$Y_{gj} = \alpha_g + \beta_{j,r_g} + \gamma_{j,c_g} + m_j(X_{gj}) + \varepsilon_{gj}, \qquad j = 1, \ldots, J. \quad (1)$$

Let N be the number of nonduplicated genes. Unlike the case without aggregation ($J = 1$), model (1) contains a tremendous amount of information about the unknown parameters, because N is very large. From theorem 5, it is known that $\beta_{j,r}$ and $\gamma_{j,c}$ can be estimated at root-n consistency from aggregated SLIM (12), where $n = IG$ is the number of replicated spots. Substituting this into (1), we have

$$Y_{gj} - \hat{\beta}_{j,r_g} - \hat{\gamma}_{j,c_g} = \alpha_g + m_j(X_{gj}) + \varepsilon_{gj}. \quad (2)$$

If the β's and γ's were known, then model (2) allows us to estimate m_j at rate $O(N^{-2/5})$, which is much faster than $n^{-2/5}$. Due to the errors in estimating the β's and γ's, the accuracy for estimating the intensity effect m_j is $O(n^{-1/2} + N^{-2/5})$. A more elegant and more efficient method is to directly modify the algorithm in section 3 of FPH to accommodate all available data. The theoretical results of such a method have not been investigated. Because the number of nonreplicated data N is far larger than n, we would expect this modification to increase the accuracy by an order of magnitude. Even in a situation where aggregation is permissible, within-slide replications are still important, because they can also be used to validate the effectiveness of normalization (see sec. 2).

Aggregation does not come without a cost. We have to assume that the gene effects are the same across the arrays. In other words, in model (12) of FPH and (1), α_g does not depend on j. When there are individual variations, such as different arrays representing genetic materials from different subjects, aggregation can possibly create biases. The robustness of aggregration is discussed in Section 5 here. In contrast, the within-array replications can still be used to normalize multiple arrays even in such a situation.

We are grateful to Professors Huang and Zhang for making connections among SLIM, aggregated SLIM, and TW–SLM, and welcome their efforts in expanding these models to accommodate various designs of microarrays. We agree with them that these three models share some common characteristics and have their own distinct personalities. Our technical results complement each other and enrich the theory of normalization.

In their discussion, Huang and Zhang suggested incorporating a weighting scheme using the standard deviation associated with the measurements in the pixels within a spot. This proposal is welcome, and it will behave better when the raw standard deviations are smoothed against the log-intensity to increase the accuracy of the weights and reduce the fluctuations of the raw weights. The resulting weights can be also combined with the weights estimated from the log-ratios conditioning on the log-intensities, as illustrated by Fan et al. (2004). We also welcome the efforts of Huang, Zhang, Korosok, and Ma for introducing array-specific location and scale parameters that take into account individual array variations.

4. THEORETICAL ISSUES

The essential ingredient of our model is its estimability, which relies heavily on the model structure. In general, when the number of nuisance parameters is proportional to sample size, even low-dimensional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ cannot be consistently estimated. We have exploited the orthogonality of the model, which turns the "curse of dimensionality" into a "blessing." In the notation (4) of FPH, the matrices $\mathbf{B}_n$ and $\mathbf{Z}_n$ are nearly orthogonal. This allows us to carry out the asymptotic theory even with the number of arrays fixed. This is in contrast with the asymptotic theory of Huang et al. (2003), in which $J \to \infty$ and $G \to \infty$ are required. The implication is that the number of arrays must be large, which is not very elegant for microarray applications, although it is hoped that the method works even when I is not too small.

