Identifying Differentially Expressed Genes in Microarray Experiments With Model-Based Variance Estimation

Xiaodong Cai, *Senior Member, IEEE,* and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Statistical tests have been employed to identify genes differentially expressed under different conditions using data from microarray experiments. The variance of gene expression levels is often required in various statistical tests; however, due to the small number of replicates, the variance estimated from the sample variance is not accurate, which causes large false positive and negative errors. More accurate and robust variance estimation is thus highly desirable to improve the performance of statistical tests. In this paper, cluster analysis was performed on the microarray data using a model-based clustering method. The variance for each gene was then estimated from cluster variances. Since cluster variances are estimated from multiple genes whose microarray data have similar variance, the proposed estimation method pools the relevant genes together; this effectively increases the number of samples in variance estimation, thereby improving variance estimation. Using simulated data, it is shown that with the novel variance estimation, the performance of the t-test, regularized t-test, and a variant of SAM test, which is called the S-test here, can be improved. Using colon microarray data of Alon *et al.*, it is demonstrated that the proposed method offers better or comparable performance compared with other gene pooling methods. Using the IHF microarray data of Arfin *et al.*, it is shown that the proposed novel variance estimation decreases the significance of those genes having a small fold change but a high significant score assigned by the t-test using the sample variance, which potentially reduces false positive probability.**

***Index Terms*—Clustering, micorarray, mixture model, statistical test, variance estimation.**

I. Introduction

MICROARRAY technology has emerged as a widely used tool in genomic research [1], [2]. It allows for simultaneously measuring the levels of thousands of different RNA molecules expressed from various genes. Studying these measurements facilitates understanding the biological processes present in living organisms. One application of microarray technology deals with identifying genes with different expression levels under different conditions or at different time points. An early method for identifying different gene expressions is based on a simple fold change, where genes whose measured expression levels differ by more than an arbitrary cutoff value between different conditions are considered to be differentially expressed [3]. More sophisticated statistical methods have been developed to account for the statistical variability of gene expressions (see [4] and references therein). The t-test provides a simple statistical method for identifying differentially expressed genes [5]. The t-test statistic is obtained by dividing the sample mean by the standard deviation estimated from the sample variance. Due to the small sample size, the sample variances estimated from each gene are unstable. If the estimated variance for a gene is much smaller than the true variance, the t-value can be large even if the fold change is very small, which causes large false discovery probability. To cope with this problem, various statistical methods have been proposed. In the method of significance analysis of microarray (SAM) [6] and the empirical Bayesian (EB) method [7], a small positive constant, which is calculated from microarray data across all genes, is added to the denominator of the gene-specific t-statistic. This somehow alleviates the aforementioned problem, because the t-value will not be "amplified" by a small estimated variance. Assuming that the variance of errors in the microarray data is an identical and independently distributed (i.i.d.) random variable, the Bayesian method can exploit the global information in all genes to form gene-specific test statistics [8]–[12]. Variance is estimated in [8], [11], and [12] using the Bayesian method, and is then used in forming the test statistic. The *a posteriori* likelihood ratio between the null and alternative hypotheses for each gene can also be derived under the Bayesian framework and be used as a test statistic [7], [9], [10].

The key to handling the problem of the unstable variance estimated from a small number of samples is to efficiently utilize the global information, while accounting for the possible heterogeneity in variance across genes, as explored in the SAM and the Bayesian method. In this paper, we develop an alternative approach to coping with this problem, using model-based variance estimation. Specifically, we first perform cluster analysis on microarray data using the model-based clustering method [13], [14]. We then estimate the variance for each gene from cluster variances. Since cluster variances are estimated from multiple genes whose microarray data have similar variance, our estimation method pools the relevant genes together; this effectively increases the number of samples in variance estimation, thereby improving variance estimation. Using simulated data, we will show that with our novel variance estimation, we can improve the performance of the t-test, regularized t-test, and

Manuscript received May 1, 2005; revised December 15, 2005. This paper appeared in part in the *Proceeding of IEEE International Workshop on Genomic Signal Processing and Statistics, 2005*. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Edward Dougherty.

