The Annals of Statistics
2010, Vol. 38, No. 5, 2723–2750
DOI: 10.1214/10-AOS802
© Institute of Mathematical Statistics, 2010

NONPARAMETRIC ESTIMATION OF GENEWISE VARIANCE FOR MICROARRAY DATA[1]

By Jianqing Fan, Yang Feng and Yue S. Niu

Princeton University, Columbia University and University of Arizona

Estimation of genewise variance arises from two important applications in microarray data analysis: selecting significantly differentially expressed genes and validation tests for normalization of microarray data. We approach the problem by introducing a two-way nonparametric model, which is an extension of the famous Neyman–Scott model and is applicable beyond microarray data. The problem itself poses interesting challenges because the number of nuisance parameters is proportional to the sample size and it is not obvious how the variance function can be estimated when measurements are correlated. In such a high-dimensional nonparametric problem, we proposed two novel nonparametric estimators for genewise variance function and semiparametric estimators for measurement correlation, via solving a system of nonlinear equations. Their asymptotic normality is established. The finite sample property is demonstrated by simulation studies. The estimators also improve the power of the tests for detecting statistically differentially expressed genes. The methodology is illustrated by the data from microarray quality control (MAQC) project.

1. Introduction. Microarray experiments are one of widely used technologies nowadays, allowing scientists to monitor thousands of gene expressions simultaneously. One of the important scientific endeavors of microarray data analysis is to detect statistically differentially expressed genes for downstream analysis [Cui, Hwang and Qiu (2005), Fan et al. (2004), Fan and Ren (2006), Storey and Tibshirani (2003), Tusher, Tibshirani and Chu (2001)]. Standard t-test and F-test are frequently employed. However, due to the cost of the experiment, it is common to see a large number of genes with

Received September 2009; revised January 2010.

[1] Supported by NSF Grant DMS-07-14554 and NIH Grant R01-GM072611.

AMS 2000 subject classifications. Primary 62G05; secondary 62P10.

Key words and phrases. Genewise variance estimation, gene selection, local linear regression, nonparametric model, correlation correction, validation test.

a small number of replications. Even in customized arrays where only several hundreds of genes expressions are measured, the number of replications is usually limited. As a result, we are facing a high-dimensional statistical problem with a large number of parameters and a small sample size.

Genewise variance estimation arises at the heart of microarray data analysis. To select differentially expressed genes among thousands of genes, the t-test is frequently employed with a stringent control of type I errors. The degree of freedom is usually small due to limited replications. The power of the test can be significantly improved if the genewise variance can be estimated accurately. In such a case, the t-test becomes basically a z-test. A simple genewise variance estimator is the sample variance of replicated data, which is not reliable due to a relatively small number of replicated genes. They have direct impact on the sensitivity and specificity of t-test [Cui, Hwang and Qiu (2005)]. Therefore, novel methods for estimating the genewise variances are needed for improving the power of the standard t-test.

Another important application of genewise variance estimation arises from testing whether systematic biases have been properly removed after applying some normalization method, or selecting the most appropriate normalization technique for a given array. Fan and Niu (2007) developed such validation tests (see Section 4), which require the estimation of genewise variance. The methods of variance estimation, like pooled variance estimator, and REML estimator [Smyth, Michaud and Scott (2005)], are not accurate enough due to the small number of replications.

Due to the importance of genewise variance in microarray data analysis, conscientious efforts have been made to accurately estimate it. Various methods have been proposed under different models and assumptions. It has been widely observed that genewise variance is to a great extent related to the intensity level. Kamb and Ramaswami (2001) proposed a crude regression estimation of variance from microarray control data. Tong and Wang (2007) discussed a family of shrinkage estimators to improve the accuracy.

Let R_{gi} and G_{gi}, respectively, be the intensities of red (Cy3) and green (Cy5) channels for the ith replication of the gth gene on a two-color microarray data. The log-ratios and log-intensities are computed, respectively, as

$$Y_{gi} = \log_2(G_{gi}/R_{gi}) \quad \text{and} \quad X_{gi} = \tfrac{1}{2}\log_2(G_{gi}R_{gi}),$$
$$i = 1, \ldots, I, g = 1, \ldots, N,$$

where I is the number of replications for each gene and N is the number of genes with replications. For the purpose of estimating genewise variance, we assume that there is no systematic biases or the systematic biases have been removed by a certain normalization method. This assumption is always made for selecting significantly differentially expressed genes or validation test under the null hypothesis. Thus, we have

$$Y_{gi} = \alpha_g + \sigma_{gi}\epsilon_{gi}$$

with α_g denoting the log-ratio of gene expressions in the treatment and control samples. Here, $(\epsilon_{g1}, \ldots, \epsilon_{gI})^T$ follows a multivariate normal distribution with $\epsilon_{gi} \sim N(0,1)$ and $\operatorname{Corr}(\epsilon_{gi}, \epsilon_{gj}) = \rho$ when $i \neq j$. It is also assumed that observations from different genes are independent. Such a model was used in Smyth, Michaud and Scott (2005).

In the papers by Wang, Ma and Carroll (2009) and Carroll and Wang (2008), nonparametric measurement-error models have been introduced to aggregate the information of estimating the genewise variance:

$$\begin{aligned} &Y_{gi} = \alpha_g + \sigma(\alpha_g)\varepsilon_{gi}, \\ &\operatorname{corr}(\varepsilon_{gi}, \varepsilon_{gi'}) = 0, \qquad g = 1, \ldots, N, i = 1, \ldots, I. \end{aligned} \tag{1}$$

The model is intended for the analysis of the Affymetrix array (one-color array) data in which α_g represents the expected intensity level, and Y_{gi} is the ith replicate of observed expression level of gene g. When it is applied to the two-color microarray data as in our setting, in which α_g is the relative expression profiles between the treatment and control, several drawbacks emerge: (a) the model is difficult to interpret as the genewide variance is a function of the log-ratio of expression profiles; (b) errors-in-variable methods have a very slow rate of convergence for the nonparametric problem and the observed intensity information X_{gi} is not used; (c) they are usually hard to be implemented robustly and depend sensitively on the distribution of $\sigma(\alpha_g)\varepsilon_{gi}$ and the i.i.d. assumption on the noise; (d) in many microarray applications, $\alpha_g = 0$ for most g and hence $\sigma(\alpha_g)$ are the same for most genes, which is unrealistic. Therefore, our model (2) below is complementary to that of Wang, Ma and Carroll (2009) and Carroll and Wang (2008), with focus on the applications to two-color microarray data.

To overcome these drawbacks in the applications to microarray data and to utilize the observed intensity information, we assume that $\sigma_{gi} = \sigma(X_{gi})$ for a smooth function $\sigma(\cdot)$. This leads to the following two-way nonparametric model:

$$Y_{gi} = \alpha_g + \sigma(X_{gi})\epsilon_{gi}, \qquad g = 1, \ldots, N, i = 1, \ldots, I, \tag{2}$$

for estimating genewise variance. This model is clearly an extension of the Neyman–Scott problem [Neyman and Scott (1948)], in which the genewise variance is a constant. The Neyman–Scott problem has many applications in astronomy. Note that the number of nuisance parameters $\{\alpha_g\}$ is proportional to the sample size. This imposes an important challenge to the nonparametric problem. It is not even clear whether the function $\sigma(\cdot)$ can be consistently estimated.

To estimate the genewise variance in their microarray data analysis, Fan et al. (2004) assumed a model similar to (2). But in the absence of other available techniques, they had to impose that the treatment effect $\{\alpha_g\}$ is also a smooth function of the intensity level so that they can apply nonparametric methods to estimate genewise variance [Ruppert et al. (1997)]. However, this assumption is not valid in most microarray applications, and the estimator of genewise variance incurs big biases unless $\{\alpha_g\}$ is sparse, a situation that Fan et al. (2004) hoped. Fan and Niu (2007) approached this problem in another simple way. When the noise in the replications is small, that is, $X_{gi} \approx \bar{X}_g$, where $\bar{X}_g$ is the sample mean for the gth gene. Therefore, they simply smoothed the pair $\{(\bar{X}_g, \bar{r}_g)\}$, where $\bar{r}_g = \sum_{i=1}^{I}(Y_{gi} - \bar{Y}_g)^2/(I-1)$. This also leads to a biased estimator, which is denoted as $\hat{\xi}^2(x)$. One asks naturally whether the function $\sigma(\cdot)$ is estimable and how it can be estimated in the general two-way nonparametric model.

We propose a novel nonparametric approach to estimate the genewise variance. We first study a benchmark case when there is no correlation between replications, that is, $\rho = 0$. This corresponds to the case with independent replications across arrays [Fan, Peng and Huang (2005), Huang, Wang and Zhang (2005)]. It is also applicable to those dealt by the Neyman–Scott problem. By noticing $\mathrm{E}\{(Y_{gi} - \bar{Y}_g)^2 | X_{gi}\}$ is a linear combination of $\sigma^2(X_{gi})$, we obtain a system of linear equations. Hence, $\sigma^2(\cdot)$ can be estimated via nonparametric regression of a proper linear combination of $\{(Y_{gi} - \bar{Y}_g)^2, i = 1, \ldots, I\}$ on $\{X_{gi}\}$. The asymptotic normality of the estimator is established. In the case that the replication correlation does not vanish, the system of equations becomes nonlinear and cannot be analytically solved. However, we are able to derive the correlation corrected estimator, based on the estimator without genewise correlation. The genewise variance function and the correlation coefficient of repeated measurements are simultaneously estimated by iteratively solving a nonlinear equation. The asymptotic normality of such estimators is established.

Model (2) can be applied to the microarrays in which within-array replications are not available. In that case, we can aggregate all the microarrays together and view them as a super array with replications [Fan, Peng and Huang (2005), Huang, Wang and Zhang (2005)]. In other words, i in (2) indexes arrays and ρ can be taken as 0, namely (2) is the across-array replication with $\rho = 0$.

The structure of this paper is as follows. In Section 2, we discuss the estimation schemes of the genewise variance and establish the asymptotic properties of the estimators. Simulation studies are given in Section 3 to verify the performance of our methods in the finite sample. Applications to the data from Microarray Quality Control (MAQC) project are showed in Section 4 to illustrate the proposed methodology. In Section 5, we give a short summary. Technical proofs are relegated to the Appendix.