To appreciate why parameters can be estimated consistently without assuming $J \to \infty$, the following argument, due to Fan, Chen, Chan, Tam, and Ren (2005), provides insights into the high-dimensional semilinear model. Suppose that we have two arrays ($J = 2$), and that the log-ratios for each array follow the model

$$Y_{gj} = \alpha_g + \boldsymbol{\beta}_j^T \mathbf{U}_{gj} + m_j(X_{gj}) + \varepsilon_{gj}, \tag{3}$$

where $\varepsilon_{gj} \sim \mathrm{N}(0, \sigma^2(X_{gj}))$, independent of the log-intensity X_{gj} and covariate $\mathbf{U}_{gj}$. This is a slight extension of model (11) of FPH and an important specific case of TW–SLM. Here $\boldsymbol{\beta}_j$ can be considered as some array-specific effect with suitable identifiability conditions. This model has similar theoretical components as model (1). Let $Y_g^* = Y_{g2} - Y_{g1}$ and $\varepsilon_g^* = \varepsilon_{g2} - \varepsilon_{g1}$. Then, by taking the difference of (3), we have

$$Y_g^* = \boldsymbol{\beta}_2^T \mathbf{U}_{g2} - \boldsymbol{\beta}_1^T \mathbf{U}_{g1} + m_2(X_{g2}) - m_1(X_{g1}) + \varepsilon_g^*. \tag{4}$$

This is now a usual semiparametric additive model. Fan, Härdle, and Mammen (1998) have shown that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be estimated at a root-G rate and that m_1 and m_2 can be estimated at a nonparametric rate. They have further demonstrated some oracle properties of $\hat{m}_1$ and $\hat{m}_2$. In summary, with a special structure, low-dimensional parameters in a high-dimensional semilinear model are estimable. Huang et al. (2003) focused on more general structures of high-dimensional nuisance parameters and hence needed to assume that $J \to \infty$ to estimate parameters consistently.

We thank Professor Zhang for raising a number of challenging questions and Professors Korosok and Ma for their comments on asymptotic extensions. All raise the issues on the asymptotics that the number of arrays J grows to infinity slowly with G. In this case, all parameters can be estimated more accurately. The row and column effects can still be estimated with root-n rate, and the intensity effect can be estimated at rate $n^{-2/5}$. The treatment effect on genes can be estimated with rate $J^{-1/2}$ for model (12) in FPH. Usually $n = IG$ is much larger than J. Hence the estimation errors for the intensity and block effects are asymptotically an order of magnitude faster than $J^{-1/2}$ and are negligible when J grows slowly. Consequently, the treatment effects $\boldsymbol{\alpha}_n$ can be estimated at rate $J^{-1/2}$, because they are the average of J arrays. But the block effect $\boldsymbol{\beta}$ does not improve much, and neither does the intensity effect. To examine the impact of the number of arrays on the estimation of parameters, we have repeated the simulation experiments in Example 1 of FPH with the different number of arrays J (with I taken as 2). The MSEs for estimating the intensity, block, and treatment effects are summarized in Table 2. Figure 1 gives the graphic presentation of the results. These results are clearly in line with our asymptotic theory. In particular, the slopes for the treatment effects are nearly 1.

Table 2. MSEs for Aggregated Estimator for Different Number of Arrays J (I = 2, n = 240/J)

	G	J = 1	J = 2	J = 4	J = 8	J = 16
m	200	.0774	.0498	.0407	.0355	.0322
	400	.0362	.0232	.0227	.0193	.0201
β	200	.0333	.0203	.0178	.0158	.0156
	400	.0152	.0098	.0082	.0081	.0075
α	200	.5797	.2776	.1368	.0686	.0331
	400	.5388	.2617	.1336	.0670	.0350

