40. A Two-Way Semilinear Model for Normalization and Analysis of Microarray Data

A proper normalization procedure ensures that the normalized intensity ratios provide meaningful measures of relative expression levels. We describe a two-way semilinear model (TW-SLM) for normalization and analysis of microarray data. This method does not make the usual assumptions underlying some of the existing methods. The TW-SLM also naturally incorporates uncertainty due to normalization into significance analysis of microarrays. We propose a semiparametric M-estimation method in the TW-SLM to estimate the normalization curves and the normalized expression values, and discuss several useful extensions of the TW-SLM. We describe a back-fitting algorithm for computation in the model. We illustrate the application of the TW-SLM by applying it to a microarray data set. We evaluate the performance of TW-SLM using simulation studies and consider theoretical results concerning the asymptotic distribution and rate of convergence of the least-squares estimators in the TW-SLM.

Microarrays are a useful tool for monitoring gene expression levels on a large scale and has been widely used in functional genomics [40.1, 2]. In a microarray experiment, cDNA segments representing the collection of genes and expression sequence tags (ESTs) to be probed are amplified by the polymerase chain reaction (PCR) and spotted in high density on glass microscope slides using a robotic system. Such slides are called microarrays. Each microarray contains thousands of reporters of the collection of genes or ESTs. The microarrays are queried in a co-hybridization assay using two fluorescently labeled biosamples prepared from the cell populations of interest. One sample is labeled with the fluorescent dye Cy5 (red), and another with the fluorescent dye Cy3 (green). Hybridization is assayed using a confocal laser scanner to measure fluorescence intensities, allowing simultaneous determination of the relative expression levels of all the genes represented on the slide [40.3]. The ability to monitor gene expressions on a large scale promises to have a profound impact on the understanding of basic cellular processes, developing better tools for disease diagnostics and treatment, cancer classification, and identifying drug targets, among others. Indeed, microarrays have already been used for detecting differentially expressed genes in different cell populations, classifying different cancer subtypes, identifying gene clusters based on co-expressions [40.4–7].

Because a microarray experiment monitors thousands of genes simultaneously, it routinely produces a massive amount of data. This and the unique nature of microarray experiments present a host of challenging statistical issues. Some of these can be dealt with using the existing statistical methods, but many are novel questions that require innovative solutions. One such question is normalization. The purpose of normalization is to remove bias in the observed expression levels and establish the baseline ratios of intensity levels from

the florescent dyes Cy3 and Cy5 across the whole dynamic range. A proper normalization procedure ensures that the intensity ratios provide meaningful measures of relative expression levels. In a microarray experiment, many factors may cause bias in the observed expression levels, such as differential efficiency of dye incorporation, differences in concentration of DNA on arrays, difference in the amount of RNA labeled between the two channels, uneven hybridizations, differences in the printing pin heads, among others.

Many researchers have considered various normalization methods; see for example [40.8–13]. For reviews of some of the existing normalization methods, see [40.14, 15]. More recently, *Fan* et al. [40.16] proposed a semilinear in-slide model (SLIM) method that makes use of replications of a subset of the genes in an array. If the number of replicated genes is small, the expression values of the replicated genes may not cover the entire dynamic range or reflect the spatial variation in an array. *Fan* et al. [40.17] 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 widely used normalization method is the local regression *lowess* [40.18] normalization proposed by *Yang* et al. [40.11]. This method estimates the normalization curves using the robust *lowess* for log-intensity ratio versus log-intensity product using all the genes in the study. The underlying assumption of this normalization method is either that the number of differentially expressed genes is relatively small or that the expression levels of up- and down-regulated genes are symmetric, so that the *lowess* normalization curves are not affected significantly by the differentially expressed genes. If it is expected that many genes will have differential expressions, *Yang* et al. [40.11] suggested using dye-swap for normalization. This approach makes the assumption that the normalization curves in the two dye-swaped slides are symmetric. Because of the slide-to-slide variation, this assumption may not always be satisfied.

Strictly speaking, an unbiased normalization curve should be estimated using genes whose expression levels remain constant and cover the whole range of the intensity. Thus *Tseng* et al. [40.12] first used a rank-based procedure to select a set of invariant genes that are likely to be non-differentially expressed, and then use these genes for *lowess* normalization. However, they pointed out that the number of invariant genes may be small and not cover the whole dynamic range of the expression values, and extrapolation is needed to fill in the gaps that are not covered by the invariant genes. In addition, a threshold value is required in this rank-based procedure. The level of the sensitivy of the final result to the threshold value may need to be evaluated on a case-by-case basis.

A common practice in microarray data analysis is to consider normalization and detection of differentially expressed genes separately. That is, the normalized values of the observed expression levels are treated as data in the subsequent analysis. However, because normalization typically includes a series of statistical adjustments to the data, there are variations associated with this process. These variations will be inherited in any subsequent analysis. It is desirable to take them into account in order to assess the uncertainty of the analysis results correctly.

We have proposed a two-way semilinear model (TW-SLM) for normalization and analysis of microarray data [40.19–21]. When this model is used for normalization, it does not require some of the assumptions that are needed in the *lowess* normalization method. Below, we first give a description of this model, and then suggest an M-estimation (including the least squares estimator as a special case) and a local regression method for estimation in this model. We describe a back-fitting algorithm for computation in the model. We then consider several useful extensions of this model. We illustrate the application of the TW-SLM by applying it to the Apo A1 data set [40.7]. We evaluate the performance of TW-SLM using simulation studies. We also state theoretical results concerning the asymptotic distribution and rate of convergence of the least squares estimator of the TW-SLM.

40.1 The Two-Way Semilinear Model

Suppose there are J genes and n slides in the study. Let R_{ij} and G_{ij} be the red (Cy 5) and green (Cy 3) intensities of gene j in slide i, respectively. Let y_{ij} be the log-intensity ratio of the red over green channels of the j-th gene in the i-th slide, and let x_{ij} be the corresponding average of the log-intensities of the red and green channels. That is,

$$y_{ij} = \log_2 \frac{R_{ij}}{G_{ij}}, \quad x_{ij} = \frac{1}{2}\log_2(R_{ij}G_{ij}),$$
$$i = 1, \ldots, n, \quad j = 1, \ldots, J\,.$$

Let $z_i \in \mathbb{R}^d$ be a covariate vector associated with the i-th slide. It can be used to code various types of designs. The TW-SLM model decomposes the observed intensity ratio y_{ij} in the following way:

$$y_{ij} = f_i(x_{ij}) + z_i' \beta_j + \sigma_{ij}\varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J, \tag{40.1}$$

where f_i is the intensity-dependent normalization curve for the i-th slide, $\beta_j \in \mathbb{R}^d$ is the effect associated with the j-th gene; σ_{ij} are the residual standard deviation, ε_{ij} have mean 0 and variance 1. We note that f_i can be considered as the log-intensity ratios in the absence of the gene effects. From a semiparametric modeling standpoint, these f_i functions are nonparametric components in the model and are to be estimated. In model (40.1), it is only made explicit that the normalization curves f_i are slide-dependent. It can also be made dependent upon regions of a slide to account for spatial effects. For example, it is straightforward to extend the model with an additional subscript in (y_{ij}, x_{ij}) and f_i and make f_i also depend on the printing-pin blocks within a slide. We describe this and two other extensions of TW-SLM in Sect. 40.4. Below, we denote the collection of the normalization curves by $\boldsymbol{f} = \{f_1, \ldots, f_n\}$ and the matrix of the gene effects by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)' \in \mathbb{R}^{J\times d}$. Let $\Omega_0^{J\times d}$ be the space of all $J \times d$ matrices $\boldsymbol{\beta}$ satisfying $\sum_{j=1}^J \beta_j = 0$. From the definition of the TW-SLM model (40.1), $\boldsymbol{\beta}$ is identifiable only up to a member in $\Omega_0^{J\times d}$.

We call (40.1) TW-SLM since it contains the two-way analysis of variation (ANOVA) model as a special case with $f_i(x_{ij}) = \alpha_i$ and $z_i = 1$. Our approach naturally leads to the general TW-SLM

$$y_{ij} = f_i(x_{ij}) + z_{ij}' \beta_j + \varepsilon_{ij}, \tag{40.2}$$

which could be used to incorporate additional prior knowledge in the TW-SLM (Sect. 40.3). The identifiability condition $\sum_j \beta_j = 0$ is no longer necessary in (40.2) unless $z_{ij} = z_i$ as in (40.1).

The TW-SLM is an extension of the semiparametric regression model (SRM) proposed by *Wahba* [40.22] and *Engle* et al. [40.23]. Specifically, if $f_1 = \cdots = f_n \equiv f$ and $J = 1$, then the TW-SLM simplifies to the SRM, which has one nonparametric component and one finite-dimensional regression parameter. Much work has been done concerning the properties of the semiparametric least squares estimator (LSE) in the SRM, see for example, *Heckman* [40.24] and *Chen* [40.25]. It has been shown that, under appropriate regularity conditions, the semiparametric least squares estimator of the finite-dimensional parameter in the SRM is asymptotically normal, although the rate of convergence of the estimator of the nonparametric component is slower than $n^{1/2}$.