2. Nonparametric estimators of genewise variance.

2.1. *Estimation without correlation.*

We first consider the specific case where there is no correlation among the replications $Y_{g1}, \dots, Y_{gI}$ of the same gene g under model (2). This is usually applicable to the across-array replication and stimulates our procedure for the more general case with the replication correlation. In the former case, we have

$$\mathrm{E}[(Y_{gi} - \bar{Y}_g)^2 | \mathbf{X}] = (I-1)^2 \sigma^2(X_{gi})/I^2 + \sum_{j \neq i} \sigma^2(X_{gj})/I^2, \qquad i = 1, \dots, I.$$

We will discuss in Section 2.2.4 the case that $I = 2$. For $I > 2$, we have I different equations with I unknowns $\sigma^2(X_{g1})$, $\sigma^2(X_{g2}), \dots, \sigma^2(X_{gI})$ for a given gene g. Solving these I equations, we can express the unknowns in terms of $\{\mathrm{E}[(Y_{gi} - \bar{Y}_g)^2 | \mathbf{X}]\}_{i=1}^{I}$, estimable quantities. Let

$$\mathbf{r}_g = ((Y_{g1} - \bar{Y}_g)^2, \dots, (Y_{gI} - \bar{Y}_g)^2)^T \quad \text{and} \quad \boldsymbol{\sigma}_g^2 = (\sigma^2(X_{g1}), \dots, \sigma^2(X_{gI}))^T.$$

Then, it can easily be shown that $\boldsymbol{\sigma}_g^2 = \mathbf{B}\mathrm{E}[\mathbf{r}_g | \mathbf{X}]$, where $\mathbf{B}$ is the coefficient matrix:

$$\mathbf{B} = ((I^2 - I)\mathbf{I} - \mathbf{E})/(I-1)(I-2)$$

with $\mathbf{I}$ being the $I \times I$ identity matrix and $\mathbf{E}$ the $I \times I$ matrix with all elements 1. Define

$$\mathbf{Z}_g = (Z_{g1}, \dots, Z_{gI})^T \overset{\triangle}{=} \mathbf{B}\mathbf{r}_{\mathbf{g}}.$$

Then we have

$$\sigma^2(X_{gi}) = \mathrm{E}[Z_{gi} | \mathbf{X}]. \tag{3}$$

Note that the left-hand side of (3) depends only on X_{gi}, not other variables. By the the double expectation formula, it follows that the variance function $\sigma^2(\cdot)$ can be expressed as the univariate regression

$$\sigma^2(x) = \mathrm{E}[Z_{gi} | X_{gi} = x], \qquad i = 1, \dots, I. \tag{4}$$

Using the synthetic data $\{(X_{gi}, Z_{gi}), g = 1, \dots, N\}$ for each given i, we can apply the local linear regression technique [Fan and Gijbels (1996)] to obtain a nonparametric estimator $\hat{\eta}_i^2(x)$ of $\sigma^2(\cdot)$. Explicitly, for a given kernel K and bandwidth h,

$$\hat{\eta}_i^2(x) = \sum_{g=1}^{N} W_{N,i}\left(\frac{X_{gi} - x}{h}\right) Z_{gi}, \qquad i = 1, \dots, I, \tag{5}$$

with

$$W_{N,i}(u) = h^{-1}K(u)\frac{S_{N,2} - uS_{N,1}}{S_{N,2}S_{N,0} - S_{N,1}^2},$$

where $K_h(u) = h^{-1}K(u/h)$ and $S_{N,l} = \sum_{g=1}^{N} K_h(X_{gi} - x)[(X_{gi} - x)/h]^l$, whose dependence on i is suppressed. Thus, we have I estimators $\hat\eta_1^2(x), \ldots, \hat\eta_I^2(x)$ for the same genewise variance function $\sigma^2(\cdot)$. Each of these I estimators $\hat\eta_i^2(x)$ is a consistent estimator of $\sigma^2(x)$. To optimally aggregate those I estimators, we need the asymptotic properties of $\boldsymbol{\eta}(x) = (\hat\eta_1^2(x), \ldots, \hat\eta_I^2(x))^T$.

Denote

$$c_K = \int_{-\infty}^{\infty} u^2 K(u)\, du, \qquad d_K = \int_{-\infty}^{\infty} K^2(u)\, du,$$

$$\sigma_1 = \mathrm{E}[\sigma(X_{gi})] \quad \text{and} \quad \sigma_2 = \mathrm{E}[\sigma^2(X_{gi})].$$

Assume that X_{gi} are i.i.d. with marginal density $f_X(\cdot)$ and ε_{gi} are i.i.d. random variables from the standard normal distribution. In the following result, we assume that I is fixed, but N diverges.

THEOREM 1. *Under the regularity conditions in the Appendix, for a fixed point x, we have*

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\eta} - (\sigma^2(x) + b(x) + o_P(h^2))\mathbf{e}) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}),$$

provided that $h \to 0$ and $Nh \to \infty$, where $\mathbf{e} = (1, 1, \ldots, 1)^T$ and

$$\boldsymbol{\Sigma} = V_1\mathbf{I} + V_2(\mathbf{E} - \mathbf{I})$$

with $b(x) = \frac{h^2}{2}c_K(\sigma^2(x))''$,

$$V_1 = \frac{d_K}{Nhf_X(x)}\left\{2\sigma^4(x) + \frac{4 + 4(I-1)(I-3)}{(I-1)(I-2)^2}\sigma_2\sigma^2(x) + \frac{2}{(I-1)(I-2)}\sigma_2^2\right\},$$

$$V_2 = \frac{1}{N}\left\{\frac{4}{(I-1)^2}\sigma^4(x) - \frac{8}{(I-1)^2}\sigma_2\sigma^2(x) + \frac{2(I-3)}{(I-1)^2(I-2)}\sigma_2^2\right\}.$$

Note that V_2 is one order of magnitude smaller than V_1. Hence, the estimators $\hat\eta_1^2(x), \ldots, \hat\eta_I^2(x)$ are asymptotically independently distributed as $N(\sigma^2(x) + b(x), V_1)$. Their dependence is only in the second order. The best linear combination of I estimators is

$$\hat\eta^2(x) = [\hat\eta_1^2(x) + \hat\eta_2^2(x) + \cdots + \hat\eta_I^2(x)]/I \tag{6}$$

with the asymptotic distribution

$$N(\sigma^2(x) + b(x), V_1/I + (1 - 1/I)V_2). \tag{7}$$

See also the aggregated estimator (16) with $\rho = 0$, which has the same asymptotic property as the estimator (8). See Remark 1 below for additional discussion.

Theorem 1 gives the asymptotic normality of the proposed nonparametric estimators under the presence of a large number of nuisance parameters $\{\alpha_g\}_{g=1}^N$. With the newly proposed technique, we do not have to impose any assumptions on α_g such as sparsity or smoothness. This kind of local linear estimator can be applied to most two-color microarray data, for instance, customized arrays and Agilent arrays.

2.2. *Variance estimation with correlated replications.*

2.2.1. *Aggregated estimator.* We now consider the case with correlated with-array replications. There is a lot of evidence that correlation among within-array replicated genes exists [Smyth, Michaud and Scott (2005), Fan and Niu (2007)]. Suppose that within-array replications have a common correlation $\operatorname{corr}(Y_{gi}, Y_{gj}|\mathbf{X}) = \rho$ when $i \neq j$. Observations across different genes or arrays are independent. Then the conditional variance of $(Y_{gi} - \bar{Y}_g)$ can be expressed as

$$\begin{aligned} &\operatorname{var}[(Y_{gi} - \bar{Y}_g)|\mathbf{X}] \\ &\quad = (I-1)^2\sigma^2(X_{gi})/I^2 + 2\rho \sum_{\substack{1 \le j < k \le I, \\ j \neq i, k \neq i}} \sigma(X_{gj})\sigma(X_{gk})/I^2 \\ &\qquad + 2(I-1)\rho \sum_{j \neq i} \sigma^2(X_{gj})/I^2 - \sum_{j \neq i} \sigma(X_{gi})\sigma(X_{gj})/I^2. \end{aligned} \tag{8}$$

This is a complex system of nonlinear equations and the analytic form cannot be found. Innovative ideas are needed.

Using the same notation as that in the previous section, it can be calculated that

$$\begin{aligned} \mathrm{E}[Z_{gi}|\mathbf{X}] = {}& \sigma^2(X_{gi}) - \frac{2}{I-1} \sum_{j \neq i} \rho\sigma(X_{gi})\sigma(X_{gj}) \\ & + \frac{2}{(I-1)(I-2)} \sum_{\substack{1 \le j < k \le I, \\ j \neq i, k \neq i}} \rho\sigma(X_{gj})\sigma(X_{gk}). \end{aligned}$$

Taking the expectation with respect to X_{gj} for all $j \neq i$, we obtain

$$\mathrm{E}[Z_{gi}|X_{gi} = x] = \sigma^2(x) - 2\rho\sigma_1\sigma(x) + \rho\sigma_1^2 \overset{\triangle}{=} \eta^2(x), \tag{9}$$

where $\sigma_1 = \mathrm{E}[\sigma(X)]$.

Here, we can directly apply the local linear approach to all aggregated data $\{(X_{gi}, Z_{gi})\}_{i,g=1}^{I,N}$, due to the same regression function (9). Let $\hat{\eta}_A^2(\cdot)$ be the local linear estimator of $\eta^2(\cdot)$, based on the aggregated data. Then

$$\hat{\eta}_A^2(x) = \sum_{g=1}^{N}\sum_{i=1}^{I} W_N\left(\frac{X_{gi}-x}{h}\right) Z_{gi} \tag{10}$$

with

$$W_N(u) = h^{-1}K(u)\frac{S_{NI,2} - uS_{NI,1}}{S_{NI,0}S_{NI,2} - S_{NI,1}^2},$$

where $S_{NI,l} = \sum_{g=1}^{N}\sum_{i=1}^{I} K_h(X_{gi}-x)[(X_{gi}-x)/h]^l$. There are two solutions to (9):

$$\hat{\sigma}_A(x,\rho)^{(1),(2)} = \hat{\rho}\hat{\sigma}_1 \pm \sqrt{\hat{\rho}^2\hat{\sigma}_1^2 - \hat{\rho}\hat{\sigma}_1^2 + \hat{\eta}_A^2(x)}, \tag{11}$$

Notice that given the sample $\mathbf{X}$ and $\mathbf{Y}$, $\hat{\sigma}_A(x,\rho)^{(1),(2)}$ are continuous in both x and ρ. For $\rho < 0$, $\hat{\sigma}_A(x,\rho)^{(1)}$ should be used since the standard deviation should be nonnegative. Since $\hat{\sigma}_A(x,\rho)^{(1)} > \hat{\sigma}_A(x,\rho)^{(2)}$ for every x and ρ, by the continuity of the solution in ρ, we can only use the same solution when ρ changes continuously. Then $\hat{\sigma}_A(x,\rho)^{(1)}$ should always be used regardless of ρ. From now on, we drop the superscript and denote

$$\hat{\sigma}_A(x) = \rho\sigma_1 + \sqrt{\rho^2\sigma_1^2 - \rho\sigma_1^2 + \hat{\eta}_A^2(x)}. \tag{12}$$

This is called the aggregated estimator. Note that in (12), ρ, σ_1 and $\sigma(\cdot)$ are all unknown.