X. Cai is with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124 USA (e-mail: x.cai@miami.edu).

G. B. Giannakis is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, 55455 USA (e-mail: georgios@ece.umn.edu).

Digital Object Identifier 10.1109/TSP.2006.873733

a variant of SAM test, which we call the S-test. Cluster analysis has been applied to microarray data to identify the genes with similar patterns of expression. The model-based and several other clustering methods have been used for this purpose [2], [15]–[18]. Different from [15]–[17], we here capitalize on model-based cluster analysis to estimate the variance for each gene and use this estimated variance to identify differentially expressed genes. Several variance estimation methods pooling genes were proposed in [19] and [20] under the assumption that genes with similar expression level have similar variance. Our model-based variance estimator pools genes by exploiting a normal mixture model for the distribution of data, thereby utilizing more information. The gene pooling methods in [12], [19], and [20] are compared in [21].

II. Methods

We consider one-color microarrays such as Affymetrix oligonucleotide arrays, although our method is also applicable to the two-color microarrays such as spotted cDNA arrays. Suppose $x_{ij}, i = 1,2,\cdots,n, j = 1,2,\cdots,m_c,m_c+1,\cdots,m$ represents the expression level of gene i in array j, where the first m_c and the last $m_t = m - m_c$ arrays or replicates are obtained under the control and treatment conditions, respectively. To test the null hypothesis H_i of equal mean expression for gene i under two conditions, one often uses a Welch two-sample t-statistic [5]

$$t_i = \frac{\bar{x}_{i,t} - \bar{x}_{i,c}}{\sqrt{s_{i,t}^2/m_t + s_{i,c}^2/m_c}} \tag{1}$$

where $\bar{x}_{i,c} = (1/m_c)\sum_{j=1}^{m_c} x_{ij}$ and $\bar{x}_{i,t} = (1/m_t)\sum_{j=m_c+1}^{m_c+m_t} x_{ij}$ denote the sample means of expression levels of gene i under control and treatment conditions, respectively, and $s_{i,c}^2 = (1/(m_c-1))\sum_{j=1}^{m_c}(x_{ij} - \bar{x}_{i,c})^2$ and $s_{i,t}^2 = (1/(m_t-1))\sum_{j=m_c+1}^{m}(x_{ij} - \bar{x}_{i,t})^2$ denote sample variances of gene i under control and treatment conditions, respectively. If data $\{x_{ij}\}$ are normally distributed with equal variance, then t_i follows a student distribution, and the p-value can be calculated from the student distribution. Otherwise, the p-value needs to be estimated from an approximate distribution of t_i, or using other techniques such as permutation methods [5], [7], [22]. Note that the denominator of (1) is the estimated standard deviation of the numerator $\bar{x}_{i,t} - \bar{x}_{i,c}$ obtained from the sample variances $s_{i,c}^2$ and $s_{i,t}^2$, and only the data x_{ij}, $j = 1,\cdots,m$ are used in estimating these sample variances. Due to the small values of m_c and m_t, the sample variances $s_{i,c}^2$ and $s_{i,t}^2$ may not be accurate estimates of the true variances, which causes large probability of false positive and negative errors. Since the data set $\{x_{ij}\}$ for n genes is available and some genes may have similar variances, if we can identify those genes with similar variances, we can pool these genes together and improve the estimate of variance for each gene. To this end, we will perform cluster analysis of the data set $\{x_{ij}\}$ and obtain more accurate estimates of variances. We will then use these estimated variances to improve the performance of the t-test, as well as other tests, including the regularized t-test and the S-test.

A. Variance Estimation Based on the Mixture Model

Defining the row vector $\mathbf{x}_i = [x_{i1},\ldots,x_{im}]$, we assume that $\mathbf{x}_i$, $i = 1,\ldots,n$ are ..i.d. random vectors following a normal mixture model, as in [15]–[17]. The probability density function (pdf) of $\mathbf{x}_i$ is given by

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i|\mathbf{u}_k,\boldsymbol{\Sigma}_k) \tag{2}$$

where $0 \le \pi_k \le 1$ with $\sum_{k=1}^{K}\pi_k = 1$ is the mixing proportion of cluster k, and $f_k(\mathbf{x}_i|\mathbf{u}_k,\boldsymbol{\Sigma}_k)$ is the normal pdf with mean $\mathbf{u}_k$ and covariance $\boldsymbol{\Sigma}_k$

$$f_k(\mathbf{x}_i|\mathbf{u}_k,\boldsymbol{\Sigma}_k) = (2\pi)^{-m/2}|\boldsymbol{\Sigma}_k|^{-1/2} \times \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{u}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \mathbf{u}_k)^T\right]. \tag{3}$$

Depending on the parameters $\{\pi_k\}_{k=1}^K$, $\{\boldsymbol{\mu}_k\}_{k=1}^K$, $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$, and K, the normal mixture model can capture a wide range of distributions.

