Published in final edited form as:

Biometrika. 2008 ; 95(2): 437–449. doi:10.1093/biomet/asn017.

Nonparametric variance estimation in the analysis of microarray data: a measurement error approach

RAYMOND J. CARROLL and

*Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.,
carroll@stat.tamu.edu*

YUEDONG WANG

*Department of Statistics and Applied Probability, University of California, Santa Barbara, California
93106, U.S.A. yuedong@pstat.ucsb.edu*

Summary

This article investigates the effects of measurement error on the estimation of nonparametric variance functions. We show that either ignoring measurement error or direct application of the simulation extrapolation, SIMEX, method leads to inconsistent estimators. Nevertheless, the direct SIMEX method can reduce bias relative to a naive estimator. We further propose a permutation SIMEX method which leads to consistent estimators in theory. The performance of both SIMEX methods depends on approximations to the exact extrapolants. Simulations show that both SIMEX methods perform better than ignoring measurement error. The methodology is illustrated using microarray data from colon cancer patients.

Some key words

Heteroscedasticity; Local polynomial regression; Measurement error; Microarray; Nonparametric regression; Permutation; Simulation-extrapolation; Variance function estimation

1. Introduction

Microarray experiments quantify expression levels on a global scale by measuring transcript abundance of thousands of genes simultaneously (Nguyen et al., 2002; Leung & Cavalieri, 2003). One of the important problems in the analysis of microarray data is to detect differentially expressed genes. Methods such as the t -test are routinely used to provide formal statistical inference. The number of replications is typically small, and this leads to unreliable variance estimators and low power of conventional statistical methods (Callow et al., 2000; Cui et al., 2005).

It has been observed that the variance increases proportionally with the intensity level. Many authors used this property to build parametric (Chen et al., 1997; Rocke & Durbin, 2001; Chen et al., 2002; Weng et al., 2006) and nonparametric (Kamb & Ramaswami, 2001; Huang & Pan, 2002; Lin et al., 2003; Jain et al., 2003) variance-mean models. Borrowing information from different genes with similar variances, these new variance estimators are more reliable and lead to more powerful tests. Nevertheless, two subtle issues have not been addressed in the literature. First, for the purpose of estimating the variance function, the mean responses of all the genes represent a large number of unknown nuisance parameters. Therefore, care needs to be taken to derive consistent estimators of the variance function. One simple approach is to fit a variance-mean model based on reduced data consisting of sample means and variances (Huang & Pan, 2002). The second subtle issue is that, because of sampling error, naive application of

nonparametric regression methods to the reduced data leads to inconsistent estimators. In Wang et al. (2006), we investigated the effects of measurement error on estimating parameters in parametric variance models and proposed consistent estimators. This article investigates the nonparametric case.

We consider the general one-way analysis of variance model,

$$Y_{i,j} = X_i + g^{1/2}(X_i)\varepsilon_{i,j}, \quad i=1, \dots, n; \quad j=1, \dots, m, \quad (1)$$

where $Y_{i,j}$ is the j th replicate at level i with mean X_i and variance $g(X_i)$, and random errors $\varepsilon_{i,j} \sim N(0, 1)$. The variance is a function of the mean and our goal is to estimate the variance-mean function g nonparametrically.

2. Measurement error, its impact and methods

2.1. When measurement error is ignored

For ease of exposition, in this section we consider locally constant regression estimation only. We expect there to be a similar impact of measurement error on more complicated

nonparametric regression methods. Let $\bar{Y}_{i\cdot} = m^{-1} \sum_{j=1}^m Y_{i,j}$ and $S_i = (m-1)^{-1} \sum_{j=1}^m (Y_{i,j} - \bar{Y}_{i\cdot})^2$ be sample means and variances, respectively. Since S_i is an unbiased estimator of $g(X_i)$, if X were observable a locally constant regression estimator is

$$\hat{g}(x_0) = \frac{n^{-1} \sum_{i=1}^n K_h(X_i - x_0) S_i}{n^{-1} \sum_{i=1}^n K_h(X_i - x_0)}, \quad (2)$$

where $K(\cdot)$ is a symmetric density function, h is the bandwidth and $K_h(v) = h^{-1}K(v/h)$.

Note that $\hat{g}(x_0)$ cannot be computed because the X -values are unobservable. Since $\bar{Y}_{i\cdot}$ is an unbiased estimator of X_i , one may replace X_i in (2) by $\bar{Y}_{i\cdot}$. This is equivalent to fitting a locally constant regression model to the reduced data $\{(\bar{Y}_{i\cdot}, S_i)\}$ (Huang & Pan, 2002). Denote the resulting estimator by $\hat{g}_N$, and note that the measurement error in $\bar{Y}_{i\cdot}$ as an estimator of X_i has been ignored. One natural question is whether or not $\hat{g}_N$ is consistent. In the remainder of this article, we assume that the X_i are independent and identically distributed with density function $f_X(\cdot)$.

To illustrate the potential effects of the measurement error, we generate data from model (1) with $g(x) = x^2/2$, $X \sim \text{Uni}[0, 1]$, $m = 9$ and $n = 250$. Figure 1 shows the true function, observations and the naive estimator $\hat{g}_N$. Obviously the measurement error makes observations more widely spread and thus causes attenuation in the naive estimator $\hat{g}_N$. This kind of behaviour is typical. In fact, the following result shows that $\hat{g}_N$ is usually not consistent.