2.2.2. *Estimation of correlation.* To estimate ρ, we assume that there are J independent arrays ($J \geq 2$). In other words, we observed data from (2) independently J times. In this case, the residual maximum likelihood (REML) estimator introduced by Smyth, Michaud and Scott (2005) is as follows:

$$\hat{\rho}_0 = \frac{\sum_{g=1}^{N} s_{B,g}^2 - \sum_{g=1}^{N} s_{W,g}^2}{\sum_{g=1}^{N} s_{B,g}^2 + (I-1)\sum_{g=1}^{N} s_{W,g}^2}, \tag{13}$$

where $s_{B,g}^2 = I(J-1)^{-1}\sum_{j=1}^{J}(\bar{Y}_{gj} - \bar{Y}_g)^2$ with $\bar{Y}_{gj} = I^{-1}\sum_{i=1}^{I} Y_{gij}$ and $\bar{Y}_g = J^{-1}\sum_{j=1}^{J}\bar{Y}_{gj}$ is the between-arrays variance and $s_{W,g}^2$ is the within-array variance:

$$s_{W,g}^2 = \frac{1}{J(I-1)}\sum_{j=1}^{J}\sum_{i=1}^{I}(Y_{gij} - \bar{Y}_{gj})^2.$$

As discussed in Smyth, Michaud and Scott (2005), the estimator $\hat{\rho}_0$ of ρ is consistent when $\operatorname{var}(Y_{gij}|\mathbf{X}) = \sigma_g$ is the same for all $i = 1, \ldots, I$ and $j = 1, \ldots, J$. However, this assumption is not valid under the model (2) and a correction is needed. We propose the following estimator:

$$\hat{\rho} = \frac{\sigma_2}{\sigma_1^2} \cdot \frac{\sum_{g=1}^N s_{B,g}^2 - \sum_{g=1}^N s_{W,g}^2}{\sum_{g=1}^N s_{B,g}^2 + (I-1)\sum_{g=1}^N s_{W,g}^2}. \tag{14}$$

The consistency of $\hat{\rho}$ is given by the following theorem.

THEOREM 2. *Under the regularity condition in the Appendix, the estimator $\hat{\rho}$ of ρ is $\sqrt{N}$-consistent:*

$$\hat{\rho} - \rho = O_P(N^{-1/2}).$$

With a consistent estimator of ρ, σ_1, σ_2 and $\sigma_A(\cdot)$ can be solved by the following iterative algorithm:

Step 1. Set $\hat{\eta}_A^2(\cdot)$ as an initial estimate of $\sigma_A^2(\cdot)$.
Step 2. With $\hat{\sigma}_A(\cdot)$, compute

$$\hat{\sigma}_1 = N^{-1}\sum_{g=1}^N \hat{\sigma}_A(X_{gi}), \qquad \hat{\sigma}_2 = N^{-1}\sum_{g=1}^N \hat{\sigma}_A^2(X_{gi}), \qquad \hat{\rho} = \hat{\rho}_0\hat{\sigma}_2/\hat{\sigma}_1^2. \tag{15}$$

Step 3. With $\hat{\sigma}_1$, $\hat{\sigma}_2$ and $\hat{\rho}$, compute $\hat{\sigma}_A(\cdot)$ using (12).
Step 4. Repeat steps 2 and 3 until convergence.

This provides simultaneously the estimators $\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\rho}$ and $\hat{\sigma}_A(\cdot)$. From our numerical experience, this algorithm converges quickly after a few iterations. When the algorithm converges, the estimator $\sigma_A^2(x)$ is given by

$$\hat{\sigma}_A(x) = \hat{\rho}\hat{\sigma}_1 + \sqrt{\hat{\rho}^2\hat{\sigma}_1^2 - \hat{\rho}\hat{\sigma}_1^2 + \hat{\eta}_A^2(x)}. \tag{16}$$

Note that the presence of multiple arrays is only used to estimate the correlation ρ for the replications. It is not needed for estimating the genewise variance function. In the case of the presence of J arrays, we can take the average of the J estimates from each array.

2.2.3. *Asymptotic properties.* Following a similar idea as the case without correlation, we can derive the asymptotic property of $\hat{\eta}_A^2(x)$.

THEOREM 3. *Under the regularity conditions in the Appendix, for a fixed point x, we have*

$$\{V^*\}^{-1/2}\{\hat{\eta}_A^2(x) - [\eta^2(x) + \beta(x)] + o_P(h^2)\} \xrightarrow{D} N(0,1),$$

provided that $h \to 0$ *and* $Nh \to \infty$, *with* $\beta(x) = \frac{h^2}{2} c_K (\eta^2(x))''$ *and*

$$V^* = \frac{1}{I} V_1' + \frac{I-1}{I} V_2',$$

where

$$V_1' = \frac{d_K}{Nhf_X(x)} \{2\sigma^4(x) - 8\rho\sigma_1\sigma^3(x) + C_2\sigma^2(x) + C_3\sigma(x) + C_4\},$$

$$V_2' = \frac{1}{N} \{D_0\sigma^4(x) + D_1\sigma^3(x) + D_2\sigma^2(x) + D_3\sigma(x) + D_4\}$$

with coefficients $C_2, \ldots, C_4, D_0, \ldots, D_4$ *defined in the Appendix.*

The asymptotic normality of $\hat{\sigma}_A^2(x)$ can be derived from that of $\hat{\eta}_A^2(x)$. More specifically, $\hat{\sigma}_A^2(x) = \varphi(\eta_A^2(x))$ with $\varphi(z) = (\rho\sigma_1 + \sqrt{\rho^2\sigma_1^2 - \rho\sigma_1^2 + z})^2$. The derivative of $\varphi(\cdot)$ with respect to z is $\psi(z) = \rho\sigma_1 / \sqrt{\rho^2\sigma_1^2 - \rho\sigma_1^2 + z} + 1$. Then, by the delta method, we have

$$\{V^*\}^{-1/2}(\hat{\sigma}_A^2(x) - \varphi(\eta^2(x) + \beta(x) + o_P(h^2))) \xrightarrow{D} N(0, \psi^2(\eta^2(x))).$$

REMARK 1. An alternative approach when correlation exists is to apply the same correlation correction idea to $\{X_{gi}, Z_{gi}\}_{g=1}^N$ for every replication i, resulting in the estimator $\hat{\sigma}_i^2(x)$. In this case, it can be proved that the best linear combination of the estimator is

$$\hat{\sigma}^2(x) = [\hat{\sigma}_1^2(x) + \hat{\sigma}_2^2(x) + \cdots + \hat{\sigma}_I^2(x)]/I. \tag{17}$$

This estimator has the same asymptotic performance as the aggregated estimator. However, we prefer the aggregated estimator due to the following reasons: the equation (16) only needs to be solved once by using the algorithm in Section 2.2.2, all data are treated symmetrically, and $\hat{\eta}_A^2(\cdot)$ can be estimated more stably.

2.2.4. *Two replications.* The aforementioned methods apply to the case when there are more than two replications. For the case $I = 2$, the equations for $\operatorname{var}[(Y_{gi} - \bar{Y}_g)|\mathbf{X}]$ collapse into one. In this case, it can be shown using the same arguments before that

$$\operatorname{var}[(Y_{gi} - \bar{Y}_g)|X_{gi} = x] = \tfrac{1}{4}\sigma^2(x) + \tfrac{1}{4}\sigma_2 - \tfrac{1}{2}\rho\sigma_1\sigma(x), \qquad i = 1, 2, \tag{18}$$

where $\sigma_2 = \mathrm{E}[\sigma^2(X_{gi})]$. In this case, the left-hand side is always equal to $\operatorname{var}[(Y_{g1} - Y_{g2})/2|X_{gi} = x]$.

Let $\hat{\eta}^2(x)$ be the local linear estimator of the function on the right-hand side by smoothing $\{(Y_{g1} - Y_{g2})^2/4\}_{g=1}^N$ on $\{X_{g1}\}_{g=1}^N$ and $\{X_{g2}\}_{g=1}^N$. Then the genewise variance is a solution to the following equation:

$$\hat{\sigma}(x) = \hat{\rho}\hat{\sigma}_1 + \sqrt{\hat{\rho}^2\hat{\sigma}_1^2 - \hat{\sigma}_2 + 4\hat{\eta}^2(x)}. \tag{19}$$

The algorithm in Section 2.2.2 can be applied directly.

3. Simulations and comparisons.

In this section, we conduct simulations to evaluate the finite sample performance of different variance estimators $\hat{\xi}^2(x)$, $\hat{\eta}^2(x)$ and $\hat{\sigma}_A^2(x)$. First, the bias problem of the naive nonparametric variance estimator $\hat{\xi}^2(x)$ is demonstrated. It is shown that this bias issue can be eliminated by our newly proposed methods. Then we consider the estimators $\hat{\eta}^2(x)$ and $\hat{\sigma}_A^2(x)$ under different configurations of the within-array replication correlation.

3.1. *Simulation design.*

In all the simulation examples, we set the number of genes $N = 2000$, each gene having $I = 3$ within-array replications and $J = 4$ independent arrays. For the purpose of investigating the genewise variance estimation, the data are generated from model (2). The details of simulation scheme are summarized as follows:

α_g: The expression levels of the first 250 genes are generated from the standard double exponential distribution. The rest are 0s. These expression levels are the same over 4 arrays in each simulation, but may vary over simulations.

X: The intensity is generated from a mixture distribution: with probability 0.7 from the distribution $0.0004(x-6)^3 I(6 < x < 16)$ and 0.3 from the uniform distribution over $[6, 16]$.

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

$\sigma^2(\cdot)$: The genewise variance function is taken as

$$\sigma^2(x) = 0.15 + 0.015(12 - x)^2 I\{x < 12\}.$$

The parameters are taken from Fan, Peng and Huang (2005). The kernel function is selected as $\frac{70}{81}(1 - |x|^3)^3 I(|x| \le 1)$. In addition, we fix the bandwidth $h = 1$ for all the numerical analysis.

For every setting, we repeat the whole simulation process for T times and evaluate the estimates of $\sigma^2(\cdot)$ over $K = 101$ grid points $\{x_k\}_{k=1}^K$ on the interval $[6, 16]$. For the kth grid point, we define

$$B_k = \bar{\sigma}^2(x_k) - \sigma^2(x_k) \qquad \text{with } \bar{\sigma}^2(x_k) = T^{-1}\sum_{t=1}^T \hat{\sigma}_t^2(x_k),$$

$$S_k = T^{-1}\sum_{t=1}^T [\hat{\sigma}_t^2(x_k) - \bar{\sigma}^2(x_k)]^2,$$

and $\mathrm{MSE}_k = B_k^2 + S_k$. Let $f(\cdot)$ be the density function of intensity X. Let

$$\mathrm{Bias}^2 = \sum_{k=1}^{K} B_k^2 f(x_k) \Big/ \sum_{k=1}^{K} f(x_k), \qquad \mathrm{VAR} = \sum_{k=1}^{K} S_k f(x_k) \Big/ \sum_{k=1}^{K} f(x_k)$$

and

$$\mathrm{MISE} = \sum_{k=1}^{K} \mathrm{MSE}_k f(x_k) \Big/ \sum_{k=1}^{K} f(x_k)$$

be the integrated squared bias (Bias^2), the integrated variance (VAR), and the integrated mean squared error (MISE) of the estimate $\hat{\sigma}^2(\cdot)$, respectively. For the tth simulation experiment, we define

$$\mathrm{ISE}_t = \sum_{k=1}^{K} (\hat{\sigma}_t^2(x_k) - \sigma^2(x_k))^2 f(x_k) \Big/ \sum_{k=1}^{K} f(x_k)$$

be the integrated squared error for the tth simulation.