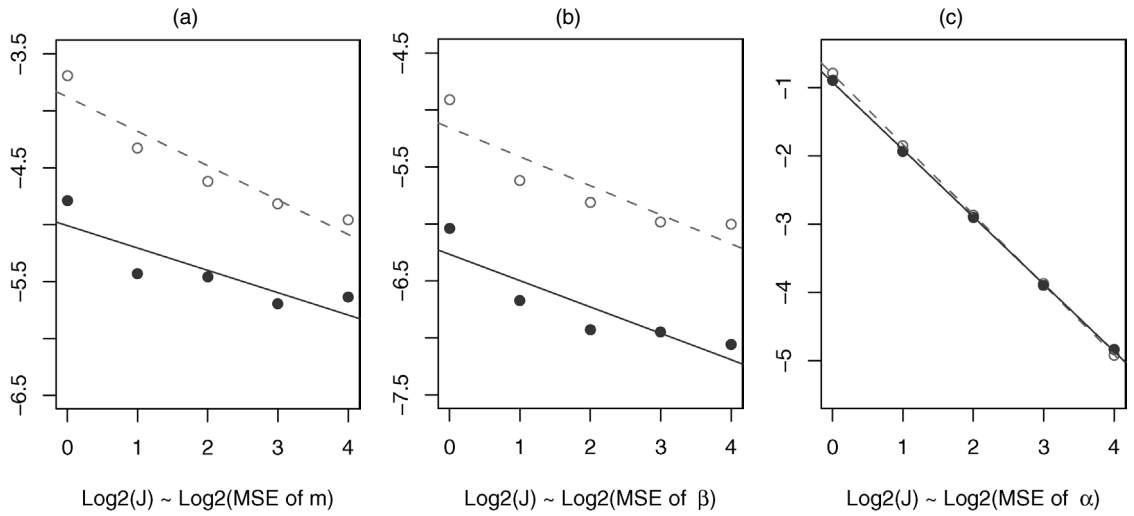

Figure 1. The Plot $\log_2(J)$ Against $\log_2(MSE)$: (a) Intensity Effect, (b) Block Effect, and (c) Treatment Effect. $G = 200$, solid lines; $G = 400$, dashed lines.

The parameters α_g, although becoming a nuisance for normalization purposes, are the key parameters of scientific interest. Their consistency requires that the number of arrays J tends to infinity. Fortunately, the microarray techniques are often used for the preliminary selection of significantly differently expressed genes. Hence the scientific question becomes whether the parameters α_g are statistically significantly different from 0 at certain significant levels (say .1%). This is an easier statistical problem than consistently estimating parameter values themselves.

Zhang raised an excellent point that log-ratios ($Y = \log R/G$) and log-intensities [$X = .5\log(RG)$] are both created from the intensity outputs of "green" (G) and "red" (R) channels, which might create some kind of spurious relation. In other words, even in the absence of the block effects and treatments [namely, $Y \sim \mathrm{N}(0, \sigma^2)$], there might be a spurious relationship,

$$Y = m(X) + \varepsilon, \tag{5}$$

for a nonzero function m. This concern is quite relevant. It can be eased if we assume that R and G are independently lognormally distributed. In this case, X and Y are independent, and the spurious relation m must be 0. In addition, if $\log(R)$ and $\log(G)$ are uncorrelated, then so are X and Y. Zhang's question can also be biologically tested. If the treatment and control arrays contain the identical biological materials, then there is no treatment effect. One can then smooth the (X, Y) pairs to see whether $m(\cdot)$ is statistically significant from 0.

Zhang, Kosorok, and Ma raise some computational issues associated with our methods. The speed of convergence depends on the implementation of a statistical estimate. Unlike optimizations in numerical analysis, the convergence in statistical computation has a somewhat different meaning. Because parameters are estimated with errors, optimization can be conducted somewhat crudely. This is why we use the word "implementation" instead of "algorithm." Our current implementation uses the local linear smoother, which is not a projection type of estimator. The algorithm will converge under some conditions on the smoothing matrix similar to those of Opsomer and Ruppert (1997). If the local linear estimator in the smoothing step in section 3 of FPH is replaced by smoothing splines, then the algorithm is truly a Gaussian–Sidel one, and it will converge under some mild conditions (Buja, Hastie, and Tibshirani 1989). Both algorithms should have about the same speed of convergence. But if the local linear fit is replaced by the polynomial splines, then the computation cost can be higher, depending on how it is implemented. The implementation of Huang et al. (2003) uses polynomial splines and can be more computationally intensive. It depends on whether they directly invert large matrices and whether they carefully exploit the sparsity of matrices created by B-splines (Eubank 1999).

5. INCORPORATING SIDE INFORMATION

Side information can be incorporated into the normalization and analysis of microarray data. Several discussants have touched on several aspects of these. For example, Zhang mentioned the possibility of using a parametric model to estimate the intensity effect, Huang and Zhang augmented the TW–SLM analysis using expressions from control genes, and Fan et al. (2004) advocated using duplicated genes. The side information should be incorporated into the design and analysis whenever possible.