Theorem 1—Under standard regularity conditions such as $h \rightarrow 0$ and $nh/\log(n) \rightarrow \infty$,

$$\hat{g}_N(x_0) \rightarrow \frac{\int g^{1/2}(x_0 + m^{-1/2}s) \varphi\{s/g^{1/2}(x_0 + m^{-1/2}s)\} f_X(x_0 + m^{-1/2}s) ds}{\int \varphi\{s/g^{1/2}(x_0 + m^{-1/2}s)\} f_X(x_0 + m^{-1/2}s) / g^{1/2}(x_0 + m^{-1/2}s) ds}, \quad (3)$$

where the convergence is in probability and φ is the density of the standard Normal distribution.

The right-hand side of (3) usually does not equal $g(x_0)$, which indicates that $\hat{g}_N$ is not consistent. It is not difficult to see that for a finite m , the necessary and sufficient condition for $\hat{g}_N$ to be consistent for all f_X is $g(x) \equiv g_0$, a constant variance function. It is obvious that the right-hand side of (3) converges to $g(x_0)$ as $m \rightarrow \infty$.

For the case that $g(x) = x^2/2$, $X \sim \text{Uni}[0, 1]$ and $m = 3, 9, 15$, we plot the limiting functions of naive estimators in Fig. 2(a). The bias decreases slowly as m increases and is not ignorable even when $m = 15$.

2.2. Direct application of the SIMEX algorithm

One popular approach for dealing with measurement error is the SIMEX method (Cook & Stefanski, 1995; Carroll et al., 2006), modified to account for heteroscedasticity (Devanarayan & Stefanski, 2002). We adapt the SIMEX method to estimate g as follows.

Step 1—Generate $Z_{b,i,j} \sim N(0, 1)$ independently, $i = 1, \dots, n, j = 1, \dots, m, b = 1, \dots, B$, where B is large. In general, B in the range 50–100 is large enough to eliminate simulation variability. Let

$$c_{b,i,j} = \frac{Z_{b,i,j} - \bar{Z}_{b,i,\cdot}}{\{\sum_{j=1}^m (Z_{b,i,j} - \bar{Z}_{b,i,\cdot})^2\}^{1/2}},$$

$$W_{b,i}(\zeta) = \bar{Y}_{i,\cdot} + (\zeta/m)^{1/2} \sum_{j=1}^m c_{b,i,j} Y_{i,j}.$$

Step 2—Apply a nonparametric regression method to the reduced data $\{W_{b,i}(\zeta), S_i\}$ for each $b = 1, \dots, B$ and then average over b . Denote the resulting estimator by $\hat{g}_S(\cdot, \zeta)$.

Step 3—Extrapolate $\hat{g}_S(\cdot, \zeta)$ back to $\zeta = -1$.

We call the above method ‘direct SIMEX’. Now suppose that we use a locally constant estimator in Step 2; that is, we replace X_i in (2) by $W_{b,i}(\zeta)$. Denote the resulting estimator after extrapolation by $\hat{g}_S(\cdot) = \lim_{\zeta \rightarrow -1} \hat{g}_S(\cdot, \zeta)$ and define

$$H_s(t) = f_X(t) g^{1/2}(t) \varphi\{s/g^{1/2}(t)\} \{s^2/g(t) - 1\}.$$

From standard results, it follows that $\hat{g}_S(\cdot, \zeta) = g_S(\cdot, \zeta) + o_p(1)$, where $g_S(\cdot, \zeta)$ corresponds to regression of S_i on $W_{b,i}(\zeta)$.

Theorem 2—Under standard regularity conditions such as $h \rightarrow 0$ and $nh/\log(n) \rightarrow \infty$, in the limit as $\zeta \rightarrow -1$,

$$\lim_{\zeta \rightarrow -1} g_S(x_0, \zeta) = g_S(x_0, -1) = g(x_0) - \frac{1}{2m(m-1)f_X(x_0)} \int s^2 H_s^{(2)}(x_0) ds. \quad (4)$$

Usually the asymptotic bias $-\int s^2 H_s^{(2)}(x_0) ds / \{2m(m-1)f_X(x_0)\}$ does not equal zero. Therefore, direct application of the SIMEX method also leads to an inconsistent estimator. In the case of

a constant variance function, $g(x) \equiv g_0$, $\hat{g}_S$ approaches $g(x_0) - f_x^{(2)}(x_0)/\{m(m-1)f_x(x_0)\}$. In this case, $\hat{g}_S$ underestimates at points where $f_x^{(2)}(x_0) > 0$ and overestimates at points where $f_x^{(2)}(x_0) < 0$.

For the case of $g(x) = x^2/2$, $X \sim \text{Uni}[0, 1]$ and $m = 3, 9, 15$, we plot the limiting functions of the direct SIMEX estimator in Fig. 2(b). The bias decreases quickly as m increases and is relatively small even when $m = 9$. Surprisingly, for $m = 3$, the direct SIMEX method actually leads to an estimator which is more biased than a naive estimator.

2.3. Permutation SIMEX method

The direct SIMEX method is not consistent in the ideal case in which the exact extrapolation function is known. Our permutation SIMEX method overcomes this problem for parametric variance functions (Wang et al., 2006). We now describe permutation SIMEX for nonparametric estimation of the variance function.