The EM algorithm of [23] can be used to fit $\mathbf{x}_i$, $i = 1,\cdots,n$ to this mixture model, and estimate parameters $\{\pi_k\}_{k=1}^K$, $\{\boldsymbol{\mu}_k\}_{k=1}^K$, and $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$, as advocated in [13], [15]–[17]. Since we need some parameters estimated by the EM algorithm in our variance estimation, we list the steps of the EM algorithm as follows [13], [14].

1) Initialize $\hat{p}_{ik}$, the probability that $\mathbf{x}_i$ belongs to cluster k, using e.g., a hierarchical clustering method [18].
2) M-step: Compute the maximum-likelihood parameter estimates given $\hat{p}_{ik}$, as follows:

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n}\hat{p}_{ik}}{n} \tag{4}$$

$$\hat{\mathbf{u}}_k = \frac{\sum_{i=1}^{n}\hat{p}_{ik}\mathbf{x}_i}{n\hat{\pi}_k} \tag{5}$$

and compute $\hat{\boldsymbol{\Sigma}}_k$ using (8), (9), or (10), depending on the model.

3) E-step: Compute $\hat{p}_{ik}$ given the parameter estimates from the M-step

$$\hat{p}_{ik} = \frac{\hat{\pi}_k f_k(\mathbf{x}_i|\hat{\mathbf{u}}_k,\hat{\boldsymbol{\Sigma}}_k)}{\sum_{j=1}^{K}\hat{\pi}_j f_j(\mathbf{x}_i|\hat{\mathbf{u}}_j,\hat{\boldsymbol{\Sigma}}_j)}. \tag{6}$$

4) Repeat the M- and E-steps until convergence criteria are satisfied.

Various models have been proposed for the structure of the covariance matrix $\boldsymbol{\Sigma}_k$ [24]. Since we are interested in estimating the cluster variance, we can use a diagonal matrix for $\boldsymbol{\Sigma}_k$. Specifically, we consider the following three models: 1) $\boldsymbol{\Sigma}_k = \lambda_k\mathbf{I}_m$, where $\mathbf{I}_m$ denotes the identity matrix of size m, 2) $\boldsymbol{\Sigma}_k = \mathrm{diag}(\lambda_{k,c}\mathbf{I}_{m_c},\lambda_{k,t}\mathbf{I}_{m_t})$, and 3) $\boldsymbol{\Sigma}_k = \mathrm{diag}(\lambda_{k1},\lambda_{k2},\cdots,\lambda_{km})$. Model 1 assumes that genes in the same cluster have the same variance for all replicates; model 2 assumes that a gene has different variances under control and treatment conditions; and model 3 allows for different

replicates to have different variances. Letting $n_k = \sum_{i=1}^{n} \hat{p}_{ik}$, we can estimate the unrestricted $\mathbf{\Sigma}_k$ as

$$\mathbf{W}_k = \frac{1}{n_k} \sum_{i=1}^{n} \hat{p}_{ik} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k). \tag{7}$$

If $[\mathbf{W}_k]_{ij}$ denotes the entry of $\mathbf{W}_k$ on the ith row and the jth column, we have the following estimate of $\mathbf{\Sigma}_k$ for model 1:

$$\begin{aligned} \hat{\lambda}_k &= \frac{1}{m} \sum_{j=1}^{m} [\mathbf{W}_k]_{jj} \\ \hat{\mathbf{\Sigma}}_k &= \hat{\lambda}_k \mathbf{I}_m \end{aligned} \tag{8}$$

the following for model 2:

$$\begin{aligned} \hat{\lambda}_{k,c} &= \frac{1}{m_c} \sum_{j=1}^{m_c} [\mathbf{W}_k]_{jj} \\ \hat{\lambda}_{k,t} &= \frac{1}{m_t} \sum_{j=m_c+1}^{m} [\mathbf{W}_k]_{jj} \\ \hat{\mathbf{\Sigma}}_k &= \operatorname{diag}\left(\hat{\lambda}_{k,c} \mathbf{I}_{m_c}, \hat{\lambda}_{k,t} \mathbf{I}_{m_t} \right) \end{aligned} \tag{9}$$

and the following for model 3:

$$\begin{aligned} \hat{\lambda}_{kj} &= [\mathbf{W}_k]_{jj}, \qquad j = 1, \cdots, m \\ \hat{\mathbf{\Sigma}}_k &= \operatorname{diag}(\hat{\lambda}_{k1}, \cdots, \hat{\lambda}_{km}). \end{aligned} \tag{10}$$