40.2 Semiparametric M-Estimation in TW-SLM

We describe two approaches of semiparametric M-estimation in the TW-SLM. The first one uses linear combinations of certain basis functions (e.g. B-splines) to approximate the normalization curves. The second one uses the local regression technique for estimation in the TW-SLM. Three important special cases in each approach include the least squares estimator, the least absolute deviation estimator, and Huber's robust estimator [40.26].

40.2.1 Basis-Based Method

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})'$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iJ})'$ and $f(\boldsymbol{x}_i) \equiv [f(x_{i1}), \ldots, f(x_{iJ})]'$ for a univariate function f. We write the TW-SLM (40.1) in vector notation as

$$\boldsymbol{y}_i = \boldsymbol{\beta} z_i + f_i(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{40.3}$$

Let $\Omega_0^{J\times d}$ be the space of all $J \times d$ matrices $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_J)'$ satisfying $\sum_{j=1}^J \beta_j = 0$. It is clear from the definition of the TW-SLM model (40.3) that $\boldsymbol{\beta}$ is identifiable only up to a member in $\Omega_0^{J\times d}$, since we may simply replace β_j by $\beta_j - \sum_{k=1}^J \beta_k/J$ and $f_i(x)$ by $f_i(x) + \sum_{k=1}^J \beta_k' z_i/J$ in (40.1). In what follows, we assume

$$\boldsymbol{\beta} \in \Omega_0^{J\times d} \equiv \left\{ \boldsymbol{\beta} : \sum_{j=1}^J \beta_j = 0 \right\}. \tag{40.4}$$

Let $b_{i1}, \ldots, b_{i,K_i}$ be K_i B-spline basis functions[40.27]. Let

$$S_i \equiv \overline{\{b_{i0}(x) \equiv 1, b_{ik}(x), k = 1, \ldots, K_i\}} \tag{40.5}$$

be the spaces of all linear combinations of the basis functions. We note that wavelet, Fourier and other types of basis functions can also be used. We approximate f_i

by

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

where $\boldsymbol{b}_i(x) = [1, b_{i1}(x), \ldots, b_{i,K_i}(x)]'$, and $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iK_i})'$ are coefficients to be estimated from the data. Let $\boldsymbol{b}f = (f_1, \ldots, f_n)$ and

$$M_s(\boldsymbol{\beta}, \boldsymbol{f}) = \sum_{i=1}^{n}\sum_{j=1}^{J} m_s\left[y_{ij} - f_i(x_{ij}) - \beta_j' z_i\right], \tag{40.6}$$

where m_s is an appropriate convex function which may also depend on a scale parameter s. Three important special cases are $m_s(t) = t^2$, $m_s(t) = |t|$, and the Huber ρ function. We define the semiparametric M-estimator of $\{\boldsymbol{\beta}, \boldsymbol{f}\}$ to be the $\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}\} \in \Omega_0^{J\times d} \times \prod_{i=1}^{n} S_i$ that minimizes $M_s(\boldsymbol{\beta}, \boldsymbol{f})$. It is often necessary to consider a scale parameter s in robust estimation. This scale parameter usually needs to be estimated jointly with $(\boldsymbol{\beta}, \boldsymbol{f})$.

One question is how to determine the number of basis functions K_i. For the purpose of normalization, it is reasonable to use the same K for all the arrays, that is, let $K_1 = \cdots = K_n \equiv K$. This will make normalization consistent across the arrays. For the cDNA microarray data, the total intensity has positive density over a finite interval, typically [0, 16]. For the cubic polynomial splines, we have used the number of knots $K = 12$, and the data percentiles as the knots in the R function bs.

40.2.2 Local Regression (Lowess) Method

We can also use the lowess method [40.18] for the estimation of TW-SLM. Let W_λ be a kernel function with window width λ. Let

$$s_p(t; \boldsymbol{\alpha}, x) = \alpha_0(x) + \alpha_1(x)t + \cdots + \alpha_p(x)t^p$$

be a polynomial in t with order p, where $p = 1$ or 2 are common choices. The objective function of the *lowess* method for the TW-SLM is

$$M_s(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{J} W_\lambda(x_{ik}, x_{ij}) m_s\Big[y_{ik} - s_p(x_{ik}; \boldsymbol{\alpha}, x_{ij}) - z_i'\beta_k\Big]. \tag{40.7}$$

Let $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ be the value that minimizes M_L. The lowess M-estimator of f_i at x_{ij} is $\hat{f}_i(x_{ij}) = s_p(x_{ij}, \hat{\boldsymbol{\alpha}}, x_{ij})$.

40.2.3 Back-Fitting Algorithm in TW-SLM

In both the basis-based and local regression methods, we use a back-fitting algorithm [40.28] to compute the semiparametric M-estimators. For the M-estimator based on the basis spaces S_i defined in (40.6), set $\boldsymbol{\beta}^{(0)} = 0$. For $k = 0, 1, 2, \ldots,$

- Step 1: compute $\boldsymbol{f}^{(k)}$ by minimizing $M_s\left(\boldsymbol{f}, \boldsymbol{\beta}^{(k)}\right)$ with respect to the space $\prod_{i=1}^{n} S_i$.
- Step 2: for the $\boldsymbol{f}^{(k)}$ computed above, obtain $\boldsymbol{\beta}^{(k+1)}$ by minimizing $M_s\left(\boldsymbol{f}^{(k)}, \boldsymbol{\beta}\right)$ with respect to $\boldsymbol{\beta}$ in $\Omega_0^{J\times d}$.

Iterate between steps 1 and 2 until the desired convergence criterion is satisfied.

For strictly convex m, e.g., $m(t) = t^2$ or $m(t) = |t|$, the algorithm converges to the unique global optimal point. The back-fitting algorithm can be also applied to the lowess M-estimators. When $m(t) = t^2$, then computation consists of a series of weighted regression problems.

40.2.4 Semiparametric Least Squares Estimation in TW-SLM

An important special case of the M-estimator is the least squares (LS) estimator, which has an explicit form in the TW-SLM [40.19, 20]. The LS objective function is

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

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

$$\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{f}}\right) = \arg\min_{(\boldsymbol{\beta}, \boldsymbol{f}) \in \Omega_0^{J\times d} \times \prod_{i=1}^{n} S_i} D^2(\boldsymbol{\beta}, \boldsymbol{f}). \tag{40.8}$$

Denote the spline basis matrix for the i-th array by

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

Define the projection matrix Q_i as

$$Q_i = B_i\left(B_i'B_i\right)^{-1}B_i', \quad i = 1, \ldots, n.$$

Let $\alpha_i = (\alpha_{i0}, \ldots, \alpha_{iK_i})'$ be the spline coefficients for the estimation of f_i and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$. We can write $D^2(\boldsymbol{\beta}, \boldsymbol{\alpha}) = D^2(\boldsymbol{\beta}, \boldsymbol{f})$. Then the problem of minimizing

$D^2(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is equivalent to solving the linear equations:

$$\hat{\boldsymbol{\beta}} \sum_{i=1}^{n} z_i z_i' + \sum_{i=1}^{n} B_i \hat{\alpha}_i z_i' = \sum_{i=1}^{n} \boldsymbol{y}_i z_i', \; B_i' B_i \hat{\alpha}_i + B_i' \hat{\boldsymbol{\beta}} z_i = B_i' \boldsymbol{y}_i \; .$$

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

Using (40.3), it can be shown that the SLSE (40.8) equals

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\| \boldsymbol{y}_i - (I_J - Q_i)\, \boldsymbol{\beta} z_i \right\|^2 . \quad (40.9)$$

In the special case when $d = 1$ (scalar β_j) and $\boldsymbol{\beta}$ is a vector in $\mathbb{R}^J$, (40.9) is explicitly

$$\hat{\boldsymbol{\beta}} = \hat{\Lambda}^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i) \boldsymbol{y}_i z_i' \right] , \quad (40.10)$$

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

$$\hat{\Lambda}_{J,n} \equiv \frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i) \otimes z_i z_i' . \quad (40.11)$$

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

For general z_i and $d \geq 1$, (40.9) is still given by (40.10) with

$$\hat{\Lambda}_{J,n} \equiv \frac{1}{n} \sum_{i=1}^{n} (I_J - Q_i) \otimes z_i z_i' . \quad (40.12)$$

The information operator (40.11) is an average of tensor products, i. e. a linear mapping from $\Omega_0^{J \times d}$ to $\Omega_0^{J \times d}$ defined by $\hat{\Lambda}\boldsymbol{\beta} \equiv n^{-1} \sum_{i=1}^{n} (I_J - Q_i) \boldsymbol{\beta} z_i z_i'$.