3.2. *The bias of naive nonparametric estimator.* A naive approach is to regard α_g in (2) as a smooth function of X_{gi}, namely, $\alpha_g = \alpha(X_{gi})$. The function $\alpha(\cdot)$ can be estimated by a local linear regression estimator, resulting in an estimated function $\hat{\alpha}(\cdot)$. The squared residuals $\{r_{gi}^2\}_{g=1}^N$ is then further smoothed on $\{X_{gi}\}_{g=1}^N$ to obtain an estimate $\hat{\xi}^2(x)$ of the variance function $\sigma^2(\cdot)$, where $r_{gi} = \hat{Y}_{gi} - \hat{\alpha}(X_{gi})$ [Ruppert et al. (1997)].

To provide a comprehensive view of the performances of the naive and the new estimators, we first compare the performances of $\hat{\xi}^2(x)$ and $\hat{\eta}^2(x)$ under the smoothness assumption of the gene effect α_g. Data from the naive nonparametric regression model is also generated with

$$\alpha(x) = \exp\left(-\frac{1}{1-(x-13)^2}\right) I\{12 < x < 14\}.$$

This allows us to understand the loss of efficiency when α_g is continuous in X_{gi}. This usually does not occur for microarray data, but can appear in other applications. Note that $\alpha(\cdot)$ is zero in most of the region and thus is reasonably sparse. Here, the number of simulations is taken to be $T = 100$. The data is generated with the assumption that $\rho = 0$, in which case the variance estimators $\hat{\eta}^2(x)$ and $\hat{\sigma}_A^2(x)$ have the same performance (see also Table 2 below). Thus, we only report the performance of $\hat{\eta}^2(x)$.

In Table 1, we report the mean integrated squared bias (Bias^2), the mean integrated variance (VAR), and the mean integrated squared error (MISE) of $\hat{\xi}^2(x)$ and $\hat{\eta}^2(x)$ with and without the smoothness assumption on the gene effect α_g. From the left panel of Table 1, we can see that when the

TABLE 1

*Mean integrated squared bias (*Bias^2*), mean integrated variance (VAR), mean integrated squared error (MISE) over 100 simulations for variance estimators $\hat{\xi}^2(x)$ and $\hat{\eta}^2(x)$. Two different gene effect functions $\alpha(\cdot)$ are implemented. All quantities are multiplied by 1000*

	Smooth gene effect			**Nonsmooth gene effect**		
	Bias^2	**VAR**	**MISE**	**Bias^2**	**VAR**	**MISE**
$\hat{\xi}^2(x)$	0.01	0.14	0.15	16.00	1.47	17.47
$\hat{\eta}^2(x)$	0.57	0.24	0.80	0.00	0.22	0.23

TABLE 2

*Mean integrated squared bias (*Bias^2*), mean integrated variance (VAR), mean integrated squared error (MISE) over 1000 simulations for different variance estimators $\hat{\eta}^2(x)$ and $\hat{\sigma}_O^2(x)$. Seven different correlation schemes are simulated: $\rho = -0.4$, $\rho = -0.2$, $\rho = 0$, $\rho = 0.2$, $\rho = 0.4$, $\rho = 0.6$ and $\rho = 0.8$. All quantities are multiplied by 1000*

		$\boldsymbol{\rho}$						
		−0.4	**−0.2**	**0**	**0.2**	**0.4**	**0.6**	**0.8**
Bias^2	$\hat{\eta}^2(x)$	5.93	1.48	0.00	1.48	5.91	13.31	23.67
	$\hat{\sigma}_A^2(x)$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	$\hat{\sigma}_O^2(x)$	0.00	0.00	0.00	0.00	0.00	0.00	0.01
VAR	$\hat{\eta}^2(x)$	0.44	0.33	0.24	0.16	0.10	0.05	0.02
	$\hat{\sigma}_A^2(x)$	0.27	0.25	0.24	0.22	0.20	0.19	0.20
	$\hat{\sigma}_O^2(x)$	0.27	0.25	0.24	0.22	0.20	0.18	0.23
MISE	$\hat{\eta}^2(x)$	6.37	1.81	0.24	1.64	6.01	13.37	23.69
	$\hat{\sigma}_A^2(x)$	0.27	0.25	0.24	0.22	0.21	0.19	0.20
	$\hat{\sigma}_O^2(x)$	0.27	0.25	0.24	0.22	0.20	0.18	0.24

smoothness assumption is valid, the estimator $\hat{\xi}^2(x)$ outperforms $\hat{\eta}^2(x)$. The reason is that the mean function $\alpha(X_{gi})$ depends on the replication and is not a constant. Therefore, model (2) fails and $\hat{\eta}^2(x)$ is biased. One should compare the results with those on the second row of the right panel where the model is right for $\hat{\eta}^2(x)$. In this case, $\hat{\eta}^2(x)$ performs much better. Its variance is about $3/2$ as large as the variance in the case that mean is generated from a smooth function $\alpha(X_{gi})$. This is expected. In the latter case, to eliminate α_g, the degree of freedom reduces from $I = 3$ to 2, whereas in the former case, $\alpha(X_{gi})$ can be estimated without losing the degree of freedom, namely the number of replications is still 3. The ratio $3/2$ is reflected in Table 1. However, when the smoothness assumption does not hold, there is serious bias in the estimator $\hat{\xi}^2(x)$, even though that α_g is still reasonably sparse. The bias is an order of magnitude larger than those in the other situations.

To see how variance estimators behave, we plot typical estimators $\hat{\xi}^2(x)$ and $\hat{\eta}^2(x)$ with median ISE value among 100 simulations in Figure 1. The solid line is the true variance function while the dotted and dashed lines represent $\hat{\xi}^2(x)$ and $\hat{\eta}^2(x)$, respectively. On the left panel of Figure 1, we can see that estimator $\hat{\xi}^2(x)$ outperforms the estimator $\hat{\eta}^2(x)$ when the smoothness assumption is valid. The region where the biases occur has already been explained above. However, $\hat{\xi}^2(x)$ will generate substantial bias when the nonparametric regression model does not hold, and at the same time, our nonparametric estimator $\hat{\eta}^2(x)$ corrects the bias very well.

3.3. *Performance of new estimators.* In this example, we consider the setting in Section 3.1 that the smoothness assumption of the gene effect α_g is not valid. For comparison purpose only, we add an oracle estimator $\hat{\sigma}^2_O(x)$ in which we assume that σ_1, σ_2 and ρ are all known. We now evaluate the performance of the estimators $\hat{\eta}^2(x)$, $\hat{\sigma}^2_A(x)$ and $\hat{\sigma}^2_O(x)$ when the correlation between within-array replications varies. To be more specific, seven different correlation settings are considered: $\rho = -0.4, -0.2, 0$, 0.2, 0.4, 0.6, 0.8, with $\rho = 0$ representing across-array replications. In this case, we increase the number of simulations to $T = 1000$. Again, we report Bias^2, VAR and MISE of the three estimators for each correlation setting in Table 2. When $\rho = 0$, all the three estimators give the same bias and variance. This is consistent with our theory. We can see clearly from the table that, when $\rho \neq 0$, the estimator $\hat{\sigma}^2_A(x)$ produces much smaller biases than $\hat{\eta}^2(x)$. In fact, when $|\rho|$ as small as 0.2, the bias of $\hat{\eta}^2(x)$ already dominates the variance.

It is worth noticing that the performance of $\hat{\sigma}^2_O(x)$ and $\hat{\sigma}^2_A(x)$ are almost always the same, which indicates that our algorithm for estimating ρ, σ_1

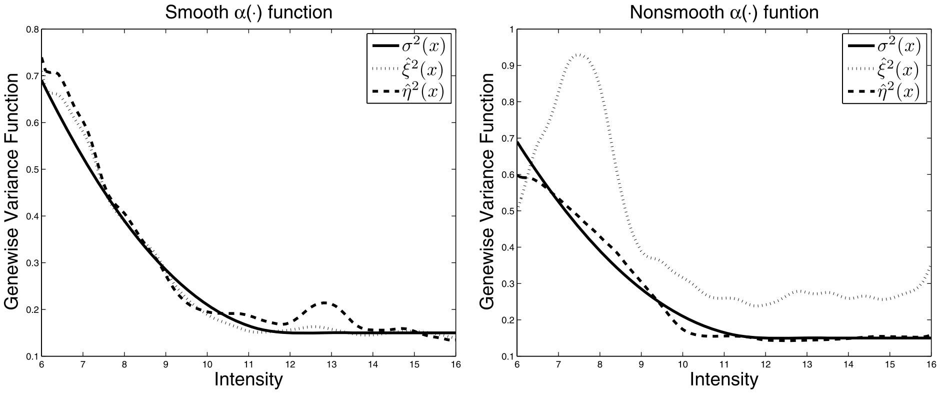

FIG. 1. *Variance estimators $\hat{\xi}^2(x)$ and $\hat{\eta}^2(x)$ with median performance when different gene effect function $\alpha(\cdot)$ are implemented. Left panel: smooth $\alpha(\cdot)$ function. Right panel: nonsmooth $\alpha(\cdot)$ function.*

and σ_2 is very accurate. To see this more clearly, the squared bias, variance and MSE of the estimator ρ, σ_1 and σ_2 in $\hat{\sigma}_A^2(x)$ under the seven correlation settings are reported in Table 3. Here, the true value of σ_1 and σ_2 is 0.4217 and 0.1857. For example, when $\rho = 0.8$, the bias of $\hat{\rho}$ is less than 0.002 for $\hat{\sigma}_A^2(x)$, which is acceptable because the convergence threshold in the algorithm is set to be 0.001.

In Figure 2, we render the estimates $\hat{\eta}^2(x)$ and $\hat{\sigma}_A^2(x)$ with the median ISE under four different correlation settings: $\rho = -0.4$, $\rho = 0$, $\rho = 0.6$ and $\rho = 0.8$. We omit the other correlation schemes since they all have similar performance. The solid lines represent the true variance function. The dotted lines and dashed lines are for $\hat{\eta}^2(x)$ and $\hat{\sigma}_A^2(x)$, respectively. For the case $\rho = 0$, the two estimators are indistinguishable. When $\rho < 0$, $\hat{\eta}^2(x)$ overestimates the genewise variance function, whereas when $\rho > 0$, it underestimates the genewise variance function.