Step 1—For $j = 1, \dots, m$,

- a. generate $Z_{b,i,k} \sim N(0, 1)$ independently, $i = 1, \dots, n$, $k = 1, \dots, m-1$, $b = 1, \dots, B$, and let

$$c_{b,i,k}^{(j)} = \frac{Z_{b,i,k} - \bar{Z}_{b,i,\cdot}}{\{\sum_{k=1}^{m-1} (Z_{b,i,k} - \bar{Z}_{b,i,\cdot})^2\}^{1/2}};$$

- b. for $i = 1, \dots, n$, $k = 1, \dots, m-1$, $b = 1, \dots, B$, let

$$\begin{aligned} W_{i,k}^{(j)} &= \begin{cases} Y_{i,k} & 1 \leq k \leq j-1, \\ Y_{i,k+1} & j \leq k \leq m-1, \end{cases} \\ W_{b,i}^{(j)}(\zeta) &= \bar{W}_{i,\cdot}^{(j)} + \{\zeta/(m-1)\}^{1/2} \sum_{k=1}^{m-1} c_{b,i,k} W_{i,k}^{(j)}, \\ S_i^{(j)} &= \{Y_{i,j} - W_{b,i}^{(j)}(\zeta)\}^2. \end{aligned}$$

Step 2—Apply a nonparametric regression method to the reduced data $\{W_{b,i}^{(j)}(\zeta), S_i^{(j)}\}$ for each combination of j and b , and then average over all j and b . For robustness, a transformation to $S_i^{(j)}$ may be applied; see §3 for an example. Denote the resulting estimator by $\hat{g}_{PS}(\cdot, \zeta)$.

Step 3—Extrapolate $\hat{g}_{PS}(\cdot, \zeta)$ back to $\zeta = -1$.

Suppose that we use a locally constant estimator in Step 2. For any fixed b , as $n \rightarrow \infty$ and $h \rightarrow 0$, $\hat{g}_{PS}(\cdot, \zeta) = g_{PS}(\cdot, \zeta) + o_p(1)$, where $g_{PS}(\cdot, \zeta)$ corresponds to regression of $\{Y_{i,j} - W_{b,i}^{(j)}(\zeta)\}^2$ on $W_{b,i}^{(j)}(\zeta)$. In the ideal case in which we know the extrapolation function, the permutation SIMEX method correctly yields the function of interest $g(x_0)$.

Theorem 3—Under standard regularity conditions such as $h \rightarrow 0$ and $nh/\log(n) \rightarrow \infty$, in the limit as $\zeta \rightarrow -1$, $g_{PS}(x_0, \zeta) \rightarrow g(x_0)$.

3. Asymptotic results for permutation simex

In this section, for robustness and flexibility in practice, we consider a locally linear regression estimator with a general estimating function, and we derive the asymptotic bias and variance for this estimator. Consider a kernel estimator for Step 2 in the permutation SIMEX method. Ignoring subscripts, if X were known, our original estimating function would have been $(Y - X)^2 - g(X)$. The problem with this approach is that one outlier will wreak havoc. Let $\lambda > 0$ and define $a_\lambda = E(|\varepsilon|^{2\lambda})$. Then another estimating function is

$$(Y - X)^{2\lambda} - a_\lambda g^\lambda(X). \quad (5)$$

In general, $\lambda = 1/3$ or $1/2$ adds more robustness with little loss of efficiency. We consider general estimating functions of the form

$$Q\{Y, X, g(X)\}, \quad (6)$$

which includes (5) with a fixed λ as a special case. The general estimating function $Q\{Y, X, g(X)\}$ can be used to build robust M -estimators.

Define $g_{PS}(x_0, \zeta)$ to be the solution to

$$0 = E[Q\{Y, W_b(\zeta), g_{PS}(x_0, \zeta)\} | W_b(\zeta) = x_0]. \quad (7)$$

For example, when Q equals (5) with $\lambda = 1$, we have

$$g_{PS}(x_0, \zeta) = E\{(Y - x_0)^2 | W_b(\zeta) = x_0\} = \int \{(x - x_0)^2 + g(x)\} f_{X|W}(x, \zeta, g) dx,$$

where $f_{X|W}(x_0, \zeta, g)$ is the conditional density of X given $W_b(\zeta) = x_0$. Since $f_{X|W}(\cdot, \zeta, g)$ depends on ζ and g in a complicated manner, the exact extrapolant $g_{PS}(x_0, \zeta)$ usually does not have a closed form. Nevertheless, since $W_b(\zeta) = X + g^{1/2}(X)\Omega$, where $\Omega \sim N\{0, (1 + \zeta)/(m - 1)\}$, it is easy to see that $g_{PS}(x_0, -1) = g(x_0)$, thus generalizing Theorem 3 to the estimating-function-based method (7).

The asymptotic theory for the permutation SIMEX estimator uses similar calculations to those in Carroll et al. (1999), but the actual details are quite different because the SIMEX algorithm we use is very different from that used by them.

For $b = 1, \dots, B$ and $j = 1, \dots, m$, define $\hat{g}_{b,j}(x_0, \zeta)$ to be the locally linear estimating equation estimator in Step 2, i.e., the solution α_0 to

$$\begin{aligned} 0 = & n^{-1} \sum_{i=1}^n K_h\{W_{b,i}^{(j)}(\zeta) - x_0\} \left[1, \{W_{b,i}^{(j)}(\zeta) - x_0\}/h \right]^T \\ & \times Q[Y_{i,j}, W_{b,i}^{(j)}(\zeta), \alpha_0 + \alpha_1 \{W_{b,i}^{(j)}(\zeta) - x_0\}/h]. \end{aligned} \quad (8)$$

Then

$$\widehat{g}_{ps}(x_0, \zeta) = (Bm)^{-1} \sum_{b=1}^B \sum_{j=1}^m \widehat{g}_{b,j}(x_0, \zeta). \quad (9)$$