The sparsity is vague but informative in the analysis of microarray data. For most studies, it is expected that only a fraction of the genes are significantly differently expressed across different tissues or samples; in other words, most of the α_g's are approximately 0, using the notation of (3). This information is vague but informative, because the number of genes such that $\alpha_g \approx 0$ is potentially large. This should be incorporated into the process of normalization and significant analysis of genes. We will pursue some problems in this direction in the future.

Aggregation can also be considered as a method of incorporating side information. Professor Tibshirani raised an excellent question on how robust such a method is. Suppose that the treatment effects on the genes vary somewhat across the arrays due to experimental conditions or individual variations. In other

Table 3. Robustness of Aggregrated Estimator; MSEs for $I = 2$, $n = 50$, and $J = 4$

	σ	$G = 100$	$G = 200$	$G = 400$	$G = 800$
m	0	.0219	.0102	.0049	.0027
	.05	.0205	.0106	.0052	.0027
	.1	.0199	.0102	.0052	.0027
	.2	.0224	.0111	.0051	.0025
β	0	.0073	.0034	.0017	.0008
	.05	.0066	.0032	.0016	.0008
	.1	.0066	.0033	.0015	.0008
	.2	.0071	.0034	.0017	.0008
α	0	.0302	.0256	.0247	.0241
	.05	.0279	.0257	.0247	.0241
	.1	.0312	.0270	.0251	.0242
	.2	.0449	.0333	.0281	.0255

words, model (1) becomes

$$Y_{gj} = \beta_{j,r_g} + \gamma_{j,c_g} + m_j(X_{gj}) + \alpha_g + \eta_{gj} + \varepsilon_{gj}, \quad j = 1, \ldots, J, \quad (6)$$

where η_{gj} are the (unobservable) individual variations on the treatment effect. If the individual variations η_{gj} behave like random noises, then the term can be absorbed into the noise term in (6). Hence the foregoing techniques continue to apply. In this sense, the aggregation is robust. To demonstrate this numerically, we consider the simulation example 2 of FPH and add the random noise $\eta_{gj} \sim \mathrm{N}(0, \sigma^2)$ to reflect individual variation on the genes. Note that the variance of the double exponential is 2, whereas the individual variations between two arrays are $\mathrm{N}(0, 2\sigma^2)$. In particular, when $\sigma = 0$, the results are taken from table 4 of FPH. Table 3 shows that the results are fairly stable.

6. EXPANDING SCOPES OF APPLICABILITY

We thank Professor Sabatti for outlining a few examples in which the curse of dimensionality can be turned into a "blessing." This is indeed a very interesting concept. The inverse type of model that she mentioned can possibly be consistently estimated with increased precision as G gets larger. The model is very important and warrants a more thorough investigation. Here we offer some heuristics to demonstrate that with some structures on p_{jt} or a_{ij}, it is potentially feasible to consistently estimate their values. For simplicity, we take $L = 1$ and $M = 2$. Then the observed data follow,

$$Y_{i1} = a_i p_1 + \varepsilon_{i1}, \qquad Y_{i2} = a_i p_2 + \varepsilon_{i2}. \quad (7)$$

For identifiability, let us assume that $p_2 = 1$. For each given p_1, by the least squares method, we have

$$\hat{a}_i = \frac{p_1 Y_{i1} + Y_{i2}}{p_1^2 + 1}. \quad (8)$$

Eliminating the nuisance parameters a_i by the profile least squares technique [i.e., substituting (8) into (7)], we obtain the synthetic nonlinear model parameterized by p_1,

$$Y_{i1} = \hat{a}_i p_1 + \varepsilon_{i1}, \qquad Y_{i2} = \hat{a}_i + \varepsilon_{i2}.$$

The least squares estimator of this synthetic nonlinear model is the solution to (by considering $\hat{a}_i$ as given)

$$\hat{p}_1 = \sum_{i=1}^{G} \hat{a}_i Y_{i1} \Big/ \sum_{i=1}^{G} \hat{a}_i^2. \quad (9)$$

This is indeed the iterative estimator of model (7), which iterates (8) and (9). Substituting (8) into (9) and using the central limit theorem, we can show heuristically that p_1 can be estimated at rate $G^{-1/2}$ under some regularity conditions.