The selection of the model and the number of clusters can be carried out by using various criteria, of which the Bayesian information criterion (BIC) is favored in some empirical studies [14] and was used in cluster analysis for microarray data [15]–[17]. The BIC value is given by

$$\text{BIC} = 2 \log \left[L(\hat{\boldsymbol{\theta}}) \right] - n_p \log(n) \tag{11}$$

where $L(\hat{\boldsymbol{\theta}})$ is the likelihood of the data with the estimated parameters contained in $\hat{\boldsymbol{\theta}}$ and n_p is the number of independent parameters. According to the BIC, the model and the number of clusters can be chosen corresponding to the first local maximum of the BIC value [14].

If the estimated number of clusters is $\hat{K}$, then the variance of x_{ij} estimated from the cluster analysis is given by the following for model 1:

$$\hat{\sigma}_{ij}^2 = \hat{\sigma}_{i,c}^2 = \hat{\sigma}_{i,t}^2 \triangleq \sum_{k=1}^{\hat{K}} \hat{p}_{ik} \hat{\lambda}_k, \qquad j = 1, \cdots, m \tag{12}$$

the following for model 2:

$$\begin{aligned} \hat{\sigma}_{ij}^2 = \hat{\sigma}_{i,c}^2 &\triangleq \sum_{k=1}^{\hat{K}} \hat{p}_{ik} \hat{\lambda}_{k,c}, \qquad j = 1, \cdots, m_c \\ \hat{\sigma}_{ij}^2 = \hat{\sigma}_{i,t}^2 &\triangleq \sum_{k=1}^{\hat{K}} \hat{p}_{ik} \hat{\lambda}_{k,t}, \qquad j = m_c + 1, \cdots, m \end{aligned} \tag{13}$$

and the following for model 3:

$$\hat{\sigma}_{ij}^2 = \sum_{k=1}^{\hat{K}} \hat{p}_{ik} \hat{\lambda}_{kj}, \qquad j = 1, \cdots, m. \tag{14}$$

Note that the actual estimated variance for x_{ij} is chosen from (12), (13), or (14) according to the BIC, as we mentioned before.

B. t-Test With Model-Based Variance Estimation

Using the estimated variance $\hat{\sigma}_{ij}^2$ for x_{ij} given in (12), (13), or (14), we obtain the estimated variance of $\bar{x}_{i,t} - \bar{x}_{i,c}$ as

$$\hat{\sigma}_i^2 = \frac{1}{m_c^2} \sum_{j=1}^{m_c} \hat{\sigma}_{ij}^2 + \frac{1}{m_t^2} \sum_{j=m_c+1}^{m} \hat{\sigma}_{ij}^2. \tag{15}$$

Replacing the denominator of (1) with $\hat{\sigma}_i$, we obtain the following t^m-statistic[1]:

$$t_i^m = \frac{\bar{x}_{i,t} - \bar{x}_{i,c}}{\hat{\sigma}_i}. \tag{16}$$

Since the distribution of t_i^m is unknown, one can use a permutation method to estimate the distribution of t_i^m under the null hypothesis and then find the p-value [5], [22]. For the t^m-statistic in (16), we will use balanced permutation to form null scores for all genes and then estimate the null distribution as advocated in [7]. To increase the granularity of the estimated null distribution, we can also fit the null scores to a normal mixture model as proposed in [25]. After we obtain an estimate of the null distribution, we can easily find the p-value.