Although the SLSE has an explicit expression, direct computation of SLSE involves inversion of a large $J \times J$ matrix. So we use the back-fitting algorithm. In this case, computation in each step of the back-fitting algorithm becomes an easier least squares problem and has explicit expressions as follows. Set $\boldsymbol{\beta}^{(0)} = \boldsymbol{0}$. For $k = 0, 1, 2, \ldots,$

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

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

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

$$\hat{\beta}_j^{(k+1)} = \left(\sum_{i=1}^{n} z_i z_i' \right)^{-1} \sum_{i=1}^{n} z_i \left[y_{ij} - f_i^{(k)}\left(x_{ij}\right) \right] , \quad j = 1, \ldots, J \; . \quad (40.13)$$

The algorithm converges to the sum of residual squares. Suppose that the algorithm meets the convergence criterion at step K. Then the estimated values of β_j are $\widehat{\beta}_j = \beta_j^{(K)}$, $j = 1, \ldots, J$, and the estimated normalization curves are

$$\hat{f}_i(x) = \boldsymbol{b}_i(x)' \alpha_i^{(K)} = \boldsymbol{b}_i(x)' (B_i' B_i)^{-1} B_i' (\boldsymbol{y}_i - \hat{\boldsymbol{\beta}} z_i) , \quad i = 1, \ldots, n \; . \quad (40.14)$$

The algorithm described above can be conveniently implemented in the statistical computing environment R [40.29]. Specifically, steps 1 and 2 can be solved by the function LM in R. The function BS can be used to create a basis matrix for the polynomial splines.

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})'$ and $f_i(\boldsymbol{x}_i) = [f_i(x_{i1}), \ldots, f_i(x_{iJ})]'$. Let $Q_i = B_i (B_i' B_i)^{-1} B_i'$. By (40.14), the estimator of $f_i(\boldsymbol{x}_i)$ is

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

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

$$\tilde{f}_i(\boldsymbol{x}_i) = Q_i \boldsymbol{y}_i = B_i \boldsymbol{\alpha}_i^{(0)} . \quad (40.15)$$

We shall call (40.15) a *spline* normalization method. Comparing $\hat{f}_i(\boldsymbol{x}_i)$ with $\tilde{f}_i(\boldsymbol{x}_i)$, we find that, if there is a relatively large percentage of differentially expressed genes, the difference between these two normalization curves can be large. The magnitude of the difference also depends on the magnitude of the gene effects.

40.3 Extensions of the TW-SLM

In this section, we describe three models that are extensions of the basic TW-SLM. These models include the multi-way SLM (MW-SLM); a model that incorporates control genes in the normalization; and a model for simultaneous location and scale normalization.

40.3.1 Multi-Way Semilinear Models

Just as TW-SLM is a semilinear extension of two-way ANOVA, for data sets with a higher-dimensional structure, multi-way ANOVA can be extended to multi-way 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 connection between ANOVA and MW-SLM is important in design of experiments and in understanding and interpretation of the contribution of different effects and identifiability conditions. The examples below are motivated by real data sets.

In model (40.1), it is only made explicit that the normalization curve f_i is array-dependent. It is straightforward to construct a 3W-SLM to normalize the data at the printing-pin block level:

$$y_{ikj} = f_{ik}(x_{ikj}) + z_i'\beta_{kj} + \epsilon_{ikj}\ , \qquad (40.16)$$

with the identifiability condition $\sum_j \beta_{kj} = 0$, where y_{ikj} and x_{ikj} are the log-intensity ratio and log-intensity product of gene j in the k-th block of array i, respectively. Model (40.16) 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 in *Huang* et al. [40.21] to analyze the Apo A1 data [40.7], 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 (40.16) since 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, we may assume no interaction between gene and block effects ($\beta_{kj} = \beta_j$) in (40.16) and reduce it to the TW-SLM with (i, k) as a single index, treating a block/array in (40.16) as an array in (40.1).

As an alternative to (40.16) we may also use constants to model the interaction between array and block effects as in ANOVA, resulting in the model

$$y_{ikj} = f_i(x_{ikj}) + \gamma_{ik} + z_i'\beta_{kj} + \epsilon_{ikj}\ , \qquad (40.17)$$

with identifiability conditions $\sum_i \gamma_{ik} = \sum_k \gamma_{ik} = 0$ and $\sum_{kj} \beta_{kj} = 0$. This can be viewed as an extension of the three-way ANOVA model $Ey_{ikj} = \mu + \alpha_{i\bullet\bullet} + \gamma_{ik\bullet} + \beta_{\bullet k\bullet} + \beta_{\bullet kj} + \beta_{\bullet\bullet j}$ without $\{i, j\}$ and three-way interactions, via $\mu + \alpha_{i\bullet\bullet} \Rightarrow f_i$ and $\beta_{\bullet k\bullet} + \beta_{\bullet kj} + \beta_{\bullet\bullet j} \Rightarrow \beta_{kj}$. Note that the main block effects are represented by f_{ik} in (40.16) and by β_{kj} in (40.17).

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 [40.12]. Suppose 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 (40.17), a 4W-SLM can be written as

$$y_{icrj} = f_i(x_{icrj}) + \gamma_{icr} + z_i'\beta_{rj} + \epsilon_{icrj} \qquad (40.18)$$

for observations with the j-th gene in the block at c-th column and r-th row of the matrix in the i-th array, with identifiability conditions $\sum_i \gamma_{icr} = \sum_r \gamma_{icr} = 0$ and $\sum_{rj} \beta_{rj} = 0$, 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 (40.18), 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}, the f_i and γ_{icr} in (40.18) can be combined into $f_{icr}(x_{icrj})$, resulting in the TW-SLM (for each fixed r) at the row level, which is equivalent to (40.16). If the replication of genes is not balanced, we may use a MW-SLM derived from an ANOVA model with incomplete/unbalanced design or the modeling methodologies described in Sect. 40.2.

From the above 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 the presence or absence of associated identifiability conditions. Furthermore, since the center of a nonparametric component, e.g. $\sum_j f_i(x_{ij})$ in a TW-SLM, is harder to interpret, identifiability conditions are usually imposed on parametric components. As a result, a nonparametric component representing an interactive effect usually represents all 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 (40.16).

40.3.2 Spiked Genes and Incorporation of Prior Knowledge in the MW-SLM

We describe three methods to incorporate prior knowledge in an MW-SLM: augmenting models, coding covariates, and imposing linear constraints. An important application of these methods is inclusion of spiked genes in normalization.

In many customized microarray experiments, it is possible to include a set of spiked genes with equal concentrations in the Cy5 and Cy3 channels. An important reason to use spiked genes is to calibrate scanning parameters, for example, intensity levels from the spiked genes can be used for tuning the laser power in each scanning channel in order to balance the Cy5 and Cy3 intensities. Spiked genes do not necessarily show an observed 1:1 ratio due to experimental variations. Because the number of spiked genes is often small, it is not adequate just to use the spiked genes as the basis for normalization.

Let y_{ik}^s and x_{ik}^s be the log-intensity ratio and product of the k-th spiked gene in the i-th array, $i = 1, \ldots, n$, $k = 1, \ldots, K$. Then we can augment the TW-SLM (40.1) as follows:

$$y_{ik}^s = f_i\left(x_{ik}^s\right) + \varepsilon_{ik}^s,\ y_{ij} = f_i\left(x_{ij}\right) + z_i'\beta_j + \varepsilon_{ij}\ . \tag{40.19}$$

The first equation is for the spiked genes, whose corresponding β_k^s are zero. Since a common f_i is used in (40.19) for each array, data from both spiked genes and genes under study contribute to the estimation of normalization curves as well as gene effects. Note that the identifiability condition $\sum_j \beta_j = 0$ in (40.1) is neither necessary nor appropriate for (40.19).

We may also use the general TW-SLM (40.2) to model spiked genes by simply setting $z_{ij} = 0$ if a spiked gene is printed at the j-th spot in the i-th array and $z_{ij} = z_i$ otherwise, where z_i are the design variable for the i-th array as in (40.1).

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

40.3.3 Location and Scale Normalization

The models we described above 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 in *Yang* et al. [40.11], 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

$$\frac{y_{ij} - f_i(x_{ij})}{\tau_i} = z_i'\beta_j + \varepsilon_{ij}, \quad i = 1, \ldots, n\ , \quad j = 1, \ldots, J\ , \tag{40.20}$$

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

40.4 Variance Estimation and Inference for β

In addition to being a standalone model for normalization, the TW-SLM can also be used for detecting differentially expressed genes. For this purpose, we need to estimate the variance of $\hat{\boldsymbol{\beta}}$. This requires the estimation of residual variances.