Sabatti also discussed an example where the blessing of dimensionality cannot be materialized unless further constraints are imposed. Her example is the Li and Wong (2001) model for summarizing probe intensities in Affymetrix oligonucleotide arrays. We agree with her intuition.

We would like to mention that SLIM has also been applied to normalize Affymetrix arrays by synthetically creating intensities from "green" and "red" channels. By treating the average intensity of the control arrays as the outputs of the green channel and regarding the intensities from treatment arrays as the outputs of the red channel, Fan et al. (2005) created "synthetic" arrays and applied a refined SLIM technique to normalize the Affymetrix arrays. These authors demonstrated that among the 12 genes picked by their methods for biological confirmations, all are biologically confirmed. In addition, among those 12 genes known to be differently expressed biologically between the treatment and control samples, without using any normalization, 2 genes are not identified in the significance analysis, yielding a missed discovery rate of $2/12 \approx 17\%$. Using the Affymetrix MAS 5.0 software directly, 29% of genes are misdiscovered. The efficacy of SLIM has clearly been demonstrated.

7. SIGNIFICANCE ANALYSIS

We appreciate the comments of Kosorok and Ma on the significant analysis of genes using marginal asymptotics. Clearly, a nice idea has been outlined. An interesting component of their model is that the individual variations on the treatment effect are allowed, namely,

$$Y^*_{gj} = \alpha_{gj} + \varepsilon_{gj}, \quad (10)$$

where $\{Y^*_{gj}\}$ are the normalized log-ratios and α_{gj} is the treatment effect on gene g, which may depend on the subject j. The balanced permutation technique (Tusher et al. 2001; Fan et al. 2004) can be used to empirically determine the distribution of a test statistic V_g (including the one outlined by Kosorok and Ma) and estimate its associated false discovery rate. The method was further improved by Fan et al. (2005) with a fuller use of the sample. The basic idea is as follows. Restrict the genes to the set $\mathcal{G} = \{g : |V_g| \le 2\}$, for example. This set of genes is unlikely to have large sample mean $\alpha_{g\cdot}$. For each subset $\mathcal{J}$ in $\{1, \ldots, J\}$, multiplying the log-ratios Y^*_{gj} for arrays $j \in \mathcal{J}$ by -1 (i.e., swapping the "treatment" with "control"), we obtain a new set of arrays consisting of modified and unmodified ones. Compute the test statistics for the new set of arrays, resulting in $\{V^*_{g,\mathcal{J}} : g \in \mathcal{G}\}$ for each given $\mathcal{J}$. Pull together all of the foregoing test statistics $\{V^*_{g,\mathcal{J}}\}$ for all subset $\mathcal{J}$ and $g \in \mathcal{G}$, and use the empirical distribution of all such test statistics as an estimate of the null distribution of V_g. With this null distribution, one can empirically compute the p values for all test statistics $\{V_g, g = 1, \ldots, G\}$. This avoids computing extreme tail probabilities entirely based on mathematical assumptions and approximations.

It is worthwhile to briefly discuss the modified t-statistic used by Tusher et al. (2001) and Fan et al. (2004), which is given by

$$t_g = \frac{\bar{Y}_{g\cdot}}{SD_g/J + s_0}, \tag{11}$$

where $\bar{Y}_{g\cdot}$ and SD_g are the mean and the standard deviation of the normalized data in (10) for each given g. Here s_0 is a positive constant that plays dual roles. Because the number of genes is large and the number of arrays is usually small, by chance alone, some of SD_g's will be close to 0. The constant s_0 bounds the denominators away from 0 and prevents us from selecting too many spuriously significant genes. Moreover, when a gene is called significant, say $|t_g| > 3$, it requires the expression ratio

$$|\bar{Y}_{g\cdot}| \geq 3(SD_g/J + s_0) \geq 3s_0.$$

In other words, the expression ratio must be large enough to be called significant. This is also biologically meaningful, because biological functions require certain amounts of different expressions. Furthermore, when selected genes to be confirmed using a different biological technique, such as reverse-transcription potymerase chain reaction, the expressions need to be sufficiently different. In other words, the constant s_0 trades-off between the statistical significance and biological functionality. A large s_0 makes a gene less likely to be called "significant" by the t-statistic, but makes biological functionality and confirmation more feasible.