C. Regularized t-Test

In the regularized t-test [8], the estimated variance for a gene is a weighted sum of the *a priori* variance and the sample variance, and the estimated variance for $\bar{x}_{i,t} - \bar{x}_{i,c}$ is given by

$$\hat{\sigma}_i^2 = \frac{v_{0,c} \tilde{\sigma}_{0,c}^2 + (m_c - 1) s_{i,c}^2}{v_{0,c} + m_c - 2} + \frac{v_{0,t} \tilde{\sigma}_{0,t}^2 + (m_t - 1) s_{i,t}^2}{v_{0,t} + m_t - 2} \tag{17}$$

where $\tilde{\sigma}_{0,c}^2$ and $\tilde{\sigma}_{0,t}^2$ are prior variances under the control and treatment conditions, respectively. If N denotes the number of samples needed to reliably estimate the variance, the parameters $v_{0,c}$ and $v_{0,t}$ in (17) are given by $v_{0,c} = N - m_c$ and $v_{0,t} = N - m_t$. The prior variances $\tilde{\sigma}_{0,c}^2$ and $\tilde{\sigma}_{0,t}^2$ are obtained as follows [8]. The mean expression levels of all genes under two conditions are ranked and then $\tilde{\sigma}_{0,c}^2$ and $\tilde{\sigma}_{0,t}^2$ are obtained by averaging the sample variances of all the neighboring genes contained in a window of size w. The regularized t-statistic is given by (16) using $\hat{\sigma}_i^2$ of (17).

With variances estimated using the mixture model, we can replace $\tilde{\sigma}_{0,c}^2$ and $\tilde{\sigma}_{0,t}^2$ in (17) with $\hat{\sigma}_{i,c}^2$ and $\hat{\sigma}_{i,t}^2$ of (12) or (13), respectively, and obtain the corresponding test statistic from (16), which we will refer to as the regularized t^m-test statistic. Compared with the original regularized t-test of [8], our regularized t^m-test only changes prior variances. Therefore, similar to [8], we will use a student distribution to calculate the p-value. Note

[1]We use t^m to emphasize that this is the t-statistic with our model-based variance estimation; we will use a similar notation for the regularized t-statistic.

that we do not consider model 3 for the regularized t^m-test, because the sample variances $s^2_{i,c}$ and $s^2_{i,t}$ are used in calculating $\hat{\sigma}^2_i$, which assumes that $\{x_{ij}\}_{j=1}^{m_c}$, as well as $\{x_{ij}\}_{j=m_c+1}^{m}$, have the same variance. Essentially, our regularized t^m-test uses a more sophisticated method to estimate the prior variances than the regularized t test of [8]. The regularized t-test chooses a window based on the mean expression levels, arbitrarily sets the window size, and does not specify the weights in averaging the sample variances of the genes in the window. In contrast, we estimate the prior variance based on cluster analysis. Our window is gene specific, and weights are chosen according to $\hat{p}_{ik}$, the estimated probability that gene i belongs to cluster k.

D. S-Statistic

The SAM of [6] and the EB method of [7] use the following statistic:

$$S_i = \frac{\bar{x}_{i,t} - \bar{x}_{i,c}}{a_{0,i} + s_i} \tag{18}$$

where s_i is the estimated standard deviation of $\bar{x}_{i,t} - \bar{x}_{i,c}$ obtained from the sample variances $s^2_{i,c}$ and $s^2_{i,t}$. The parameter $a_{0,i}$ is chosen as the ninetieth percentile of $\{s_i\}_{i=1}^{n}$ in [7], while it is calculated according to the coefficient of variation of S_i in [6]. For simplicity, we will choose $a_{0,i}$ as [7] and call S_i in (18) the S-statistic. With variances estimated from cluster analysis, we can replace s_i in (18) with $\hat{\sigma}_i$ of (15), calculate $a_{0,i}$ from $\{\hat{\sigma}_i\}_{i=1}^{n}$, and denote this test statistics as S^m-statistic. Using the S-statistic in (18), the EB approach calculates the posterior probability that a gene is differentially expressed, while the SAM identifies significant genes by analyzing the false discovery rate (FDR). For both S- and S^m-statistics, we here will simply use balanced permutation to form null scores in the same way as (18), estimate the null distribution from these null scores [7], and then calculate the p-value for each gene.