Let $\tau = \int x^2 K(x) dx$, $\gamma = \int K^2(x) dx$ and

$$v(x_0, \zeta) = -f_W(x_0, \zeta) E \left[Q_g \{Y_{1,1}, W_{b,1}^{(1)}(\zeta), \widehat{g}_{b,1}(x_0, \zeta)\} | W_{b,1}^{(1)}(\zeta) = x_0 \right],$$

where $f_W(\cdot, \zeta)$ is the marginal density of $W_{b,1}^{(1)}(\zeta)$ and $Q_g(\cdot, \cdot, g) = (\partial/\partial g)Q(\cdot, \cdot, g)$. Let $g_{ps}^{(2)}(x_0, \zeta) = (\partial^2/\partial x_0^2)g_{ps}(x_0, \zeta)$. For any fixed b and j , we have the following asymptotic expansion (Carroll et al., 1998):

$$\begin{aligned} \widehat{g}_{b,j}(x_0, \zeta) &= g_{ps}(x_0, \zeta) + (\tau h^2/2) g_{ps}^{(2)}(x_0, \zeta) \\ &+ \{nv(x_0, \zeta)\}^{-1} \sum_{i=1}^n K_h \{W_{b,i}^{(j)}(\zeta) - x_0\} Q[Y_{i,j}, W_{b,i}^{(j)}(\zeta), g_{ps} \{W_{b,i}^{(j)}(\zeta), \zeta\}] \\ &+ o_p\{h^2 + \log(n)(nh)^{-1/2}\}. \end{aligned} \quad (10)$$

Then, for fixed B ,

$$\begin{aligned} \widehat{g}_{ps}(x_0, \zeta) &= g_{ps}(x_0, \zeta) + (\tau h^2/2) g_{ps}^{(2)}(x_0, \zeta) \\ &+ \{Bmn v(x_0, \zeta)\}^{-1} \sum_{b=1}^B \sum_{j=1}^m \sum_{i=1}^n K_h \{W_{b,i}^{(j)}(\zeta) - x_0\} Q[Y_{i,j}, W_{b,i}^{(j)}(\zeta), g_{ps} \{W_{b,i}^{(j)}(\zeta), \zeta\}] \\ &+ o_p\{h^2 + \log(n)(nh)^{-1/2}\}. \end{aligned} \quad (11)$$

The expansion (11) is justified as long as B is finite. We will use (11) as our departure point.

In practice, since the ideal exact extrapolant function $g_{ps}(x, \zeta)$ is unknown, we use an approximate extrapolant. Let $\zeta_0, \dots, \zeta_K$ be a grid of ζ values with $\zeta_0 = 0$. For simplicity, we consider linear estimators of the form

$$\widehat{g}_{ps}(x_0, -1) = \sum_{k=0}^K d_k \widehat{g}_{ps}(x_0, \zeta_k), \quad (12)$$

where the d_k 's depend on grid $\zeta_0, \dots, \zeta_K$ only. Obviously (12) includes polynomial extrapolants as special cases.

Remark 1

The case $m = 3$ differs from that of $m > 3$ because, when $m = 3$, $c_{b,i,k}^{(j)} = \pm 2^{-1/2}$. Therefore, $W_{b,i}(\zeta)$ is a discrete random variable which takes on two values $W_{b,i}(\zeta) = \bar{W}_{i,1} \pm \zeta^{1/2}(\bar{W}_{i,1} - \bar{W}_{i,2})/2$, each with probability $1/2$. This causes a technical difference between $m = 3$ and $m > 3$ that is reflected in the asymptotic variance.

Remark 2

As in Carroll et al. (1999), our asymptotics are for fixed B . In those asymptotics, there is a term of order $O\{(Bnh)^{-1}\}$. Since B can be made as large as desired, this term is insignificant in practical contexts, and will be ignored in the statement of the main result.

We define

$$\begin{aligned}\beta(x_0, \zeta) &= \text{var}(Q[Y_i, W_i(s, \zeta), g_{PS}\{W_i(s, \zeta), \zeta\}] | W_i(s, \zeta) = x_0), \\ R(x_0) &= \frac{f_w(x_0, 0)}{v^2(x_0, 0)} \beta(x_0, 0).\end{aligned}$$

Theorem 4

Under standard regularity conditions such as $h \rightarrow 0$, $nh/\log(n) \rightarrow \infty$ and $B \rightarrow \infty$, the leading term in the asymptotic bias is

$$\text{bias}\{\widehat{g}_{PS}(x_0, -1)\} = \sum_{k=0}^K d_k g_{PS}(x_0, \zeta_k) - g(x_0) + (\tau h^2/2) \sum_{k=0}^K d_k g_{PS}^{(2)}(x_0, \zeta_k). \quad (13)$$

When $m > 3$, the leading term in the asymptotic variance is given by

$$\text{var}\{\widehat{g}_{PS}(x_0, -1)\} = \gamma d_0^2 R(x_0) (nmh)^{-1}. \quad (14)$$

When $m = 3$, the leading term in the asymptotic variance is given by

$$\text{var}\{\widehat{g}_{PS}(x_0, -1)\} = \frac{\gamma}{nh} \left(\frac{f_w(x_0, 0) \beta(x_0, 0) d_0^2}{v^2(x_0, 0)} + \frac{1}{2} \sum_{k=1}^K \frac{f_w(x_0, \zeta_k) \beta(x_0, \zeta_k) d_k^2}{v^2(x_0, \zeta_k)} \right). \quad (15)$$

In theory, the permutation SIMEX method leads to a consistent estimator. However, in practice, usually the linear estimator (12) is not an exact extrapolant, and this leads to an asymptotic

approximation bias $\sum_{k=0}^K d_k g_{PS}(x_0, \zeta_k) - g(x_0)$.

4. Application

Alon et al. (1999) collected colon adenocarcinoma tissues from patients and paired normal colon tissue from some of these patients. Gene expression in 40 tumour and 22 normal colon tissue samples was analyzed with an Affimetrix oligonucleotide array complementary to more than 6500 human genes. The dataset contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues.