4. Application to human total RNA samples using Agilent arrays.

Our real data example comes from Microarray Quality Control (MAQC) project [Patterson et al. (2006)]. The main purpose of the original paper is on comparison of reproducibility, sensitivity and specificity of microarray measurements across different platforms (i.e., one-color and two-color) and testing sites. The MAQC project use two RNA samples, Stratagene Universal Human Reference total RNA and Ambion Human Brain Reference total RNA. The two RNA samples have been assayed on three kinds of arrays: Agilent, CapitalBio and TeleChem. The data were collected at five sites. Our study focuses only on the Agilent arrays. At each site, 10 two-color Agilent microarrays are assayed with 5 of them dye swapped, totaling 30 microarrays.

Table 3

Squared bias, variance and MSE of $\hat{\rho}$, $\hat{\sigma}_1$ and $\hat{\sigma}_2$ in the estimate $\hat{\sigma}_A^2(x)$. All quantities are multiplied by 10^6

		$\boldsymbol{\rho}$						
$\boldsymbol{\hat{\sigma}_A^2(x)}$		**−0.4**	**−0.2**	**0**	**0.2**	**0.4**	**0.6**	**0.8**
$\hat{\rho}$	Bias^2	0.07	0.04	0.01	0.00	0.00	0.00	3.90
	VAR	7.90	16.91	28.65	36.17	35.68	27.21	20.44
	MSE	7.97	16.95	28.66	36.17	35.68	27.21	24.35
$\hat{\sigma}_1$	Bias^2	0.24	0.23	0.19	0.14	0.11	0.05	2.47
	VAR	11.65	11.52	11.79	12.46	13.64	15.55	18.66
	MSE	11.89	11.75	11.99	12.60	13.75	15.59	21.12
$\hat{\sigma}_2$	Bias^2	0.14	0.14	0.12	0.09	0.08	0.05	0.67
	VAR	10.34	10.17	10.45	11.12	12.24	13.96	16.16
	MSE	10.47	10.31	10.57	11.20	12.32	14.00	16.83

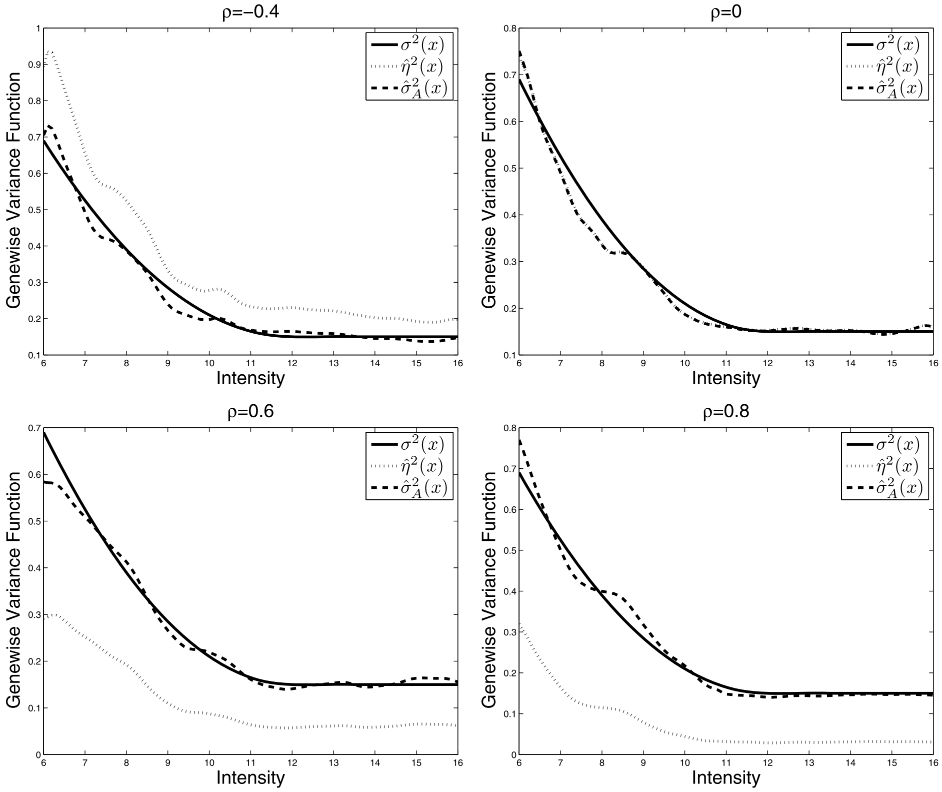

FIG. 2. *Median performance of variance estimators* $\hat{\eta}^2(x)$*,* $\hat{\sigma}^2(x)$ *and* $\hat{\sigma}_A^2(x)$ *when* $\rho = -0.4$*, 0, 0.6 and 0.8.*

4.1. *Validation test.* In the first application, we revisit the validation test as considered in Fan and Niu (2007). For the purpose of the validation tests, we use gProcessedSignal and rProcessedSignal values from Agilent Feature Extraction software as input. We follow the preprocessing scheme described in Patterson et al. (2006) and get 22,144 genes from a total of 41,675 noncontrol genes. Among those, 19 genes with each having 10 replications are used for validation tests. Under the null hypothesis of no experimental biases, a reasonable model is

$$Y_{gi} = \alpha_g + \varepsilon_{gi}, \qquad \varepsilon_{gi} \sim N(0, \sigma_g^2), \qquad i = 1, \ldots, I, g = 1, \ldots, G. \tag{20}$$

We use the notation G to denote the number of genes that have I replications. For our data, $G = 19$ and $I = 10$. Note that G can be different from N, the total number of different genes. The validation test statistics in Fan and Niu (2007) include weighted statistics

$$T_1 = \sum_{g=1}^{G} \left\{ \sum_{i=1}^{I} (Y_{gi} - \bar{Y}_g)^2 \Big/ \sigma_g^2 \right\}, \qquad T_2 = \sum_{g=1}^{G} \left\{ \sum_{i=1}^{I} |Y_{gi} - \bar{Y}_g| \Big/ \sigma_g \right\},$$

and unweighted test statistics

$$T_3 = \left\{ \sum_{g=1}^{G} \sum_{i=1}^{I} (Y_{gi} - \bar{Y}_g)^2 - (I-1) \sum_{g=1}^{G} \sigma_g^2 \right\} \left\{ 2(I-1) \sum_{g=1}^{G} \sigma_g^4 \right\}^{-1/2},$$

$$T_4 = \left\{ \sum_{g=1}^{G} \sum_{i=1}^{I} |Y_{gi} - \bar{Y}_g| - \lambda_I \sum_{g=1}^{G} \sigma_g \right\} \Big/ \left\{ \kappa_I \left(\sum_{g=1}^{G} \sigma_g^2 \right)^{1/2} \right\},$$

where $\lambda_I = \sqrt{2I(I-1)/\pi}$ and $\kappa_I^2 = \operatorname{var}(\sum_{i=1}^{I} |\varepsilon_{gi} - \bar{\varepsilon}_g|/\sigma_g)$. Under the null hypothesis, the test statistic T_1 is χ^2 distributed with degree of freedom $(I-1)G$ and T_2, T_3 and T_4 are all asymptotically normally distributed. As a result, the corresponding p-values can be easily computed.

Here, we apply the same statistics T_1, T_2, T_3 and T_4 but we replace the pooled sample variance estimator by the aggregated local linear estimator

$$\hat{\sigma}_g^2 = \sum_{i=1}^{I} \hat{\sigma}_A^2(X_{gi}) \hat{f}(X_{gi}) \Big/ \sum_{i=1}^{I} \hat{f}(X_{gi}),$$

where $\hat{f}$ is the estimated density function of X_{gi}. The difference between the new variance estimator and the simple pooled variance estimator is that we consider the genewise variance as a nonparametric function of the intensity level. The latter estimator may drag small variances of certain arrays to much higher levels by averaging, resulting in a larger estimated genewise variance and smaller test statistics or bigger p-values.

In the analysis here, we first consider all thirty arrays. The estimated correlation among replicated genes is $\hat{\rho} = 0.69$. The p-values based on the newly estimated genewise variance are depicted in Table 4. As explained in Fan and Niu (2007), T_4 is the most stable test among the four. It turns out that none of the arrays needs further normalization, which is the same as Fan and Niu (2007). Furthermore, we separate the analysis into two groups: the first group using 15 arrays without dye-swap, which has the estimated correlation $\hat{\rho} = 0.66$, and the second group using 15 arrays with dye-swap, resulting in an estimated correlation $\hat{\rho} = 0.34$. The p-values are summarized in Table 5. Results show that array AGL-2-D3 and array AGL-2-D5 need further normalization if 5% significance level applies. The difference is due to decreased estimated ρ for the dye swap arrays and p-values are sensitive to the genewise variance. We also did analysis by separating data into 6

Table 4
Comparison of p-values for $T_1, \ldots, T_4$ for MAQC project data considering all 30 arrays together

	***p*-values**			
Slide name	$\boldsymbol{T_1}$	$\boldsymbol{T_2}$	$\boldsymbol{T_3}$	$\boldsymbol{T_4}$
AGL-1-C1	1.0000	1.0000	1.0000	1.0000
AGL-1-C2	1.0000	1.0000	1.0000	1.0000
AGL-1-C3	1.0000	1.0000	1.0000	1.0000
AGL-1-C4	1.0000	1.0000	1.0000	1.0000
AGL-1-C5	1.0000	1.0000	1.0000	1.0000
AGL-1-D1	1.0000	1.0000	1.0000	1.0000
AGL-1-D2	1.0000	1.0000	1.0000	1.0000
AGL-1-D3	1.0000	1.0000	1.0000	1.0000
AGL-1-D4	1.0000	1.0000	1.0000	1.0000
AGL-1-D5	1.0000	1.0000	1.0000	1.0000
AGL-2-C1	1.0000	1.0000	1.0000	1.0000
AGL-2-C2	1.0000	1.0000	1.0000	1.0000
AGL-2-C3	1.0000	1.0000	1.0000	1.0000
AGL-2-C4	1.0000	1.0000	1.0000	1.0000
AGL-2-C5	1.0000	1.0000	1.0000	1.0000
AGL-2-D1	1.0000	0.9999	0.9996	0.9999
AGL-2-D2	0.8387	0.9011	0.8953	0.9182
AGL-2-D3	0.3525	0.1824	0.3902	0.1905
AGL-2-D4	1.0000	1.0000	1.0000	1.0000
AGL-2-D5	0.8820	0.8070	0.8848	0.7952
AGL-3-C1	1.0000	1.0000	1.0000	1.0000
AGL-3-C2	1.0000	1.0000	1.0000	1.0000
AGL-3-C3	1.0000	1.0000	1.0000	1.0000
AGL-3-C4	1.0000	1.0000	1.0000	1.0000
AGL-3-C5	1.0000	1.0000	1.0000	1.0000
AGL-3-D1	1.0000	1.0000	1.0000	1.0000
AGL-3-D2	1.0000	1.0000	1.0000	1.0000
AGL-3-D3	1.0000	1.0000	1.0000	1.0000
AGL-3-D4	1.0000	1.0000	1.0000	1.0000
AGL-3-D5	1.0000	1.0000	1.0000	1.0000

groups: with and without dye swap, and three sites of experiments. Due to the small sample size, the six estimates of ρ range from 0.08 to 0.74, and we also find that array AGL-2-D3 needs further normalization.