Tibshirani gave a good guideline for choosing a test. We agree with his basic principle. The number of significant genes is related to missed discovery rates. For two testing procedures with the same control of false discovery rate, the one that picks more significant genes tends to have less missed discovery rates. The former is related to the level of significance, and the latter is related to the power of a test. Missed discovery rates are important for scientific investigation. Once important genes fail to be discovered, the main goal of biological experiments can be defeated.

Tibshirani mentioned that the effectiveness of normalization method should also be compared with the true expected signals from spiking experiments. The validation tests outlined in Section 2 are also very useful for validating the effectiveness of normalization. Finally, we would like to mention that figure 5 in FHP merely shows that the conditional heteroscedasticity has been successfully removed and that the marginal data are normally distributed.

ADDITIONAL REFERENCES

Buja, A., Hastie, T. J., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models" (with discussion), *The Annals of Statistics*, 17, 453–555.

Donoho, D. L., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing* (2nd ed.), New York: Marcel Dekker.

Fan, J., Chen, Y., Chan, H. M., Tam, P. H., and Ren, Y. (2005), "Removing Intensity Effects and Identifying Significant Genes for Affymetrix Arrays in MIF-Suppressed Neuroblastoma Cells," unpublished manuscript.

Fan, J., Härdle, W., and Mammen, E. (1998), "Direct Estimation of Additive and Linear Components for High-Dimensional Data," *The Annals of Statistics*, 26, 943–971.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a), "Summaries of Affymetrix GeneChip Probe-Level Data," *Nucleic Acids Research*, 31, e15.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b), "Exploration, Normalization, and Summaries of High-Density Oligonucleotide Array Probe-Level Data," *Biostatistics*, 4, 249–264.

Kerr, M. K., and Churchill, G. A. (2001), "Experimental Design for Gene Expression Microarrays," *Biostatistics*, 2, 183–201.

Li, C., and Wong, W. H. (2001), "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection," *Proceedings of the National Academy of Science*, 98, 31–36.

Lönnstedt, I., and Speed, T. (2002), "Replicated Microarray Data," *Statistica Sinica*, 12, 31–46.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003), *The Analysis of Gene Expression Data*, New York: Springer-Verlag.

Sebastiani, P., Gussoni, E., Kohane, I. S., and Ramoni, M. F. (2003), "Statistical Challenges in Functional Genomics" (with discussion), *Statistical Science*, 18, 33–70.

Speed, T. P. (2003), *Statistical Analysis of Gene Expression Microarray Data*, New York: Chapman & Hall/CRC.

Storey, J. D. (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *The Annals of Statistics*, 23, 2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Ser. B, 66, 187–205.

Svrakic, N. M., Nesic, O., Dasu, M. R. K., Herndon, D., and Perez-Polo, J. R. (2003), "Statistical Approach to DNA Chip Analysis, *Recent Progress in Hormone Research*, 58, 75–93.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003), "Class Prediction by Nearest Shrunken Centroids, With Applications to DNA Microarrays," *Statistical Science*, 18, 104–117.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Science*, 98, 5116–5121.

Yang, Y. H., and Speed, T. (2002), "Design Issues for cDNA Microarray Experiments," *Nature Reviews Genetics*, 3, 579–588.

Zhang, H. P., Yu, C. Y., and Singer, B. (2003), "Cell and Tumor Classification Using Gene Expression Data: Construction of Forests," *Proceedigs of the National Academy of Science*, 100, 4168–4172.