To remove possible artifacts due to arrays, as in Huang & Pan (2002), observations on each array are standardized by subtracting the median expression level and dividing by the interquartile range of the expression levels on that array. To avoid negative values in the expression level, we then subtract the smallest value across all tumours and all genes from the dataset. Huang & Pan (2002) used these data to show that better estimation of variances

increases detection power. We limit ourselves to the estimation of the variance function in this article.

We randomly selected a subset with five tumours, tumours 3, 7, 10, 16 and 38, and a subset with ten tumours, tumours 3, 8, 11, 16, 17, 21, 22, 23, 27 and 37. Figure 3 plots sample variance versus sample mean, naive estimators, direct SIMEX estimators and permutation SIMEX estimators. We used a locally linear regression estimator with estimating function (5), $\lambda = 1/2$, and bandwidth $h = 0.8$. For comparison, we also fitted the full data with all 40 tumours. Since it is based on a relatively large number of replications, the fit with the full data can be treated as a benchmark. Modifications to the naive estimators have been made by the SIMEX methods in the regions where the variances are not small. This indicates that, in practice, biases in regions where the variances are large may become nonnegligible and corrections by the SIMEX methods may be necessary. We note that, rather than offering formal conclusions, the fits here serve the purpose of illustration only.

5. Simulations

To evaluate the finite-sample performance of the proposed methods, we used the colon data in §4 to create simulation settings. We computed sample means and variances from all 40 tumour samples. Let $\mathcal{X}$ be the collection of sample means with a few very large values excluded. Based on these sample means and variances, we estimated the variance function nonparametrically using the estimating function (5) with $\lambda = 1/2$ and a penalized spline estimator (Ruppert et al., 2003). Denote the estimated variance function by $\tilde{g}$. Let $\mathcal{R}$ be centred and scaled residuals of model (1) where the X_i 's were replaced by sample means and g was replaced by $\tilde{g}$. Figure 4 shows the histogram and density estimate of $\mathcal{X}$, the function $\tilde{g}$ and the Q-Q plot of $\mathcal{R}$. Note that the distribution of $\mathcal{R}$ is asymmetric and has a heavy right tail. The direct and permutation SIMEX methods assume normality of the random errors, even though it is known that the SIMEX approach is relatively robust to modest departures from the normality assumption. Therefore, the assumptions made in our theory do not hold and this simulation provides a challenge to our methods.

We generated data from model (1) with X_i sampled with replacement from $\mathcal{X}$, $g = \tilde{g}$, and $\varepsilon_{i,j}$ sampled with replacement from $\mathcal{R}$. We used a factorial design with $n = 1000, 2000, 3000$ and $m = 3, 6, 9$. For all simulations, we set $B = 50$ and repeated the simulations 500 times.

We used the estimating function (5) with $\lambda = 1/2$. The local linear regression estimator is implemented using the `locpoly` function in the R package `KernSmooth`. The `dpill` function for selecting a data-driven bandwidth sometimes failed, and we therefore fixed the bandwidth at 0.25. Based on our experience, a linear function was used as the extrapolant.

Define the integrated mean squared error as $\text{IMSE} = \int E\{\hat{g}(x) - g(x)\}^2 f_X(x) dx$, where f_X is the density function of X . We used a grid of $K = 100$ equally spaced points in the range of $\mathcal{X}$ to perform extrapolation. We approximate IMSE by

$\text{MSE} = (nK)^{-1} \Delta \sum_{k=1}^K \sum_{i=1}^n \{\hat{g}(x_i) - g(x_i)\}^2 \tilde{f}_X(x_i)$ where the x_i 's are grid points, Δ is the grid length, and $\tilde{f}_X$ is the kernel density estimate shown in Fig. 4(a). The squared biases and variances are defined and approximated similarly. Table 1 lists squared biases, variances and values of MSE. The direct SIMEX method reduces both bias and variance in the naive estimator for all situations. The permutation SIMEX reduces variances even further. It also reduces bias in the naive estimator except when both n and m are large. Note that biases in the naive and direct SIMEX method are inherent, see Theorems 1 and 2, while biases in the permutation SIMEX are caused by the approximation to the unknown exact extrapolant function. The bias in the naive and direct SIMEX methods decreases as m increases. On the other hand, the bias in the permutation SIMEX method decreases quite slowly as n and/or m increase. When both n and

m are large, the values of MSE for the permutation SIMEX method is dominated by bias caused by the approximate extrapolant function. Nevertheless, with the current simple linear extrapolant function, in terms of MSE, the permutation SIMEX method performs better than or as well as the direct SIMEX method when $n = 1000$ and $n = 2000$. The performance of the permutation SIMEX method may be improved further by using better extrapolant functions.

6. Discussion

The performance of both SIMEX methods depends on the approximate extrapolants used in practice and their potential may not be fully realized; more research on better extrapolants is necessary. Nevertheless, even with simple polynomial extrapolants, both SIMEX methods perform better than the naive application of nonparametric regression methods. In practice, one may compare performances of various methods using simulations similar to those in §5, and select a method accordingly.

A referee suggested that caution needs to be exercised when using the estimated variances to construct test statistics for detecting differentially expressed genes. It is possible that the mean-variance relationship holds on average, but does not provide a reliable method for estimating variances. There are other methods for improving the estimation of variances without assuming a mean-variance relationship (Baldi & Long, 2001; Lönnstedt & Speed, 2002; Smyth, 2004; Tong and Wang, 2007). Comparisons between difference methods remains an important future research topic.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Carroll's research was supported by grants from the U.S. National Cancer Institute, and by the Texas A&M Center for Environmental and Rural Health via a grant from the U.S. National Institute of Environmental Health Sciences. Wang's research was supported by a grant from the National Science Foundation. We thank the editor and two referees for constructive comments that substantially improved an earlier draft.