III. Results

A. Simulated Data

We used simulated data to compare the performance of different tests discussed in Section II. In our simulations, we chose the number of genes $n = 1000$ and considered the following cases for the number of replicates: 1) $m_c = m_t = 4$ and 2) $m_c = m_t = 8$. The number of differentially expressed genes was chosen to be $n_d = 50$, and we divided 950 nondifferentially expressed genes into ten clusters, each with 95 genes. For $n_d = 50$ differentially expressed genes, we considered the following two cases: 1) they are in a single cluster, and 2) they are divided into three clusters, with cluster sizes equal to 10, 15, and 25, respectively. The variance of each gene in different arrays are equal, and as in [9] and [26], the variance λ_k for the genes in cluster k was generated using an inverse Gamma distribution, i.e., the pdf of $x = \lambda_k^{-1}$ is given by

$$g(x; a, b) = \frac{x^{a-1}\exp(-x/b)}{\Gamma(a) b^a}, \quad x > 0. \tag{19}$$

We chose $a = 2.4$ and $b = 1.4$, because these values fit certain real microarray data, as shown in [26]. The mean of cluster k, μ_k under the control condition was generated from a normal random variable with zero mean and unit variance. For the nondifferentially expressed genes, the mean under the treatment condition is equal to the mean under the control condition. For the differentially expressed genes in cluster k, the mean under the treatment condition is equal to $\mu_k + \nu_k$, where ν_k was generated from a random variable uniformly distributed over the intervals [0.5, 3] and $[-3, -0.5]$. The microarray data $\{x_{ij}\}$ were then generated from i.i.d. normal random variables with variances and means generated using the aforementioned method.

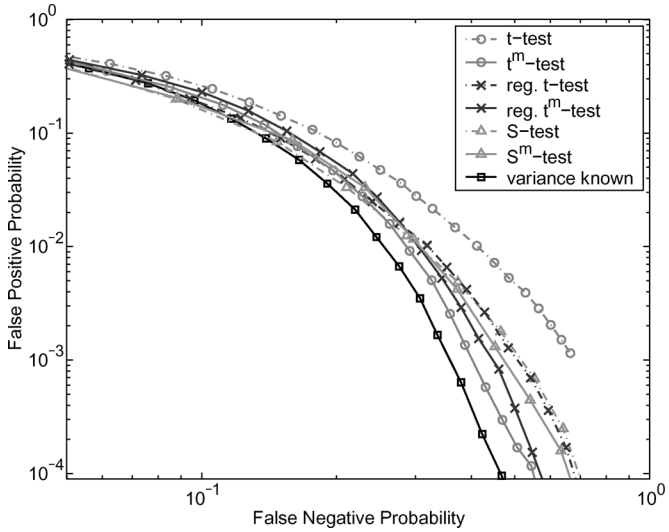

Fig. 1. Performance comparison ($m_c = m_t = 4$, $n_d = 50$); differentially expressed genes are in a single cluster.

We generated 200 data sets for each case. For each gene in these data sets, we calculated the statistics of t- and t^m-, regularized t- and t^m-, and S- and S^m-tests. Cluster analysis was carried out using a program modified from that of [27]. For the regularized t- and t^m-tests, we chose $N = 10$, and we used a window size of $w = 101$ for the regularized t-test. Note that $N = 10$ and $w = 101$ are the default values used in Cyber-T of [8]. Since we know which gene is differentially expressed and which is not, we can compare the test statistic with a cutoff value and then find the number of false positive or negative errors for this particular cutoff value. If $\mathbf{c} = [c_1, \cdots, c_M]$ denotes the vector containing M cutoff values, we can obtain the vectors $\mathbf{n}^+ = [n_1^+, \cdots, n_M^+]$ and $\mathbf{n}^- = [n_1^-, \cdots, n_M^-]$ containing the numbers of false positive and negative genes, respectively, corresponding to $\mathbf{c}$ for all 200 data sets; and then we obtain the false positive probability (FPP) and false negative probability (FNP) by dividing $\mathbf{n}^+$ and $\mathbf{n}^-$ by nn_d and $n(n - n_d)$, respectively. This allows us to compare the FPP and FNP of different tests without calculating p-values. In our simulations, we effectively use independent microarray data of $950 \times 200 = 1.9 \times 10^5$ genes to estimate FPP and use independent data of $50 \times 200 = 1000$ genes to estimate FNP. These numbers are sufficiently large to obtain a quite accurate estimate of FPP and FNP, for an FPP greater than 10^{-4} and an FNP greater than 10^{-2}.