4.2. *Gene selection.* To detect the differentially expressed genes, we follow the filter instruction and get 19,802 genes out of 41,000 unique noncontrol genes as in Patterson et al. (2006), that is, $I = 1$. The dye swap result was averaged before doing the one-sample t-test. Thus, at each site, we have five microarrays.

TABLE 5

Comparison of p-values for $T_1, \ldots, T_4$ for MAQC project data considering the arrays with and without dye-swap separately

	***p*-values**			
Slide name	$\boldsymbol{T_1}$	$\boldsymbol{T_2}$	$\boldsymbol{T_3}$	$\boldsymbol{T_4}$
AGL-1-C1	1.0000	1.0000	1.0000	1.0000
AGL-1-C2	1.0000	1.0000	1.0000	1.0000
AGL-1-C3	1.0000	1.0000	0.9999	1.0000
AGL-1-C4	1.0000	1.0000	1.0000	1.0000
AGL-1-C5	1.0000	1.0000	0.9999	1.0000
AGL-1-D1	1.0000	1.0000	1.0000	1.0000
AGL-1-D2	1.0000	1.0000	1.0000	1.0000
AGL-1-D3	1.0000	1.0000	1.0000	1.0000
AGL-1-D4	1.0000	1.0000	1.0000	1.0000
AGL-1-D5	1.0000	1.0000	1.0000	1.0000
AGL-2-C1	1.0000	1.0000	0.9943	1.0000
AGL-2-C2	1.0000	1.0000	1.0000	1.0000
AGL-2-C3	1.0000	1.0000	1.0000	1.0000
AGL-2-C4	0.0152	0.9493	0.3931	0.9136
AGL-2-C5	1.0000	1.0000	0.8060	1.0000
AGL-2-D1	0.7806	0.8074	0.6622	0.6584
AGL-2-D2	0.2170	0.2984	0.1651	0.2217
AGL-2-D3	**0.0002**	**0.0000**	**0.0001**	**0.0000**
AGL-2-D4	1.0000	1.0000	1.0000	1.0000
AGL-2-D5	**0.1236**	**0.0662**	**0.0669**	**0.0300**
AGL-3-C1	1.0000	1.0000	0.9996	1.0000
AGL-3-C2	1.0000	1.0000	0.9988	1.0000
AGL-3-C3	1.0000	1.0000	0.9977	1.0000
AGL-3-C4	1.0000	1.0000	1.0000	1.0000
AGL-3-C5	1.0000	1.0000	0.9999	1.0000
AGL-3-D1	1.0000	1.0000	1.0000	1.0000
AGL-3-D2	1.0000	1.0000	1.0000	1.0000
AGL-3-D3	1.0000	1.0000	1.0000	1.0000
AGL-3-D4	1.0000	1.0000	1.0000	1.0000
AGL-3-D5	1.0000	1.0000	1.0000	1.0000

For each site, significant genes are selected based on these 5 dye-swaped average arrays. For all $N = 19{,}802$ genes, there are no within-array replications. However, model (2) is still reasonable, in which i indexes the array. Hence, the "within-array correlation" becomes "between-array correlation" and is reasonably assumed as $\rho = 0$.

In our nonparametric estimation for the variance function, all the 19,802 genes are used to estimate the variance function, which gives us enough reason to believe that the estimator $\hat{\sigma}_A^2(x)$ is close to the inherent true variance function $\sigma^2(x)$.

We applied both the t-test and z-test to each gene to see if the logarithm of the expression ratio is zero, using the five arrays collected at each location. The number of differentially expressed genes detected by using the two different tests under three Fold Changes (FC) and four significant levels are given in Table 6. Large numbers of genes are identified as differentially expressed, which is expected when comparing a brain sample and a tissue pool sample [Patterson et al. (2006)]. We can see clearly that the z-test associated with our new variance estimator $\hat{\sigma}_A^2(x)$ leads to more differentially expressed genes. For example, at site 1, using $\alpha = 0.001$, among the fold changes at least 2, t-test picks 8231 genes whereas z-test selects 8875 genes. This gives an empirical power increase of $(8875 - 8231)/19{,}802 \approx 3.25\%$ in the group with observed fold change at least 2.

To verify the accuracy of our variance estimation in the z-test, we compare the empirical power increase with the expected theoretical power increase. The expected theoretical power increase is computed as

$$\mathrm{ave}\{\mathrm{P}_z(\mu_g/\sigma_g) - \mathrm{P}_{t_{n-1}}(\mu_g/\sigma_g)\}, \tag{21}$$

taking the average of power increases across all $\mu_g \neq 0$. However, in the absence of the availability, we replace μ_g by its estimate, which is the sample average of $n = 5$ observed log-expression ratios. Table 7 depicts the results at three different sites, in which the columns "Theo" refer to the expected theoretical power increase defined by (21), with μ_g replaced by $\bar{Y}_g$ and σ_g replaced by its estimate from the genewise variance function, and the columns "Emp" refer to the empirical power increase.

There are two things worth noticing here. First, for expected theoretical power increase, we use the sample mean $\bar{Y}_g = \mu_g + \bar{\epsilon}_g$ instead of the real gene effect μ_g, which is not observable, so it inevitably involves the error term $\bar{\epsilon}_g$.

Table 6
Comparison of the number of significantly differentially expressed genes

		$p < 0.05$		$p < 0.01$		$p < 0.005$		$p < 0.001$	
		t-test	**z-test**	**t-test**	**z-test**	**t-test**	**z-test**	**t-test**	**z-test**
Agilent 1	FC > 1.5	12692	12802	12464	12752	12313	12722	11744	12646
	FC > 2	8802	8879	8654	8872	8556	8869	8231	8858
	FC > 4	3493	3493	3431	3493	3376	3493	3231	3493
Agilent 2	FC > 1.5	12282	12678	11217	12587	10502	12536	8270	12421
	FC > 2	8644	8877	7908	8875	7452	8861	6125	8828
	FC > 4	3600	3649	3188	3649	2964	3649	2422	3649
Agilent 3	FC > 1.5	12502	12692	11994	12576	11694	12519	10788	12374
	FC > 2	8689	8832	8344	8810	8150	8800	7591	8762
	FC > 4	3585	3603	3378	3602	3278	3602	2985	3600

TABLE 7
Comparison of expected theoretical and empirical power difference (in percentage)

	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.005$		$\alpha = 0.001$	
	Theo	**Emp**	**Theo**	**Emp**	**Theo**	**Emp**	**Theo**	**Emp**
Agilent 1	2.52	0.61	6.08	3.75	8.06	5.59	13.66	11.74
Agilent 2	4.03	7.56	10.11	17.61	13.61	22.86	23.75	37.63
Agilent 3	3.02	2.56	7.14	7.39	9.42	10.19	15.94	18.18
Average	3.19	3.58	7.78	9.58	10.36	12.88	17.79	22.51

Second, the power functions $\mathrm{P}_z(\mu)$ and $\mathrm{P}_t(\mu)$ depend sensitively on μ and the tails of the assumed distribution. Despite of these, the expected theoretical and empirical power increases are in the same bulk and the averages are very close. This provides good evidence that our genewise variance estimation has an acceptable accuracy.

We also apply SIMEX and permutation SIMEX methods in Carroll and Wang (2008) to the MAQC data, to illustrate its utility. As mentioned in the Introduction, their model is not really intended for the analysis of two-color microarray data. Should we only use the information on log-ratios (Y), the model is very hard to interpret. In addition, one might question why the information on X (observed intensity levels) is not used at all. Nevertheless, we apply the SIMEX methods of Carroll and Wang (2008) to only the log-ratios Y in the two-color data and produce similar tables to the Tables 6 and 7.

From the results, we have the following understandings. First, all the numbers for z-test in Tables 8 and 9 at all significance levels are approximately the same. In fact, the p-values are very small so that numbers at different significance levels are about the same. That indicates that both SIMEX and permutation SIMEX method are tending to estimate genewise variance very small, making the test statistics large for all the time. On the other hand, our method estimates the genewise variance moderately so that the numbers are not exactly the same for different significance levels. Second, in the implementation, we found that the SIMEX and permutation SIMEX is computationally expensive (more than one hour) while our method only takes a few minutes. Third, from Tables 10 and 11 we can see that the expected theoretical power increase and the empirical ones are reasonably close, which are in lines with our method.

5. Summary.

The estimation of genewise variance function is motivated by the downstream analysis of microarray data such as validation test and selecting statistically differentially expressed genes. The methodology proposed here is novel by using across-array and within-array replications. It

Table 8

Comparison of the number of significantly differentially expressed genes using SIMEX method

		$p < 0.05$		$p < 0.01$		$p < 0.005$		$p < 0.001$	
		t-test	z-test	t-test	z-test	t-test	z-test	t-test	z-test
Agilent 1	FC > 1.5	12692	12820	12464	12820	12313	12820	11744	12820
	FC > 2	8802	8879	8654	8879	8556	8879	8231	8879
	FC > 4	3493	3493	3431	3493	3376	3493	3231	3493
Agilent 2	FC > 1.5	12282	12721	11217	12721	10502	12721	8270	12721
	FC > 2	8644	8878	7908	8878	7452	8878	6125	8878
	FC > 4	3600	3649	3188	3649	2964	3649	2422	3649
Agilent 3	FC > 1.5	12502	12760	11994	12760	11694	12760	10788	12760
	FC > 2	8689	8836	8344	8836	8150	8836	7591	8836
	FC > 4	3585	3603	3378	3603	3278	3603	2985	3603

Table 9

Comparison of the number of significantly differentially expressed genes using permutation SIMEX

		$p < 0.05$		$p < 0.01$		$p < 0.005$		$p < 0.001$	
		t-test	z-test	t-test	z-test	t-test	z-test	t-test	z-test
Agilent 1	FC > 1.5	12692	12820	12464	12820	12313	12820	11744	12820
	FC > 2	8802	8879	8654	8879	8556	8879	8231	8879
	FC > 4	3493	3493	3431	3493	3376	3493	3231	3493
Agilent 2	FC > 1.5	12282	12721	11217	12721	10502	12721	8270	12721
	FC > 2	8644	8878	7908	8878	7452	8878	6125	8878
	FC > 4	3600	3649	3188	3649	2964	3649	2422	3649
Agilent 3	FC > 1.5	12502	12760	11994	12760	11694	12760	10788	12760
	FC > 2	8689	8836	8344	8836	8150	8836	7591	8836
	FC > 4	3585	3603	3378	3603	3278	3603	2985	3603