We have considered the model in which the residual variances depend smoothly on the total intensity values, and such dependence may vary from array to array [40.21]. The model is

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

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

We use the method proposed by *Ruppert* et al. [40.30] and *Fan* and *Yao* [40.31] in estimating the variance function in a nonparametric regression model. For each $i = 1, \ldots, n$, we fit a smooth curve through the scatter plot $(x_{ij}, \hat{\epsilon}_{ij}^2)$, where $\hat{\epsilon}_{ij}^2 = (y_{ij} - \hat{f}_i(x_{ij}) - z_i'\hat{\beta}_j)^2$. This is equivalent to fitting the nonparametric regression

model

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

for $i = 1, \ldots, n$, where τ_{ij} is the residual term in this model. We use the same spline bases as in the estimation of f_i (40.14). The resulting spline estimator $\hat{\sigma}_i^2$ can be expressed as

$$\hat{\sigma}_i^2(x) = \boldsymbol{b}_i'(x)(B_i'B_i)^{-1}B_i'\hat{\boldsymbol{\epsilon}}_i^2, \tag{40.21}$$

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

We can now approximate the variance of $\hat{\beta}_j$ as follows [40.21]. Let $Z_n = \sum_{i=1}^n z_i z_i'$. Based on (40.13), we have

$$\begin{aligned}\mathrm{var}(\hat{\beta}_j) &\approx Z_n^{-1}\left[\sum_{i=1}^n z_i z_i' \mathrm{var}(\epsilon_{ij})\right] Z_n^{-1} \\ &\quad + Z_n^{-1}\left[\sum_{i=1}^n z_i z_i' \mathrm{var}\big[\hat{f}_i(x_{ij})\big]\right] Z_n^{-1} \\ &\equiv \Sigma_{\epsilon,j} + \Sigma_{f,j}\,.\end{aligned}$$

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

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

$$\Sigma_{\epsilon,j} = Z_n^{-1}\left(\sum_{i=1}^n z_i z_i' \sigma_{ij}^2\right) Z_n^{-1}\,.$$

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

$$\hat{\Sigma}_{\epsilon,j} = Z_n^{-1}\left(\sum_{i=1}^n z_i z_i' \hat{\sigma}_{ij}^2\right) Z_n^{-1}\,.$$

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

$$\begin{aligned}\hat{f}_i(\boldsymbol{x}_i) &= Q_i\big(\boldsymbol{y}_i - \hat{\boldsymbol{\beta}} z_i\big) \approx Q_i\big(\boldsymbol{y}_i - \boldsymbol{\beta} z_i\big) \\ &= Q_i[\boldsymbol{\epsilon}_i + f_i(\boldsymbol{x}_i)]\,.\end{aligned}$$

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

$$\mathrm{var}\left[\hat{f}_i(\boldsymbol{x}_i)\right] \approx Q_i \mathrm{var}(\boldsymbol{\epsilon}_i) Q_i,$$

and

$$\mathrm{var}\left[\hat{f}_i(x_{ij})\right] = \boldsymbol{e}_j' Q_i \mathrm{var}(\boldsymbol{\epsilon}_i) Q_i \boldsymbol{e}_j\,,$$

where $\boldsymbol{e}_j$ is the unit vector whose j-th element is 1. Let $\hat{\Sigma}_i$ be an estimator of $\mathrm{var}(\boldsymbol{\epsilon}_i)$. We estimate $\Sigma_{f,j}$ by

$$\hat{\Sigma}_{f,j} = Z_n^{-1}\boldsymbol{e}_j'\left(\sum_{i=1}^n Q_i \hat{\Sigma}_i Q_i\right)\boldsymbol{e}_j Z_n^{-1}\,.$$

Finally, we estimate $\mathrm{var}\big(\hat{\beta}_j\big)$ by

$$\hat{\Sigma}_{\beta,j} = \hat{\Sigma}_{\epsilon,j} + \hat{\Sigma}_{f,j}\,. \tag{40.22}$$

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

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

As is shown in Sect. 40.6, for large J, the distribution of t_j can be approximated by the standard normal distribution under the null $c'\beta_j = 0$. However, to be conservative, we use a t distribution with an appropriate number of degrees of freedom to approximate the null distribution of t_j when $c'\beta_j = 0$. For example, for a direct comparison design, the number of degrees of freedom is $n-1$. For a reference design in a two sample comparison, the variances for the two groups can be estimated separately, and then Welch's correction for the degrees of freedom can be used. Resampling methods such as permutation or bootstrap can also be used to evaluate the distribution of t_j.

Another approach is to estimate σ_{ij}^2 jointly with $(\boldsymbol{f}, \boldsymbol{\beta})$. This approach is computationally more intensive but may yield more efficient estimates of $(\boldsymbol{\beta}, \boldsymbol{f})$ and σ_{ij}^2. Consider an approximation to σ_{ij}^2 using the spline basis functions:

$$\sigma_{ij}^2 = \sigma_i^2(x_{ij}) = \sum_{k=1}^{K_i} \gamma_{ik} b_k(x_{ij})\,. \tag{40.23}$$

Let $\boldsymbol{\gamma}$ be the collection of the γ_{ik}. Assuming normality for ε_{ij}, the negative likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\gamma}) = -\prod_{i=1}^n \prod_{j=1}^J \frac{1}{\sigma_{ij}}\phi\left(\frac{y_{ij} - f_i(x_{ij}) - \beta_j' z_i}{\sigma_{ij}}\right), \tag{40.24}$$

where ϕ is the density of $N(0,1)$. For robust M-estimation, we define the M-estimation objective function as

$$M_s(\boldsymbol{\beta}, \boldsymbol{f}, \boldsymbol{\gamma}) = \sum_{i=1}^n \sum_{j=1}^J \sigma_{ij} m_s\left(\frac{y_{ij} - f_i(x_{ij}) - \beta_j' z_i}{\sigma_{ij}}\right). \tag{40.25}$$

Again, we can use a back-fitting algorithm for computing the M-estimators, but with an extra step in each iteration for $\boldsymbol{\gamma}$.

40.5 An Example and Simulation Studies

40.5.1 Apo A1 Data

We now illustrate the TW-SLM for microarray data by the Apo A1 data set of *Callow* et al. [40.7]. The analysis described here is from *Huang* et al. [40.21]. The purpose of this experiment is to identify differentially expressed genes in the livers of mice with very low high-density lipoprotein (HDL) cholesterol levels compared to inbred mice. The treatment group consists of eight mice with the apo A1 gene knocked out and the control group consists of eight C57BL/6 mice. For each of these mice, target cDNA is obtained from mRNA by reverse transcription and labeled using a red fluorescent dye (Cy5). The reference sample (green fluorescent dye Cy3) used in all hybridizations was obtained by pooling cDNA from the eight control mice. The target cDNA is hybridized to microarrays containing 5548 cDNA probes. This data set was analyzed by *Callow* et al. [40.7] and *Dudoit* et al. [40.32]. Their analysis uses lowess normalization and the two-sample t-statistic. Eight genes with multiple comparison adjusted permutation p-value ≤ 0.01 are identified.

We apply the proposed normalization and analysis method to this data set. As in *Dudoit* et al. [40.32], we use printing-tip-dependent normalization. The TW-SLM model used here is

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

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

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

where σ^2_{ik} are unknown smooth functions. We apply the printing-pin-dependent normalization and estimation approach described in Sect. 40.3.2. The covariate $z_i = (1, 0)'$ for the treatment group (apo A1 knock-out mice) and $z_i = (0, 1)'$ for the control group (C57BL/6 mice). The coefficient $\beta_{kj} = (\beta_{kj1}, \beta_{kj2})$. The contrast $\beta_{kj1} - \beta_{kj2}$ measures the expression difference for the j-th gene in the k-th block between the two groups.

To compare the proposed method with the existing ones, we also analyzed the data using the lowess normalization method as in *Dudoit* et al. [40.32], and a lowess-like method where, instead of using local regression, splines are used in estimating the normalization curves described in (40.15) at the end of Sect. 40.2. We refer to this method as the spline (normalization) method below.

As examples of the normalization results, Fig. 40.1 displays the M–A plots and printing-tip-dependent normalization curves in the 16 printing-pin blocks of the array from one knock-out mouse. The solid line is the normalization curve based on the TW-SLM model, and the dashed line is the lowess normalization curve. The degrees of freedom used in the spline basis function in the TW-SLM normalization is 12, and following *Dudoit* et al. [40.32], the span used in the lowess normalization is 0.40. We see that there are differences between the normalization curves based on the two methods. The lowess normalization curve attempts to fit each individual M–A scatter plot, without taking into account the gene effects. In comparison, the TW-SLM normalization curves do not follow the plot as closely as the lowess normalization. The normalization curves estimated using the spline method with exactly the same basis functions used in the TW-SLM closely resemble those estimated using the lowess method. Because they are indistinguishable by eye, these curves are not included in the plots.