References

- Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering of tumour and normal colon tissues probed by oligonucleotide arrays. *Proc Nat Acad Sci* 1999;96:6745–50. [PubMed: 10359783]
- Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 2001;17:509–19. [PubMed: 11395427]
- Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res* 2000;10:2022–29. [PubMed: 11116096]
- Carroll RJ, Maca JD, Ruppert D. Nonparametric regression in the presence of measurement error. *Biometrika* 1999;86:541–54.
- Carroll RJ, Ruppert D, Welsh AH. Local estimating equation. *J Am Statist Assoc* 1998;93:214–27.
- Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, C. *Measurement Error in Nonlinear Models: A Modern Perspective*. Vol. 2. New York: Chapman & Hall; 2006.
- Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Optics* 1997;2:364–74.
- Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM. Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics* 2002;18:1207–15. [PubMed: 12217912]

- Cook J, Stefanski LA. A simulation extrapolation method for parametric measurement error models. *J Am Statist Assoc* 1995;89:1314–28.
- Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 2005;6:59–75. [PubMed: 15618528]
- Devanarayan V, Stefanski LA. Empirical simulation extrapolation for measurement error models with replicate measurements. *Statist Prob Lett* 2002;59:219–25.
- Huang X, Pan W. Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Funct Integr Genomics* 2002;2:126–33. [PubMed: 12185460]
- Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee J. Local-pooled error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* 2003;19:1945–51. [PubMed: 14555628]
- Kamb A, Ramaswami A. A simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnol* 2001;1:1–8. [PubMed: 11316465]
- Leung Y, Cavalieri D. Fundamentals of cDNA microarray data analysis. *TRENDS Genet* 2003;11:649–59. [PubMed: 14585617]
- Lin, Y.; Nadler, ST.; Attie, AD.; Yandell, BS. Adaptive gene picking with microarray data: detecting important low abundance signals. In: Parmigiani, G.; Garrett, ES.; Irizarry, RA.; Zeger, SL., editors. *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer; 2003. p. 291–312.
- Lönnstedt I, Speed T. Replicated microarray data. *Statist Sinica* 2002;12:31–46.
- Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA microarray experiments: Biological and technological aspects. *Biometrics* 2002;58:701–17. [PubMed: 12495124]
- Rocke DM, Durbin B. A model for measurement error for gene expression arrays. *J Comp Biol* 2001;8:557–69.
- Ruppert, D.; Wand, M.; Carroll, R. *Semiparametric Regression*. New York: Cambridge University Press; 2003.
- Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist Appl Genet Mol Biol* 2004;3:Article 3.
- Tong T, Wang Y. Optimal shrinkage estimation of variances with applications to microarray data analysis. *J Am Statist Assoc* 2007;102:113–22.
- Wang, Y.; Ma, Y.; Carroll, RJ. Variance estimation in the analysis of microarray data. 2006. Available at <http://www.pstat.ucsb.edu/faculty/yuedong>
- Weng L, Dai H, Zhan Y, He Y, Stepaniants SB, Bassett DE. Rosetta error model for gene expression analysis. *Bioinformatics* 2006;22:1111–21. [PubMed: 16522673]

Appendix

Proofs

This Appendix includes proof sketches. Derivations that are largely algebraic in nature are included in supplemental materials available from the first author.

Proof of Theorem 1

Note that $E(S_i|X_i, \bar{Y}_i) = g(X_i)$. Therefore,

$$E\{K_h(\bar{Y}_{i\cdot} - x_0)S_i\} = E[E\{K_h(\bar{Y}_{i\cdot} - x_0)S_i|X_i, \bar{Y}_{i\cdot}\}] = E\{K_h(\bar{Y}_{i\cdot} - x_0)g(X_i)\}.$$

It can be shown that

$$\begin{aligned}
& \lim_{h \rightarrow 0} E \{K_h(\bar{Y}_{i.} - x_0)g(X_i)\} \\
&= \int g^{1/2}(x_0 + m^{-1/2}s) \varphi \{s/g^{1/2}(x_0 + m^{-1/2}s)\} f_X(x_0 + m^{-1/2}s) ds, \\
& \lim_{h \rightarrow 0} E \{K_h(\bar{Y}_{i.} - x_0)\} \\
&= \int \varphi \{s/g^{1/2}(x_0 + m^{-1/2}s)\} f_X(x_0 + m^{-1/2}s) / g^{1/2}(x_0 + m^{-1/2}s) ds.
\end{aligned} \tag{A1}$$

The theorem follows from the fact that, as $nh/\log(n) \rightarrow \infty$ and $h \rightarrow 0$,

$$\widehat{g}_N(x_0) \rightarrow \frac{\lim_{h \rightarrow 0} E \{K_h(\bar{Y}_{i.} - x_0)g(X_i)\}}{\lim_{h \rightarrow 0} E \{K_h(\bar{Y}_{i.} - x_0)\}}.$$

Proof of Theorem 2

Detailed calculations available in the supplementary material show that

$$E\{S_i|X_i, W_{b,i}(\zeta)\} = g(X_i) + s(\zeta) \left[\frac{\{W_{b,i}(\zeta) - X_i\}^2}{(1+\zeta)/m} - g(X_i) \right], \tag{A2}$$

where $s(\zeta) = \zeta/(m-1)(1+\zeta)$. Then

$$\begin{aligned}
& E[K_h\{W_{b,i}(\zeta) - x_0\}S_i] = E(E[K_h\{W_{b,i}(\zeta) - x_0\}S_i|W_{b,i}(\zeta), X_i]) \\
&= E[K_h\{W_{b,i}(\zeta) - x_0\}g(X_i)] \\
&+ s(\zeta)E \left(K_h\{W_{b,i}(\zeta) - x_0\} \left[\frac{\{W_{b,i}(\zeta) - X_i\}^2}{(1+\zeta)/m} - g(X_i) \right] \right).
\end{aligned}$$