Table 10

Comparison of expected theoretical and empirical power difference using SIMEX method (in percentage)

	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.005$		$\alpha = 0.001$	
	Theo	Emp	Theo	Emp	Theo	Emp	Theo	Emp
Agilent 1	2.43	2.06	7.17	5.42	10.30	7.34	20.71	13.44
Agilent 2	7.16	3.41	19.20	12.06	26.17	16.90	43.46	30.42
Agilent 3	4.18	2.88	11.71	7.38	16.45	9.89	31.38	17.57
Average	4.59	2.78	12.69	8.29	17.64	11.38	31.85	20.48

TABLE 11

Comparison of expected theoretical and empirical power difference using permutation SIMEX method (in percentage)

	$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.005$		$\alpha = 0.001$	
	Theo	**Emp**	**Theo**	**Emp**	**Theo**	**Emp**	**Theo**	**Emp**
Agilent 1	1.89	2.86	5.66	6.43	8.19	8.59	16.75	15.07
Agilent 2	4.84	7.37	13.44	17.22	18.97	22.50	36.90	37.26
Agilent 3	2.89	4.91	8.34	10.13	11.87	13.11	23.44	21.31
Average	3.20	5.05	9.15	11.26	13.01	14.74	25.70	24.55

does not require any specific assumptions on α_g such as sparsity or smoothness, and hence reduces the bias of the conventional nonparametric estimators. Although the number of nuisance parameters is proportional to the sample size, we can estimate the main interest (variance function) consistently. By increasing the degree of freedom largely, both the validation tests and z-test using our variance estimators are more powerful in identifying arrays that need to be normalized further and more capable of selecting differentially expressed genes.

Our proposed methodology has a wide range of applications. In addition to the microarray data analysis with within-array replications, it can be also applied to the case without within-array replications, as long as the model (2) is reasonable. Our two-way nonparametric model is a natural extension of the Neyman–Scott problem. Therefore, it is applicable to all the problems where the Neyman–Scott problem is applicable.

There are possible extensions. For example, the SIMEX idea can be applied on our model in order to take into account the measurement error. We can also make adaptations to our methods when we have a prior correlation structure among replications other than the identical correlation assumption.

APPENDIX

The following regularity conditions are imposed for the technical proofs:

1. The regression function $\sigma^2(x)$ has a bounded and continuous second derivative.
2. The kernel function K is a bounded symmetric density function with a bounded support.
3. $h \to 0, Nh \to \infty$.
4. $\mathrm{E}[\sigma^8(X)]$ exists and the marginal density $f_X(\cdot)$ is continuous.

We need the following conditional variance–covariance matrix of the random vector $\mathbf{Z}_g$ in our asymptotic study.

LEMMA 1. *Let $\mathbf{\Omega}$ be the variance–covariance matrix of $\mathbf{Z}_g$ conditioning on all data $\mathbf{X}$. Then, respectively, the diagonal and off-diagonal elements are*

$$\begin{aligned}\Omega_{ii} &= 2\sigma^4(X_{gi}) + \frac{2}{(I-1)^2(I-2)^2}\sum_{k\neq l}\sigma^2(X_{gk})\sigma^2(X_{gl}) \\ &\quad + \frac{4(I-3)}{(I-1)(I-2)^2}\sigma^2(X_{gi})\sum_{j\neq i}\sigma^2(X_{gj}), \qquad i=1,\ldots,I,\end{aligned} \tag{22}$$

$$\begin{aligned}\Omega_{ij} &= \frac{4}{(I-1)^2}\sigma^2(X_{gi})\sigma^2(X_{gj}) \\ &\quad + \frac{2}{(I-1)^2(I-2)^2}\sum_{\substack{k\neq l \\ k,l\neq i,j}}\sigma^2(X_{gk})\sigma^2(X_{gl}) \\ &\quad - \frac{4}{(I-1)^2(I-2)}\sum_{k\neq i,j}\sigma^2(X_{gk})(\sigma^2(X_{gi})+\sigma^2(X_{gj})), \\ &\qquad\qquad i,j=1,\ldots,I, i\neq j.\end{aligned} \tag{23}$$

PROOF. Let $\mathbf{A}$ be the variance–covariance matrix of $\mathbf{r}_g$ conditioning on all data $\mathbf{X}$. By direct computation, the diagonal elements are given by

$$\begin{aligned}A_{ii} &= \mathrm{var}[(Y_{gi}-\bar{Y}_g)^2|\mathbf{X}] \\ &= \frac{2(I-1)^4}{I^4}\sigma^4(X_{gi}) + \frac{4(I-1)^2}{I^4}\sum_{k\neq i}\sigma^2(X_{gi})\sigma^2(X_{gk}) + \frac{2}{I^4}\sum_{k\neq i}\sigma^4(X_{gk}) \\ &\quad + \frac{4}{I^4}\sum_{l,k\neq i,l<k}\sigma^2(X_{gl})\sigma^2(X_{gk}), \qquad i=1,\ldots,I,\end{aligned} \tag{24}$$

and the off-diagonal elements are given by

$$\begin{aligned}A_{ij} &= \mathrm{cov}\{[(Y_{gi}-\bar{Y}_g)^2,(Y_{gj}-\bar{Y}_g)^2]|\mathbf{X}\} \\ &= \frac{2(I-1)^2}{I^4}[\sigma^4(X_{gi})+\sigma^4(X_{gj})] + \frac{4(I-1)^2}{I^4}\sigma^2(X_{gi})\sigma^2(X_{gj}) \\ &\quad - \frac{4(I-1)}{I^4}\sum_{k\neq i,j}\sigma^2(X_{gk})(\sigma^2(X_{gi})+\sigma^2(X_{gj})) \\ &\quad + \frac{4}{I^4}\sum_{k,l\neq i,j;l<k}\sigma^2(X_{gl})\sigma^2(X_{gk}) + \frac{2}{I^4}\sum_{k\neq i,j}\sigma^4(X_{gk}).\end{aligned} \tag{25}$$

Using $\mathbf{\Omega} = \mathbf{BAB}^T$, we can obtain the result by direct computation. □

The proofs for Theorems 1 and 3 follow a similar idea. Since Theorem 1 does not involve a lot of coefficients, we will show the proof of Theorem 1 and explain the difference in the proof of Theorem 3.

PROOF OF THEOREM 1. First of all, the bias of $\eta_i^2(x)$ comes from the local linear approximation. Since $\{(X_{gi}, Z_{gi})\}_{g=1}^N$ is an i.i.d. sequence, by (4) and the result in Fan and Gijbels (1996), it follows that

$$\mathrm{E}\{\eta_i^2(x)|\mathbf{X}\} = \sigma^2(x) + b(x) + o_P(h^2).$$

Similarly, the asymptotic variance of $\eta_i^2(x)$ also follows from Fan and Gijbels (1996).

We now prove the off-diagonal elements in matrix $\mathrm{var}[\boldsymbol{\eta}|\mathbf{X}]$

$$\mathrm{cov}[(\hat\eta_i^2(x), \hat\eta_j^2(x))|\mathbf{X}] = V_2 + o_P(1/N). \tag{26}$$

Recalling that $\hat\eta_i^2(x) = \sum_{g=1}^N W_{N,i}((X_{gi} - x)/h)Z_{gi}$, we have

$$\mathrm{cov}[(\hat\eta_i^2(x), \hat\eta_j^2(x))|\mathbf{X}] = \sum_{g=1}^N W_{N,i}\left(\frac{X_{gi}-x}{h}\right) W_{N,j}\left(\frac{X_{gj}-x}{h}\right) \mathrm{cov}[(Z_{gi}, Z_{gj})|\mathbf{X}].$$

The equality follows by the fact that $\mathrm{cov}[(Z_{gi}, Z_{g'j})|\mathbf{X}] = 0$ when $g \neq g'$. Recall $\Omega_{ij} = \mathrm{cov}[(Z_{gi}, Z_{gj})|\mathbf{X}]$ and define $R_{N,g} = N \cdot W_{N,j}((X_{gj} - x)/h)\Omega_{ij}$. Thus,

$$N \cdot \mathrm{cov}[(\hat\eta_i^2(x), \hat\eta_j^2(x))|\mathbf{X}] = \sum_{g=1}^N W_{N,i}\left(\frac{X_{gi}-x}{h}\right) R_{N,g}. \tag{27}$$

The right-hand side of (27) can be seen as local linear smoother of the synthetic data $\{(X_{gi}, R_{N,g})\}_{g=1}^N$. Although $R_{N,g}$ involves N at the first glance, its conditional expectation $\mathrm{E}[R_{N,g}|X_{gi} = x]$ and conditional variance $\mathrm{var}[R_{N,g}|X_{gi} = x]$ do not grow with N. Since $\{(X_{gi}, R_{N,g})\}_{g=1}^N$ is an i.i.d. sequence, by the results in Fan and Gijbels (1996), we obtain

$$N \cdot \mathrm{cov}[(\hat\eta_i^2(x), \hat\eta_j^2(x))|\mathbf{X}] = \mathrm{E}[R_{N,g}|X_{gi} = x] + o_P(1).$$

To calculate $\mathrm{E}[R_{N,g}|X_{gi} = x]$, we apply the approximation $W_{N,i}(u) = K(u)(1 + o_P(1))/(Nhf_X(x))$ in the example of Fan and Gijbels [(1996), page 64] and have the following arguments

$$\begin{aligned}
&\mathrm{E}[R_{N,g}|X_{gi} = x] \\
&\quad = \mathrm{E}\left[N \cdot \frac{1}{Nhf_X(x)} hK_h(X_{gj} - x)\Omega_{ij}|X_{gi} = x\right](1 + o_P(1)) \\
&\quad = (f_X(x))^{-1} \int K(u)\Omega_{ij}|_{X_{gi}=x}(x + hu, \mathbf{s}) f_X(x + hu)\, du\, d\mathbf{s} + o_P(1) \\
&\quad = NV_2 + o_P(1),
\end{aligned}$$

where $\mathbf{s}$ represents all the integrating variables corresponding to $X_{g1},\dots,X_{gI}$ except X_{gi} and X_{gj}. That justifies (26).