Figure 40.2 displays the volcano plots of $-\log_{10}$ p-values versus the mean differences of log-expression values between the knock-out and control groups. In the first (left panel) volcano plot, both the normalization and estimation of $\boldsymbol{\beta}$ are based on the TW-SLM. We estimated the variances for $\hat{\beta}_{kj1}$ and $\hat{\beta}_{kj2}$ separately. These variances are estimated based on (40.21), which assumes that the residual variances depend smoothly on the total log-intensities. We then used Welch's correction for the degrees of freedom in calculating the p-values. The second (middle panel) plot is based on the lowess normalization method and use the two-sample t-statistics as in *Dudoit* et al. [40.32], but the p-values are obtained based on Welch's correction for the degrees of freedom. The third (right panel) plot is based on the spline normalization method and uses the same two-sample t-statistics as in the lowess method. The eight solid circles in the lowess volcano plot are the significant genes that were identified by *Dudoit* et al. [40.32]. These eight genes are also plotted as solid circles in the TW-SLM and spline volcano plots, and are significant based on the TW-SLM and spline methods, as can be seen from the volcano plots. Comparing the three volcano plots, we see that: (i) the $-\log_{10}$ p-values based on the TW-SLM method tend to be higher than those based on the lowess and

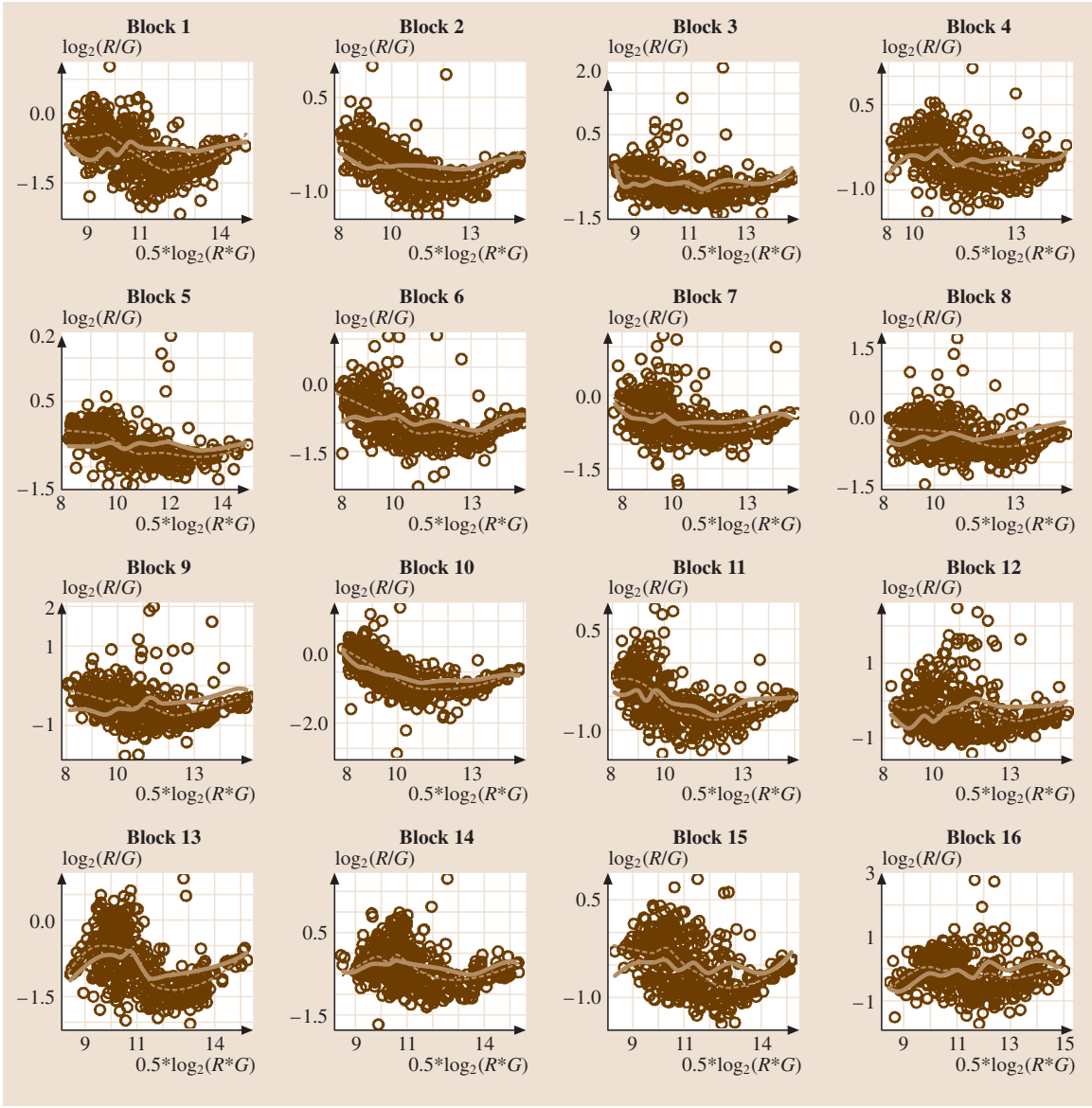

Fig. 40.1 Apo AI data: comparison of normalization curves in the 16 blocks of the array from one knock-out mouse in the treatment group. *Solid line*: normalization curve based on TW-SLM; *dashed line*: normalization curve based on lowess

spline methods; (ii) the p-values based on the lowess and spline methods are comparable.

Because we use exactly the same smoothing procedure in the TW-SLM and spline methods, and because the results between the lowess and spline methods are very similar, we conclude that the differences between the TW-SLM and lowess volcano plots are mostly due to the different normalization methods and two difference approaches for estimating the variances. We first examine the differences between the TW-SLM normalization values and the lowess as well as the spline normalization values. We plot the three pairwise scatter plots of estimated mean expression differences based on the TW-SLM, lowess, and spline normalization methods, see

Fig. 40.3. In each scatter plot, the solid line is the fitted linear regression line. For the TW-SLM versus lowess comparison (left panel), the fitted regression line is

$$y = 0.00029 + 1.090x\,. \tag{40.26}$$

The standard error of the intercept is 0.0018, so the intercept is negligible. The standard error of the slope is 0.01. Therefore, on average, the mean expression differences based on the TW-SLM normalization method are about 10% higher than those based on the lowess normalization method. For the TW-SLM versus spline comparison (middle panel), the fitted regression line and the standard errors are virtually identical to (40.26) and its associated standard errors. For the spline versus lowess comparison (right panel), the fitted regression line is

$$y = 0.00027 + 1.00257x\,. \tag{40.27}$$

The standard error of the intercept is 0.000 25, and the standard of the slope is 0.0015. Therefore, the mean expression differences based on the lowess and spline normalization methods are essentially the same, as can also be seen from the scatter plot in the right panel in Fig. 40.3.

Figure 40.4 shows the histograms of the standard errors obtained based on intensity-dependent smoothing defined in (40.21) using the residuals from the TW-SLM normalization (top panel), and the standard errors calculated for individual genes using the lowess and spline methods (middle and bottom panels). The standard errors (SE) based on the individual genes have a relatively large range of variation, but the range of standard errors based on intensity-dependent smoothing shrinks towards the middle. The SEs based on the smoothing method are more tightly centered around the median value of about 0.13. Thus, the analysis based on the smooth estimate of the error variances is less susceptible to the problem of artificially small p-values resulting from random small standard errors calculated from individual genes.

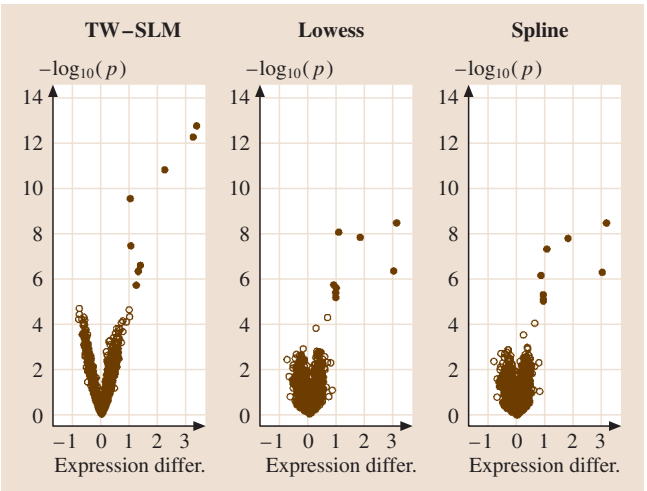

Fig. 40.2 Volcano plots: scatter plot of $-\log_{10}(p-\text{value})$ versus estimated mean expression value. The left panel shows the volcano plot based on the TW-SLM; the middle panel shows the plot based on the lowess method; the right panel shows the result based on the spline method

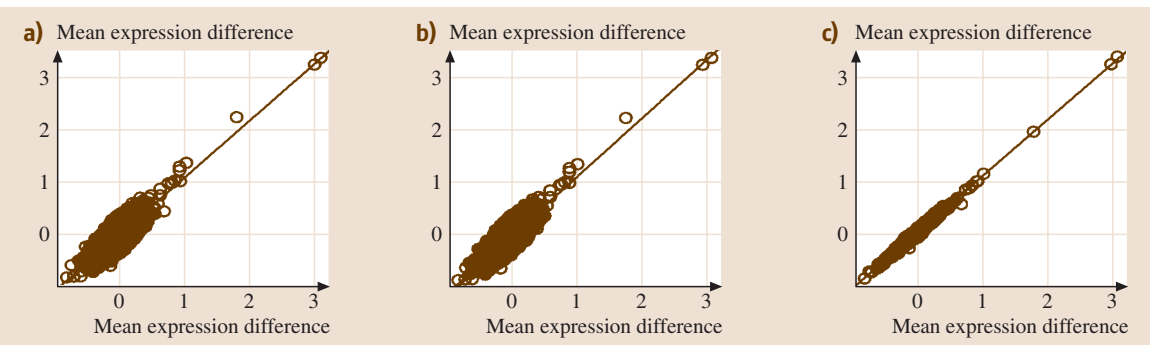

Fig. 40.3a–c Comparison of normalized expression values. *Left panel*: the scatter plot of normalized mean expression differences based on TW-SLM versus those based on lowess. *Middle panel*: The scatter plot of normalized mean expression differences based on TW-SLM versus those based on the spline method. *Right panel*: The scatter plot of normalized mean expression differences based on spline versus those based on lowess. **(a)** TW-SLM versus lowess, **(b)** TW-SLM versus spline, **(c)** spline versus lowess