Detailed calculations show that

$$\begin{aligned}
& \lim_{\zeta \rightarrow -1} \lim_{h \rightarrow 0} E[K_h\{W_{b,i}(\zeta) - x_0\}g(X_i)] = f_X(x_0)g(x_0), \\
& \lim_{\zeta \rightarrow -1} \lim_{h \rightarrow 0} s(\zeta)E \left(K_h\{W_{b,i}(\zeta) - x_0\} \left[\frac{\{W_{b,i}(\zeta) - X_i\}^2}{(1+\zeta)/m} - g(X_i) \right] \right)
\end{aligned} \tag{A3}$$

$$= - \frac{1}{2m(m-1)} \int s^2 H_s^{(2)}(x_0) ds. \tag{A4}$$

Note that, by replacing $g(x)$ by 1.0 in (A3), we obtain

$$\lim_{\zeta \rightarrow -1} \lim_{h \rightarrow 0} E[K_h\{W_{b,i}(\zeta) - x_0\}] = f_X(x_0). \tag{A5}$$

Putting all this together, we have shown that, as $nh/\log(n) \rightarrow \infty$ and $h \rightarrow 0$,

$$\begin{aligned}\widehat{g}_s(x_0) &\rightarrow \frac{\lim_{h \rightarrow 0} E [K_h \{W_{b,i}(\zeta) - x_0\} S_i]}{\lim_{h \rightarrow 0} E [K_h \{W_{b,i}(\zeta) - x_0\}]} \\ &= g(x_0) - \frac{1}{2m(m-1)f_X(x_0)} \int s^2 H_s^{(2)}(x_0) ds.\end{aligned}$$

Proof of Theorem 3

This follows by the same arguments as in the proof of Theorem 2.

Proof of Theorem 4

Equation (13) is a consequence of (11) and (12). We thus only need to show (14)–(15). Let $\tilde{Y} = (Y_{1,1}, \dots, Y_{1,m}, \dots, Y_{n,1}, \dots, Y_{n,m})$ be the collection of all observations, and define

$$\widehat{\Lambda}(x_0) = \widehat{g}_{ps}(x_0) - \sum_{k=0}^K d_k \left\{ g_{ps}(x_0, \zeta_k) + (\tau h^2/2) g_{ps}^{(2)}(x_0, \zeta_k) \right\}.$$

The asymptotically equivalent form for this comes from (11) and is given as

$$\begin{aligned}\widehat{\Lambda}(x_0) &= \sum_{k=0}^K d_k \{Bmn v(x_0, \zeta_k)\}^{-1} \\ &\times \sum_{b=1}^B \sum_{j=1}^m \sum_{i=1}^n K_h \{W_{b,i}^{(j)}(\zeta_k) - x_0\} Q[Y_{i,j}, W_{b,i}^{(j)}(\zeta_k), g_{ps}\{W_{b,i}^{(j)}(\zeta_k), \zeta_k\}].\end{aligned}$$

Of course,

$$\text{var}\{\widehat{\Lambda}(x_0)\} = \text{var} \left[E\{\widehat{\Lambda}(x_0) | \tilde{Y}\} \right] + E \left[\text{var}\{\widehat{\Lambda}(x_0) | \tilde{Y}\} \right].$$

Following arguments similar to the Appendix of Carroll et al. (1999), it can be shown that

$$E \left[\text{var}\{\widehat{\Lambda}(x_0) | \tilde{Y}\} \right] = O\{(nhB)^{-1}\}. \quad (\text{A6})$$

As described in Remark 2, since we can make B as large as possible, we will ignore terms of order $O\{(nhB)^{-1}\}$, in which case

$$\text{var}\{\widehat{\Lambda}(x_0)\} \simeq \text{var} \left[E\{\widehat{\Lambda}(x_0) | \tilde{Y}\} \right]. \quad (\text{A7})$$

Now write

$$\begin{aligned}
\widehat{\Lambda}(x_0) &= D_1 + D_2, \\
D_1 &= d_0 \{mn v(x_0, 0)\}^{-1} \sum_{j=1}^m \sum_{i=1}^n K_h\{W_i^{(j)}(0) - x_0\} Q[Y_{i,j}, W_i^{(j)}(0), g\{W_i^{(j)}(0), 0\}], \\
D_2 &= \sum_{k=1}^K d_k \{Bmn v(x_0, \zeta_k)\}^{-1} \\
&\times \sum_{b=1}^B \sum_{j=1}^m \sum_{i=1}^n K_h\{W_{b,i}^{(j)}(\zeta_k) - x_0\} Q[Y_{i,j}, W_{b,i}^{(j)}(\zeta_k), g\{W_{b,i}^{(j)}(\zeta_k), \zeta_k\}].
\end{aligned} \tag{A8}$$

Note that D_1 does not depend on b since, when $\zeta = 0$, no additional error is added on.

It is easy to check that

$$\text{var}(D_1) = \frac{\gamma d_0^2 R(x_0)}{nmh} + o\{(nh)^{-1}\}.$$

To deal with (A8), as in Remark 1, we need to consider two different situations: $m > 3$ and $m = 3$.