We plotted FPP versus FNP of various tests in Figs. 1–4 for the four cases we described previously. If the variance of each gene is perfectly known, we can obtain a statistic by replacing the denominator of (1) with the true variance of $\bar{x}_{i,t} - \bar{x}_{i,c}$. The

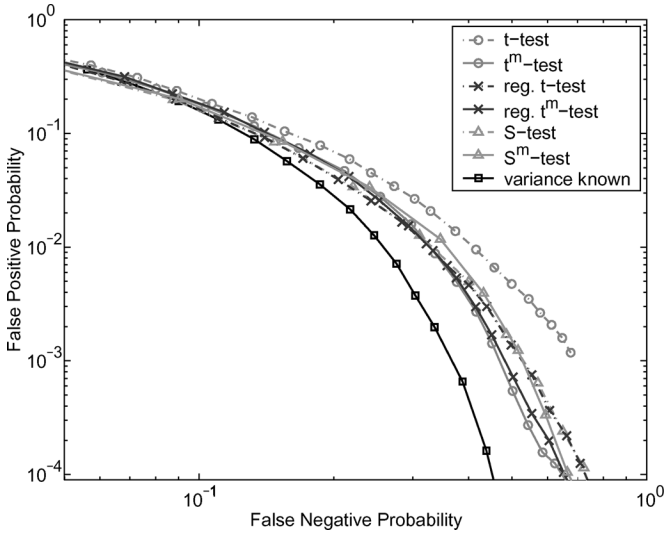

Fig. 2. Performance comparison ($m_c = m_t = 4$, $n_d = 50$); differentially expressed genes are divided into three clusters with sizes 10, 15, and 25, respectively.

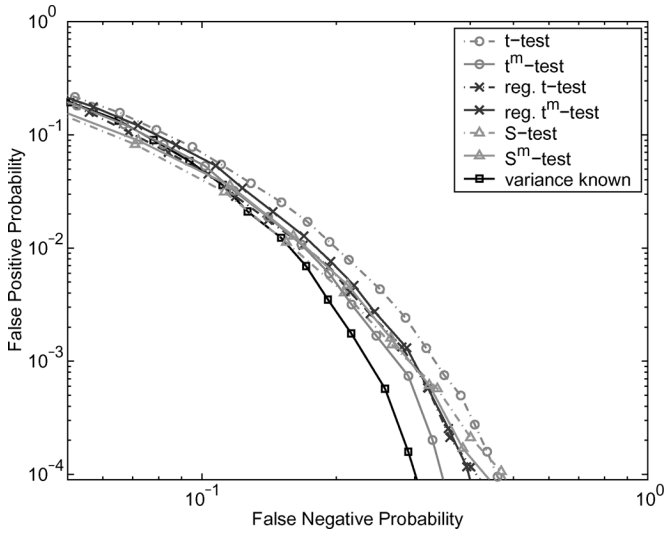

Fig. 3. Performance comparison ($m_c = m_t = 8$, $n_d = 50$); differentially expressed genes are in a single cluster.

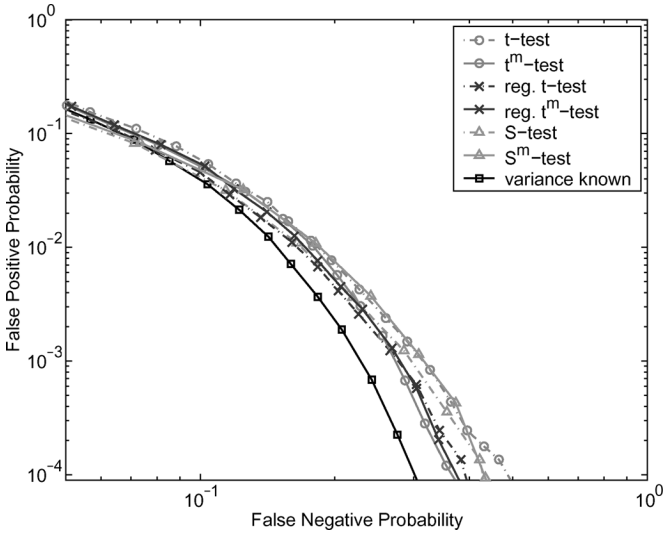

Fig. 4. Performance comparison ($m_c = m_t = 8$, $n_d = 50$); differentially expressed genes are divided into three clusters with sizes 10, 15, and 25, respectively.