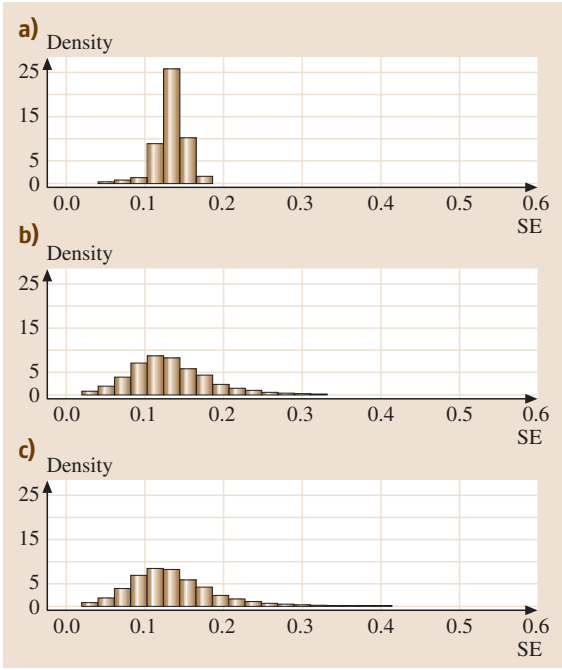

Fig. 40.4a–c Comparison of variance estimation methods. *Top panel*: The histogram of SE estimated based on smoothing as described in Sect. 40.3.2. *Middle panel*: SE estimated based on individual genes using the lowess method. *Bottom panel*: SE estimated based on individual genes using the spline method. **(a)** TW-SLM: SE based on smoothing **(b)** lowess: SE based on individual genes **(c)** spline: SE based on individual genes

40.5.2 Simulation Studies

We use simulation to compare the TW-SLM, lowess, and spline normalization methods with regard to the mean square errors (MSE) in estimating expression levels β_j. The simulation models and results are from *Huang* et al. [40.21]. Let α_1 and α_2 be the percentages of up- and down-regulated genes, respectively, and let $\alpha = \alpha_1 + \alpha_2$. We consider four models in our simulation.

- Model 1: there is no dye bias. So the true normalization curve is set at the horizontal line at 0. That is $f_i(x) \equiv 0$, $1 \le i \le n$. In addition, the expression levels of up- and down-regulated genes are symmetric and $\alpha_1 = \alpha_2$.
- Model 2: as in model 1, the true normalization curves $f_i(x) \equiv 0$, $1 \le i \le n$, but the percentages of up- and down-regulated genes are different. We set $\alpha_1 = 3\alpha_2$.
- Model 3: there are nonlinear and intensity-dependent dye biases. The expression levels of up- and down-regulated genes are symmetric and $\alpha_1 = \alpha_2$.
- Model 4: there is nonlinear and intensity-dependent dye bias. The percentages of up- and down-regulated genes are different. We set $\alpha_1 = 3\alpha_2$.

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

- Generate β_j. For most of the genes, we simulate $\beta_j \sim N(0, \tau_j^2)$. The percentage of such genes is $1-\alpha$. For up-regulated genes, we simulate $\beta_j \sim N(\mu, \tau_{\mathrm{U}j}^2)$ where $\mu > 0$. For down-regulated genes, we simulate $\beta_j \sim N(-\mu, \tau_{\mathrm{D}j}^2)$. We use $\tau_j = 0.6$, $\mu = 2$, $\tau_{\mathrm{U}j} = \tau_{\mathrm{D}j} = 1$.
- Generate x_{ij}. We simulate $x_{ij} \sim 16 \times Beta(a, b)$, where $a = 1$, $b = 2.5$.
- Generate ϵ_{ij}. We simulate $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$, where $\sigma_{ij} = \sigma(x_{ij})$. Here $\sigma(x) = 0.3 * x^{-1/3}$. So the error variance is higher at lower intensity range than at higher intensity range.
- In models 1 and 2, the log-intensity ratios are computed as $y_{ij} = f_i(x_{ij}) + \beta_j + \epsilon_{ij}$.

In models 3 and 4, the log-intensity ratios are computed according to a printing-tip-dependent model with $y_{ij} = \beta_j + f_{ik(j)}(x_{ij}) + \varepsilon_{ij}$, where the function $k(j)$ indicates the printing-pin block. This is equivalent to the model used in the analysis of the Apo A1 data in Sect. 40.5.1, with $z_i = 1$ there. We use

$$f_{ik} = \frac{a_{ik1} x^2 \sin(x/\pi)}{1 + a_{ik2} x^2} ,$$

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

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

Tables 40.1–40.4 present the summary statistics of the MSEs for estimating the relative expression levels β_j in the four models described above. In Table 40.1

Table 40.1 Simulation results for model 1. 10 000 × Summary of MSE. The true normalization curve is the horizontal line at 0. The expression levels of up- and down-regulated genes are symmetric: $\alpha_1 = \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$

		Min.	1st quartile	Median	Mean	3rd quartile	Max.
$\alpha = 0.01$	TW-SRM	3.74	51.59	75.08	88.88	106.20	4980.00
	Lowess	3.38	50.72	72.77	87.89	105.10	7546.00
	Splines	7.08	58.93	85.35	98.25	121.10	4703.00
$\alpha = 0.06$	TW-SRM	6.53	50.03	74.74	93.92	107.30	5120.00
	Lowess	9.09	50.89	73.93	91.87	106.10	6230.00
	Splines	8.95	61.34	89.03	105.60	126.10	6480.00

Table 40.2 Simulation results for model 2. 10 000 × Summary of MSE. The true normalization curve is the horizontal line at 0. But the percentages of up- and down-regulated genes are different: $\alpha_1 = 3\alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$

		Min.	1st quartile	Median	Mean	3rd quartile	Max.
$\alpha = 0.01$	TW-SRM	5.36	58.04	71.01	83.17	102.50	1416.00
	Lowess	8.86	67.69	95.80	107.40	131.00	1747.00
	Splines	8.91	65.53	94.40	110.40	135.10	1704.00
$\alpha = 0.06$	TW-SRM	6.66	47.85	68.55	78.49	97.50	1850.40
	Lowess	6.45	59.54	87.08	99.00	123.90	1945.10
	Splines	6.74	59.23	86.58	98.67	123.30	1813.10

Table 40.3 Simulation results for model 3. 10 000 × Summary of MSE. There are nonlinear and intensity-dependent dye biases. The expression levels of up- and down-regulated genes are symmetric: $\alpha_1 = \alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$

		Min.	1st quartile	Median	Mean	3rd quartile	Max.
$\alpha = 0.01$	TW-SRM	5.56	46.15	66.72	87.23	93.91	1898.00
	Lowess	6.71	51.07	74.23	88.79	107.50	3353.00
	Splines	5.90	53.83	76.91	88.64	108.60	1750.00
$\alpha = 0.06$	TW-SRM	6.64	57.26	85.79	102.80	126.40	2290.00
	Lowess	7.39	57.19	85.47	107.70	128.10	2570.00
	Splines	9.37	69.26	102.80	122.80	148.50	2230.00

Table 40.4 Simulation results for model 4. 10 000 × Summary of MSE. There are nonlinear and intensity-dependent dye biases. The percentages of up- and down-regulated genes are different: $\alpha_1 = 3\alpha_2$, where $\alpha_1 + \alpha_2 = \alpha$

		Min.	1st quartile	Median	Mean	3rd quartile	Max.
$\alpha = 0.01$	TW-SRM	5.89	51.26	74.53	85.89	107.20	2810.00
	Lowess	9.29	68.30	101.60	118.60	140.00	4088.00
	Splines	9.68	67.85	98.82	119.80	141.00	2465.00
$\alpha = 0.06$	TW-SRM	4.96	54.12	79.92	98.79	122.70	2130.00
	Lowess	6.49	71.54	113.90	130.90	169.50	2474.00
	Splines	5.77	65.46	107.57	128.40	171.60	1898.00

for simulation model 1, in which the true normalization curve is the horizontal line at 0 and the expression levels of up- and down-regulated genes are symmetric, the TW-SLM normalization tends to have slightly higher MSEs than the lowess method. The spline method has higher MSEs than both the TW-SLM and lowess methods. In Table 40.2, when there is no longer symmetry in the expression levels of up- and down-regulated genes, the TW-SLM method has smaller MSEs than both the lowess and spline methods. In Table 40.3 for simulation model 3, there is nonlinear intensity-dependent dye bias, but there is symmetry between the up- and down-regulated genes. The TW-SLM has comparable but slightly smaller MSEs than the lowess method. The spline method has higher MSEs than both the TW-SLM and lowess methods. In Table 40.4

for simulation model 4, there is nonlinear intensity-dependent dye bias, and the percentages of up- and down-regulated genes are different, the TW-SLM has considerably smaller MSEs. We have also examined biases and variances. There are only small differences in variances among the TW-SLM, lowess, and spline methods. However, the TW-SLM method generally has smaller biases.