When $m > 3$, $W_{b,i}(\zeta)$ is a continuous random variable. Using a similar argument to that in Carroll et al. (1999), when $\zeta > 0$, we have

$$\text{var} \left(E\{nv(x_0, \zeta)\}^{-1} \sum_{i=1}^n E \left[K_h\{W_{b,i}(\zeta) - x_0\} Q[Y_i, W_{b,i}(\zeta), g\{W_{b,i}(\zeta), \zeta\}] \tilde{Y} \right] \right) = O(n^{-1}).$$

Thus $\text{var}\{E(D_2|\tilde{Y})\} = O(n^{-1})$ and $\text{cov}\{E(D_1|\tilde{Y}), E(D_2|\tilde{Y})\} = o\{(nh)^{-1}\}$. This combined with (A7) completes the proof of (14) when $m > 3$.

In the case of $m = 3$, $c_{b,i,k}$ takes on two values only: $c_{b,i,k} = \pm 2^{-1/2}$. Therefore, given $\tilde{Y}_i$, $W_{b,i}(\zeta)$ is a discrete random variable which takes on two values $W_{b,i}(\zeta) = \tilde{W}_i \pm \zeta^{1/2}(\tilde{W}_{i,1} - \tilde{W}_{i,2})/2$, each with probability 1/2; see Remark 1. Let

$$W_i(s, \zeta) = \overline{W}_i + s\zeta^{1/2}(\tilde{W}_{i,1} - \tilde{W}_{i,2})/2, \quad s = \pm 1.$$

Then, again using (A7) and ignoring higher-order terms, we obtain

$$\begin{aligned}
&E\{\widehat{\Lambda}(x_0) | \tilde{Y}\} \\
&= \frac{1}{2} \sum_{k=0}^K d_k n^{-1} \sum_{i=1}^n \sum_{s=\pm 1} \{v(x_0, \zeta_k)\}^{-1} K_h\{W_i(s, \zeta_k) - x_0\} Q[Y_i, W_i(s, \zeta_k), g\{W_i(s, \zeta_k), \zeta_k\}] \\
&= (2n)^{-1} \sum_{i=1}^n \sum_{s=\pm 1} \sum_{k=0}^K \{d_k / v(x_0, \zeta_k)\} K_h\{W_i(s, \zeta_k) - x_0\} Q[Y_i, W_i(s, \zeta_k), g\{W_i(s, \zeta_k), \zeta_k\}].
\end{aligned}$$

The terms inside the summation over i are independent with mean zero. This means that $\text{var} E\{\widehat{\Lambda}(x_0) | \tilde{Y}\} = E[E\{\widehat{\Lambda}(x_0) | \tilde{Y}\}]^2$. Detailed calculations show that

$$\begin{aligned} \text{var} \left[E \left\{ \widehat{\Lambda}(x_0) | \tilde{Y} \right\} \right] &= \frac{\gamma f_W(x_0, 0) \beta(x_0, 0) d_0^2}{v^2(x_0, 0) nh} \\ &+ \frac{1}{2} \sum_{k=1}^K \frac{\gamma f_W(x_0, \zeta_k) \beta(x_0, \zeta_k) d_0^2}{v^2(x_0, \zeta_k) nh} + o\{h^4 + (nh)^{-1}\}. \end{aligned} \quad (\text{A9})$$

Once again using (A7), we complete the proof of (15).

Fig. 1.
 Plots of (a) sample variances vs true X_i , (b) sample variances vs sample means $\bar{Y}_{i..}$, (c) true variance function, solid line, and $\hat{g}_N$, dashed line.

Fig. 2.
Plots of the limiting functions of (a) the naive estimator and (b) the direct SIMEX estimator for $m = 3, 9, 15$ and the true function.

Fig. 3.
Tumour study. Plots of subsets based on (a) five and (b) ten tumours.

Fig. 4. Simulation study. (a) Histogram of χ and its kernel density estimate, (b) sample means plotted against sample variances together with the estimate $\hat{g}$ of the variance function, (c) Q-Q plot of residuals.

Table 1
Simulation study. Squared biases, variances and mean squared errors.

	<i>n</i> = 1000			<i>n</i> = 2000			<i>n</i> = 3000		
	<i>m</i> = 3	<i>m</i> = 6	<i>m</i> = 9	<i>m</i> = 3	<i>m</i> = 6	<i>m</i> = 9	<i>m</i> = 3	<i>m</i> = 6	<i>m</i> = 9
	Squared bias								
Naïve	0.4928	0.4511	0.1883	0.1455	0.0914	0.0623	0.0808	0.0510	0.0409
DSIMEX	0.2962	0.2106	0.1312	0.0981	0.0695	0.0487	0.0543	0.0370	0.0326
PSIMEX	0.2077	0.1502	0.1426	0.0913	0.0692	0.0680	0.0619	0.0540	0.0565
Variance									
Naïve	0.4604	0.4314	0.1765	0.1209	0.0770	0.0523	0.0562	0.0364	0.0309
DSIMEX	0.2832	0.1979	0.1229	0.0913	0.0623	0.0429	0.0497	0.0314	0.0262
PSIMEX	0.2007	0.1266	0.1085	0.0806	0.0403	0.0291	0.0508	0.0249	0.0182
Mean squared error									
Naïve	0.9533	0.8825	0.3648	0.2664	0.1684	0.1146	0.1371	0.0874	0.0718
DSIMEX	0.5794	0.4085	0.2541	0.1895	0.1317	0.0916	0.1040	0.0684	0.0587
PSIMEX	0.4084	0.2769	0.2511	0.1719	0.1095	0.0971	0.1127	0.0789	0.0747

DSIMEX, direct SIMEX approach; PSIMEX, permutation SIMEX approach.