To prove the multivariate asymptotic normality

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\eta}-(\sigma^2(x)+b(x)+o_P(h^2))\mathbf{e})\xrightarrow{D}N(0,\mathbf{I}_I), \tag{28}$$

we employ Cramér–Wold device: for any unit vector $\mathbf{a}=(a_1,\dots,a_I)^T$ in $\mathbb{R}^I$,

$$F^*\overset{\triangle}{=}\{\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}\}^{-1/2}\left\{\sum_{i=1}^I a_i\sum_{g=1}^N W_{N,i}\left(\frac{X_{gi}-x}{h}\right)(Z_{gi}-\sigma^2(Z_{gi}))\right\}\xrightarrow{D}N(0,1).$$

Denote by $Q_{g,i}=W_{N,i}((X_{gi}-x)/h)(Z_{gi}-\sigma^2(X_{gi}))$ and $\widetilde{Q}_g=\sum_{i=1}^I a_iQ_{g,i}$. Note that the sequence $\{\widetilde{Q}_g\}_{g=1}^N$ is i.i.d. distributed. To show the asymptotic normality of F^*, it is sufficient to check Lyapunov's condition:

$$\lim_{N\to\infty}\frac{\sum_{g=1}^N \mathrm{E}[|\widetilde{Q}_g|^4|\mathbf{X}]}{(\sum_{g=1}^N \mathrm{E}[|\widetilde{Q}_g|^2|\mathbf{X}])^2}=0.$$

To facilitate the presentation, we first note that sequences $\{Q_{g,i}\}_{g=1}^N$ are i.i.d. and satisfy Lyapunov's condition for each fixed i. Denote $\delta_{N,i}^2=\sum_{g=1}^N \mathrm{E}[|Q_{g,i}|^2|\mathbf{X}]$. And recall that $\delta_{N,i}^2=\operatorname{var}[\hat{\eta}_i^2(x)|\mathbf{X}]=O_P((Nh)^{-1})$. Let c^* be a generic constant which may vary from one line to another. We have the following approximation:

$$\begin{aligned}\sum_{g=1}^N \mathrm{E}[|Q_{g,i}|^4|\mathbf{X}]&=c^*N^{-3}\mathrm{E}\{K_h^4(X_{gi}-x)[(Z_{gi}-\sigma^2(X_{gi}))^4|\mathbf{X}]\}(1+o_P(1))\\&=O_P((Nh)^{-3}).\end{aligned}$$

Therefore, $\sum_{g=1}^N \mathrm{E}[|Q_{g,i}|^4|\mathbf{X}]=o(\delta_{N,i}^4)$. By the marginal Lyapunov conditions, we have the following inequality:

$$\sum_{g=1}^N \mathrm{E}[\widetilde{Q}_g^4|\mathbf{X}]\le c^*\sum_{i=1}^I\sum_{g=1}^N \mathrm{E}[|Q_{g,i}|^4|\mathbf{X}]=c^*I\cdot o_P((Nh)^{-2})=o_P((Nh)^{-2}).$$

For the denominator, we have the following arguments:

$$\begin{aligned}\sum_{g=1}^N \mathrm{E}[|\widetilde{Q}_g|^2|\mathbf{X}]&=\sum_i a_i^2\sum_{g=1}^N \mathrm{E}[Q_{g,i}^2|\mathbf{X}]+\sum_{i\ne j}a_ia_j\sum_{g=1}^N \mathrm{E}[Q_{g,i}Q_{g,j}|\mathbf{X}]\\&=\sum_i a_i^2\operatorname{var}[\hat{\eta}_i^2(x)|\mathbf{X}]+\sum_{i\ne j}a_ia_j\operatorname{cov}[(\hat{\eta}_i^2(x),\hat{\eta}_j^2(x))|\mathbf{X}]\\&\overset{*}{=}O_P((Nh)^{-1})+O_P(N^{-1})\\&=O_P((Nh)^{-1}).\end{aligned}$$

Note that the second to last equality holds by the asymptotic conditional variance–covariance matrix $\mathbf{\Sigma}$. Therefore Lyapunov's condition is justified. That completes the proof. □

Proof of Theorem 2. First of all, for each given g,

$$\mathrm{E}s_{B,g}^2 = I\operatorname{var}(\bar{Y}_{gj}) = \sigma_2 + \rho(I-1)\sigma_1^2.$$

Note that by (8), we have

$$\begin{aligned}\mathrm{E}(Y_{gij}-\bar{Y}_{gj})^2 &= I^{-2}[I(I-1)\sigma_2 + \rho(I-1)(I-2)\sigma_1^2 - 2(I-1)^2\rho\sigma_1^2]\\ &= I^{-1}(I-1)(\sigma_2-\rho\sigma_1^2).\end{aligned}$$

Thus, for all g, we have

$$\mathrm{E}s_{W,g}^2 = \sigma_2 - \rho\sigma_1^2.$$

Since $\{s_{B,g}^2\}$ and $\{s_{W,g}^2\}$ are i.i.d. sequences across the N genes, by the central limit theorem, we have

$$\begin{aligned}\frac{1}{N}\sum_{g=1}^N s_{B,g}^2 &= \sigma_2 + \rho(I-1)\sigma_1^2 + O_P(N^{-1/2}),\\ \frac{1}{N}\sum_{g=1}^N s_{W,g}^2 &= \sigma_2 - \rho\sigma_1^2 + O_P(N^{-1/2}).\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\rho}_0 &= \frac{\sigma_2 + \rho(I-1)\sigma_1^2 - \sigma_2 + \rho\sigma_1^2 + O_P(N^{-1/2})}{\sigma_2 + \rho(I-1)\sigma_1^2 + (I-1)(\sigma_2 - \rho\sigma_1^2) + O_P(N^{-1/2})}\\ &= \rho\sigma_1^2/\sigma_2 + O_P(N^{-1/2}).\end{aligned}$$ □

Proof of Theorem 3. Note that

$$\begin{aligned}\operatorname{var}[\hat{\eta}_A^2(x)|\mathbf{X}] &= \sum_{g=1}^N\sum_{i=1}^I W_N^2\left(\frac{X_{gi}-x}{h}\right)\operatorname{var}[Z_{gi}|\mathbf{X}]\\ &\quad + \sum_{g=1}^N\sum_{i\neq j}^I W_N\left(\frac{X_{gi}-x}{h}\right)W_N\left(\frac{X_{gj}-x}{h}\right)\operatorname{cov}[(Z_{gi},Z_{gj})|\mathbf{X}].\end{aligned}$$

Following similar steps in the proof of Theorem 1, one can verify $\operatorname{var}[\hat{\eta}_A^2(x)|\mathbf{X}] = V_1'/I + (1-1/I)V_2' + o_P((Nh)^{-1})$, where the coefficients $C_2,\ldots,C_4,D_0$,

$\ldots, D_4$ are as follows:

$$C_2 = \frac{4(1+\rho^2)\sigma_2 + [4\rho(I-2) + 4\rho^2(2I-3)]\sigma_1^2}{I-1},$$

$$C_3 = -\frac{8\rho^2(I-3)\sigma_1^3 + 8(\rho^2+\rho)\sigma_1\sigma_2}{I-1},$$

$$C_4 = \frac{2}{(I-1)(I-2)}\{(1+\rho^2)\sigma_2^2 + 2(\rho^2+\rho)(I-3)\sigma_1^2\sigma_2 \\ + (I-3)(I-4)\rho^2\sigma_1^4\},$$

$$D_0 = 2\left(\rho^2 - \frac{4\rho}{I-1} + \frac{2(1+\rho^2)}{(I-1)^2}\right),$$

$$D_1 = \frac{8}{(I-1)^2}\{(2I-4)\rho - (I^2-4I+5)\rho^2\}\sigma_1,$$

$$D_2 = \frac{4}{(I-1)^2(I-2)}\{(I-3)^2\rho^2 + ((I-2)^2+1)\rho - 2(I-2)\}\sigma_2 \\ + \frac{4(I-3)}{(I-1)^2(I-2)}\{(3(I-2)(I-3)+2)\rho^2 - 2(I-2)\rho\}\sigma_1^2,$$

$$D_3 = -\frac{8(I-3)^2}{(I-1)^2(I-2)}\{(\rho^2+\rho)\sigma_1\sigma_2 + (I-4)\rho^2\sigma_1^3\},$$

$$D_4 = \frac{4}{(I-1)^2(I-2)^2} \\ \times \left\{(1+\rho^2)\binom{I-2}{2}\sigma_2^2 \right. \\ \left. + 6(\rho^2+\rho)\binom{I-2}{3}\sigma_1^2\sigma_2 + 12\rho^2\binom{I-2}{4}\sigma_1^4\right\}. \qquad \square$$

Acknowledgments. The authors thank the Editor, the Associate Editor and two referees, whose comments have greatly improved the scope and presentation of the paper.

REFERENCES

Carroll, R. J. and Wang, Y. (2008). Nonparametric variance estimation in the analysis of microarray data: A measurement error approach. *Biometrika* **95** 437–449. MR2422697

Cui, X., Hwang, J. T. and Qiu, J. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **6** 59–75.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London. MR1383587

Fan, J. and Niu, Y. (2007). Selection and validation of normalization methods for c-DNA microarrays using within-array replications. *Bioinformatics* **23** 2391–2398.

FAN, J., PENG, H. and HUANG, T. (2005). Semilinear high-dimensional model for normalization of microarray data: A theoretical analysis and partial consistency (with discussion). *J. Amer. Statist. Assoc.* **100** 781–813. MR2201010

FAN, J. and REN, Y. (2007). Statistical analysis of DNA microarray data in cancer research. *Clinical Cancer Research* **12** 4469–4473.

FAN, J., TAM, P., VANDE WOUDE, G. and REN, Y. (2004). Normalization and analysis of cDNA micro-arrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc. Natl. Acad. Sci. USA* **101** 1135–1140.

HUANG, J., WANG, D. and ZHANG, C. (2005). A two-way semi-linear model for normalization and significant analysis of cDNA microarray data. *J. Amer. Statist. Assoc.* **100** 814–829. MR2201011

KAMB, A. and RAMASWAMI, A. (2001). A simple method for statistical analysis of intensity differences in microarray-deried gene expression data. *BMC Biotechnology* **1** 8.

NEYMAN, J. and SCOTT, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113

PATTERSON, T. ET AL. (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology* **24** 1140–1150.

RUPPERT, D., WAND, M. P., HOLST, U. and HÖSSJER, O. (1997). Local polynomial variance function estimation. *Technometrics* **39** 262–273. MR1462587

SMYTH, G., MICHAUD, J. and SCOTT, H. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21** 2067–2075.

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856

TONG, T. and WANG, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *J. Amer. Statist. Assoc.* **102** 113–122. MR2293304

TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98** 5116–5121.

WANG, Y., MA, Y. and CARROLL, R. J. (2009). Variance estimation in the analysis of microarray data. *J. Roy. Statist. Soc. Ser. B* **71** 425–445.

J. FAN
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu

Y. FENG
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1255 AMSTERDAM AVENUE, 10TH FLOOR
NEW YORK, NEW YORK 10027
USA
E-MAIL: fy2158@columbia.edu

Y. S. NIU
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ARIZONA
617 N. SANTA RITA AVE.
P.O. BOX 210089
TUCSON, ARIZONA 85721-0089
USA
E-MAIL: yueniu@math.arizona.edu