40.6 Theoretical Results

In this section, we provide theoretical results concerning the distribution of $\hat{\boldsymbol{\beta}}$ and the rate of convergence for the normalization of f_i. The proofs can be found in *Huang* et al. [40.21]. Our results are derived under subsets of the following four conditions. We assume that the data from different arrays are independent, and impose conditions on the n individual arrays. Our conditions depend on n only through the uniformity requirements across the n arrays, so that all the theorems in this section hold in the case of fixed $n \geq 2$ as the number of genes $J \to \infty$ as well as the case of $(n, J) \to (\infty, \infty)$ with no constraint on the order of n in terms of J. In contrast, *Huang* and *Zhang* [40.20] focused on applications with large number of arrays. The results in this section hold for any basis functions b_{ik} in (40.5), e.g. spline, Fourier, or wavelet bases, as long as Q_i in (40.9) are projections from $\mathbb{R}^J$ to $\{f(\boldsymbol{x}_i) : f \in S_i\}$ with $Q_i \boldsymbol{e} = \boldsymbol{e}$, where $\boldsymbol{e} = (1, \ldots, 1)'$. Furthermore, with minor modifications in the proof, the results hold when Q_i are replaced by nonnegative definite smoothing matrices A_i with their largest eigenvalues bounded by a fixed constant, see [40.20, 21].

Condition I: *In (40.3), $\boldsymbol{x}_i$, $i = 1, \ldots, n$, are independent random vectors, and for each i $\left(x_{ij}, j \leq J\right)$ are exchangeable random variables. Furthermore, for each $i \leq n$, the space S_i in (40.5) depends on design variables $\left(\boldsymbol{x}_k, z_k, k \leq n\right)$ only through the values of $\boldsymbol{x}_i$ and $(z_k, k \leq n)$.*

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

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

Condition II is satisfied by common designs such as the reference and direct comparison designs. Since $\sum_{i=1}^n Z_n^{-1} z_i z_i' = I_d$, $\sum_{i=1}^n z_i' Z_n^{-1} z_i = d$. In balanced designs or orthogonal designs with replications, $Z_n \propto I_d$, n is a multiplier of d, and $z_i' Z_n^{-1} z_i = \kappa^* = d/n < 1$ for all $i \leq n$. In particular, (40.6) describes a balanced design with $d = 1$, so that condition II holds as long as $n \geq 2$.

Condition III: *For the projections Q_i in (40.9), $K_{J,n}^* \equiv \max_{i \leq n} E\left[\mathrm{tr}(Q_i) - 1\right] = O\left(J^{1/2}\right)$.*

An assumption on the maximum dimensions of the approximation spaces is usually required in nonparametric smoothing. Condition III assumes that the ranks of the projections Q_i are uniformly of the order $O\left(J^{1/2}\right)$ to avoid overfitting, and more important, to avoid colinearity between the approximation spaces for the estimation of $(f_i(\boldsymbol{x}_i), i \leq n)$ and the design variables for the estimation of $\boldsymbol{\beta}$. Clearly, $E\left[\mathrm{tr}(Q_i) - 1\right] \leq K_i$ for the K_i in (40.5).

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

Condition IV demands that the ranges of the projections Q_i be sufficiently large that the approximation errors for $f_i(\boldsymbol{x}_i)$ are uniformly $O(1)$ in an average sense. Although this is the weakest possible condition on Q_i for the consistent estimation of $f_i(\boldsymbol{x}_i)$, the combination of conditions III and IV does require careful selection of spaces S_i in (40.5) and certain condition on the tail probability of x_{ij}. See the two examples in *Huang* et al. [40.21] that illustrate this point.

40.6.1 Distribution of $\hat{\beta}$

We now describe the distribution of $\hat{\boldsymbol{\beta}}$ in (40.9) conditionally on all the covariates and provide an upper bound for the conditional bias of $\hat{\boldsymbol{\beta}}$.

Let $\hat{\Lambda}$ be the information operator in (40.12). Define

$$\tilde{\beta}_{J,n} = -\Pi_{J,n}\boldsymbol{\beta} + \hat{\Lambda}_{J,n}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}(I_J - Q_i) f_i(\boldsymbol{x}_i) z_i'\right], \tag{40.28}$$

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

$$V_{J,n} = \frac{1}{n}\sum_{i=1}^{n} V_i \otimes z_i z_i' ,$$
$$V_i = (I_J - Q_i)\mathrm{var}(\boldsymbol{\epsilon}_i)(I_J - Q_i) . \qquad (40.29)$$

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

$$B^{(k+1)} \leftarrow n(\boldsymbol{b} - \Pi_{J,n}\boldsymbol{b})Z_n^{-1} + \sum_{i=1}^{n} Q_i B^{(k)} z_i z_i' Z_n^{-1} \qquad (40.30)$$

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

Theorem 40.1
Let $\hat{\boldsymbol{\beta}}_{J,n}$, $\hat{\Lambda}$ and $V_{J,n}$ be as in (40.9), (40.12) and (40.29) respectively. Suppose that given $\{\boldsymbol{x}_i, i \le n\}$, $\boldsymbol{\epsilon}_i$ are independent normal vectors. Then, conditionally on $\{\boldsymbol{x}_i, i \le n\}$,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N\left(\tilde{\boldsymbol{b}}_{J,n}, \frac{1}{n}\hat{\Lambda}_{J,n}^{-1} V_{J,n} \hat{\Lambda}_{J,n}^{-1}\right) . \qquad (40.31)$$

In particular, for all $\boldsymbol{b} \in \Omega_0^{J\times d}$, $\lim_{k\to\infty} B^{(k)} = \hat{\Lambda}^{-1}\boldsymbol{b}$ with the $B^{(k)}$ in (40.30), and

$$\sigma_{J,n}^2(\boldsymbol{b}) \equiv \mathrm{var}\left[\mathrm{tr}\left(\boldsymbol{b}'\hat{\boldsymbol{\beta}}\right)\middle|\{\boldsymbol{x}_i, i \le n\}\right]$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} z_i'\left(\hat{\Lambda}^{-1}\boldsymbol{b}\right)' V_i \left(\hat{\Lambda}^{-1}\boldsymbol{b}\right) z_i . \qquad (40.32)$$

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

Theorem 40.2
Suppose conditions I to IV hold. If $c_{J,n}/\rho_{J,n}^* \to \infty$, then

$$\sup\left\{E\min\left(1, \frac{\mathrm{tr}^2\left(\boldsymbol{b}'\tilde{\boldsymbol{b}}_{J,n}\right)}{\mathrm{tr}\left(\boldsymbol{b}Z_n^{-1}\boldsymbol{b}'\right)c_{J,n}}\right) : \boldsymbol{b} \in \Omega_0^{J\times d}, \boldsymbol{b} \neq 0\right\} = O(1). \qquad (40.33)$$

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

$$\sup_{\boldsymbol{b}\in\Omega_0^{J\times d}, \boldsymbol{b}\neq 0}\left\{\sup_{x\in\mathbb{R}}\left|P\left(\frac{\mathrm{tr}\left(\boldsymbol{b}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)}{\sigma_{J,n}(\boldsymbol{b})} \le x\right) - \Phi(x)\right|\right\} = O(1) , \qquad (40.34)$$

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

This result states that, under conditions I to IV, appropriate linear combinations of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, such as contrasts, have an approximate normal distribution with mean zero and the approximation is uniform over all linear combinations. Therefore, this result provides theoretical justification for inference procedures based on $\hat{\boldsymbol{\beta}}$, such as those described in Sect. 40.3. Without the normality condition, (40.29) is expected to hold under the Lindeberg condition as $(n, J) \to (\infty, \infty)$, even in the case $n = O(J)$ [for example $n = O(\log J)$]. We assume the normality here so that (40.29) holds for fixed n as well as large n.

40.6.2 Convergence Rates of Estimated Normalization Curves $\hat{f}_i$

Normalization is not only important in detecting differentially expressed genes, it is also a basic first step for other high-level analysis, including classification and cluster analysis. Thus, it is of interest in itself to study the behavior of the estimated normalization curves. Here we study the convergence rates of $\hat{f}_i$.

Since $\hat{f}_i(\boldsymbol{x}_i) = Q_i(\boldsymbol{y}_i - \hat{\boldsymbol{\beta}}z_i)$, it follows from (40.3) that

$$\hat{f}_i(\boldsymbol{x}_i) = Q_i[f_i(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i] - Q_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})z_i . \qquad (40.35)$$

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

Theorem 40.3
Suppose conditions I to IV hold and $\mathrm{var}(\boldsymbol{\epsilon}_i) \le (\sigma^*)^2 I_J$ for certain $0 < \sigma^* < \infty$. Then, for certain $\epsilon_{J,M}$ with $\lim_{M\to\infty}\lim_{J\to\infty}\epsilon_{J,M} \to 0$,

$$\max_{i\le n} P\left\{\|\hat{f}_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_i)\|^2/J > M\left[\rho_{J,n}^* + (\sigma^*)^2 K_{J,n}^*/J\right]\right\} \le \epsilon_{J,M} .$$

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

In the case of $\mathrm{var}(\boldsymbol{\epsilon}_i) = \sigma^2 I_J$, $\max_{i\le n} E\|Q_i(\boldsymbol{y}_i - \boldsymbol{\beta}z_i) - f_i(\boldsymbol{x}_i)\|^2/J \ge \max\left(\rho^*_{J,n}, \sigma^2 K^*_{J,n}/J\right)$ is the convergence rate for the ideal fits $Q_i(\boldsymbol{y}_i - \boldsymbol{\beta}z_i)$ for $f_i(\boldsymbol{x}_i)$. Theorem 3 asserts that $\hat{f}_i(\boldsymbol{x}_i)$ have the same convergence rates as the ideal fits. Thus, $\hat{f}_i(\boldsymbol{x}_i)$ achieve or nearly achieve the optimal rate of convergence for normalization under standard conditions. See the two examples in *Huang* et al. [40.21].

40.7 Concluding Remarks

The basic idea of TW-SLM normalization is to estimate the normalization curves and the relative gene effects simultaneously. The TW-SLM normalization does not assume that the normalization is constant as in the global normalization method, nor does it make the assumptions that the percentage of differentially expressed genes is small or that the up- and down-regulated genes are distributed symmetrically, as are required in the lowess normalization method [40.11]. This model puts normalization and significant analysis of gene expression in the framework of a high dimensional semiparametric regression model. We used a back-fitting algorithm to compute the semiparametric M-estimators in the TW-SLM. For identification of differentially expressed genes, we used an intensity-dependent variance model. This variance model is a compromise between the constant residual variance assumption used in the ANOVA method and the approach in which the variances of all the genes are treated as being different. We described two nonparametric methods for variance estimation. The first method is to smooth the scatter plot of the squared residuals versus the total intensity. The second one is to estimate the variance function jointly with the normalization curves and gene effects. For the example we considered in Sect. 40.6, the proposed method yields reasonable results when compared with the published results. Our simulation studies show that the TW-SLM normalization has better performance in terms of the mean squared errors than the lowess and spline normalization methods. Thus the proposed TW-SRM for microarray data is a powerful alternative to the existing normalization and analysis methods.

The TW-SLM is qualitatively different from the SRM. For microarray data, the number of genes J is always much greater than the number of arrays n. This fits the description of the well-known small-n large-p difficulty (we use p instead of J to be consistent with the phrase used in the literature). Furthermore, in the TW-SLM, both n and J play the dual role of sample size and number of parameters. That is, for estimating $\boldsymbol{\beta}$, J is the number of parameters, n is the sample size. But for estimating $\boldsymbol{f}$, n is the number of (infinite dimensional) parameters, J is the sample size for each f_i. On one hand, sufficiently large n is needed for the inference of $\boldsymbol{\beta}$. But a large n makes normalization more difficult, because then more nonparametric curves need to be estimated. On the other hand, sufficiently large J is needed for accurate normalization, but then estimation of $\boldsymbol{\beta}$ becomes more difficult. We are not aware of any other semiparametric models [40.33] in which both n and J play such dual roles of sample size and number of parameters. Indeed, here the difference between the sample size and the number of parameters is no longer as clear as in a conventional statistical model. This reflects a basic feature of microarray data in which self-calibration in the data is required when making statistical inference.

References

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

40.2 P.O. Brown, D. Botstein: Exploring the new world of the genome with microarrays, Nat. Genet. **21**(1), 33–37 (1999)

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

40.4 M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein: Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA **95**(25), 14863–14868 (1998)

40.5 T.R. Golub, D.K. Slonim, P: Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander: Molecular classification of cancer:

Class discovery and class prediction by gene expression monitoring, Science **286**(5439), 531–537 (1999)

40.6 A. A. Alizadeh, M. B. Eisen, E. R. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature **403**(6769), 503–511 (2000)

40.7 M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, E. M. Rubin: Microarray expression profiling identifies genes with altered expression in HDL deficient mice, Gen. Res. **10**, 2022–2029 (2000)

40.8 Y. Chen, E. R. Dougherty, M. L. Bittner: Ratio-based decisions and the quantitative analysis of cDNA microarray images, J. Biomed. Opt. **2**, 364–374 (1997)

40.9 M. K. Kerr, M. Martin, G. A. Churchill: Analysis of variance for gene expression microarray data, J. Comp. Biol. **7**, 819–837 (2000)

40.10 T. B. Kepler, L. Crosby, K. T. Morgan: Normalization and analysis of DNA microarray data by self-consistency and local regression, Genome Biol. **3**(7), research0037.1–research0037.12 (2002)

40.11 Y. H. Yang, S. Dudoit, P. Luu, T. P. Speed: Normalization for cDNA microarray data. In: *Microarrays: Optical Technologies and Informatics*, Proceedings of SPIE, Vol. 4266, ed. by M. L. Bittner, Y. Chen, A. N. Dorsel, E. R. Dougherty (Int. Soc. Opt. Eng., San Diego 2001) pp. 141–152

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

40.13 D. B. Finkelstein, J. Gollub, R. Ewing, F. Sterky, S. Somerville, J. M. Cherry: Iterative linear regression by sector. In: *Methods of Microarray Data Analysis. Papers from CAMDA 2002*, ed. by S. M. Lin, K. F. Johnson (Kluwer Academic, Dordrecht 2001) pp. 57–68

40.14 J. Quackenbush: Microarray data normalization and transformation, Nat. Gen. (Suppl.) **32**, 496–501 (2002)

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

40.16 J. Fan, P. Tam, G. Vande Woude, Y. Ren: Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a Cytokine, Proc. Natl. Acad. Sci. **101**, 1135–1140 (2004)

40.17 J. Fan, H. Peng, T. Huang: Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency, J. Am. Stat. Assoc. **100**, 781–796 (2005)

40.18 W. S. Cleveland: Robust locally weighted regression and smoothing scatterplots, J. Am. Stat. Assoc. **74**, 829–836 (1979)

40.19 J. Huang, H.-C. Kuo, I. Koroleva, C.-H. Zhang, M. B. Soares: A Semi-linear Model for Normalization and Analysis of cDNA Microarray Data, Tech Report **321** (2003) Depart. of Stat., Univ. Iowa, Iowa City

40.20 J. Huang, C.-H. Zhang: Asymptotic analysis of a two-way semiparametric regression model for microarray data, Stat. Sin. **15**, 597–618 (2005)

40.21 J. Huang, D. L. Wang, C-H. Zhang: A two-way semilinear model for normalization and significant analysis of microarray data, J. Am. Stat. Assoc. **100**, 814–829 (2005)

40.22 G. Wahba: Partial spline models for semiparametric estimation of functions of several variables. In: *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar Tokyo (Inst. Stat. Mathematics, Tokyo 1984) pp. 319–329

40.23 R. F. Engle, C. W. J. Granger, J. Rice, A. Weiss: Semiparametric estimates of the relation between weather and electricity sales, J. Am. Stat. Assoc. **81**, 310–320 (1986)

40.24 P. Heckman: Spline smoothing in partly linear model, J. R. Stat. Soc. Ser. B **48**, 244–248 (1986)

40.25 H. Chen: Convergence rates for a parametric component in a partially linear model, Ann. Stat. **16**, 136–146 (1988)

40.26 P. Huber: *Robust Statistics* (Wiley, New York 1981)

40.27 L. Schumaker: *Spline Functions: Basic Theory* (Wiley, New York 1981)

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

40.29 R Development Core Team: *R: A Language and Environment for Statistical Computing* (R Foundation Stat. Computing, Vienna 2003) http//www.R-project.org.

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

40.31 J. Fan, Q. Yao: Efficient estimation of conditional variance functions in stochastic regression, Biometrika **85**, 645–660 (1998)

40.32 S. Dudoit, Y. H. Yang, M. J. Callow, T. P. Speed: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Stat. Sin. **12**, 111–140 (2000)

40.33 P. J. Bickel, C. A. J. Klaassen, Y. Ritov, J. A. Wellner: *Efficient and Adaptive Estimation for Semiparametric Models* (Johns Hopkins Univ. Press, Baltimore 1993)