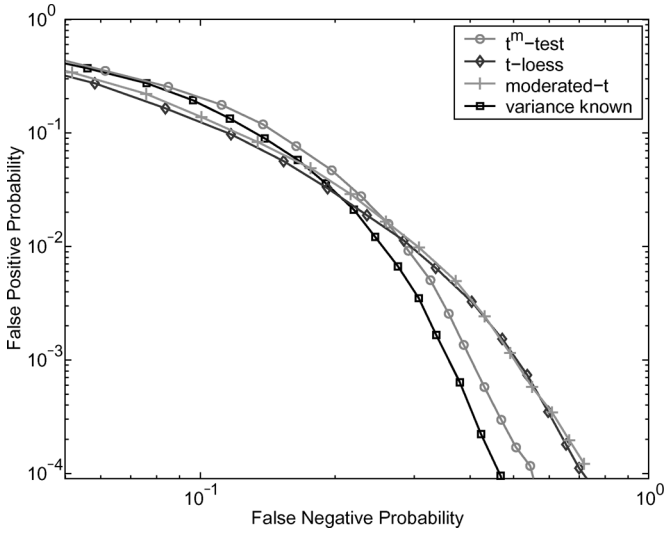

Fig. 5. Comparison of gene pooling methods ($m_c = m_t = 4$, $n_d = 50$); differentially expressed genes are in a single cluster; nondifferentially expressed genes are in ten clusters.

FPP versus FNP of such an ideal test is also plotted in these figures, which can serve as a lower bound for the performance of all other tests. Comparing the curves of t- and t^m-tests, we see that with our novel variance estimation based on cluster analysis, the FPP is lower for a fixed FNP, or equivalently, the FNP is lower for a fixed FPP. Hence, our model-based variance estimation improves the performance of the t-test compared with the variance estimation based on sample variances used in the original t-test. Similarly, our regularized t^m- or S^m-tests offer better performance than that of the original regularized t- or S-tests. Since the regularized t- and S-tests have already used some global information to improve performance, performance improvement of our regularized t^m- and S^m-tests, relative to the regularized t- and S-tests, is relatively small. Comparing Figs. 1 and 2 with Figs. 3 and 4, we see that for a fixed FPP, increasing the number of replicates reduces the FNP, or equivalently, improves the detection power, as expected.

We also compared the performance of our method with that of other two gene pooling methods: the moderated t-test in [12] and the smoothed sample variance using a loess smoother in [19]. We used the Limma program of [12] to calculate the moderated t-statistic and obtain FPP and FNP. In smoothing the sample variance, we used a window size of 100 samples, as recommended in [19]. Figs. 5 and 6 plot FPP versus FNP when $m = 4$; nondifferentially expressed genes are in ten clusters, and differentially expressed genes are in one and three clusters, respectively. These are the same settings used in Figs. 1 and 2, respectively. It is seen that our method outperforms the other two gene pooling methods, when FPP < 0.01 in Fig. 5 and when FPP < 0.002 in Fig. 6. Considering multiple test adjustment, a p-value less than 0.002 will be used in practice. Hence, our method outperforms alternative gene pooling methods in the practical region of FPP, in these two cases. We also ran simulations using the following setting: both differentially and

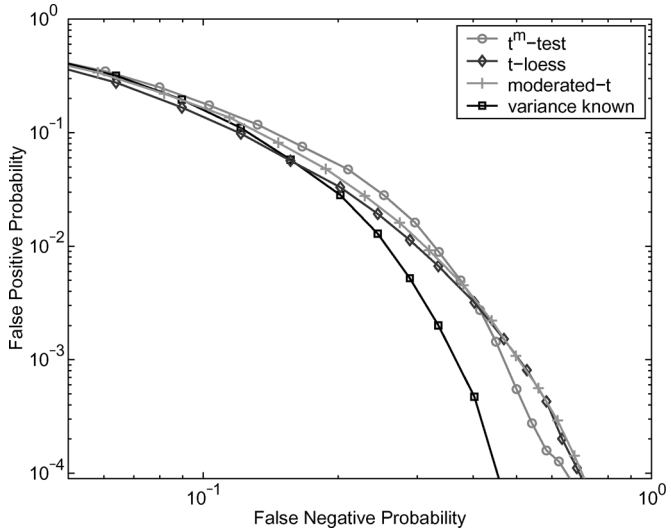

Fig. 6. Comparison of gene pooling methods ($m_c = m_t = 4$, $n_d = 50$); differentially expressed genes are divided into three clusters with sizes 10, 15, and 25, respectively; nondifferentially expressed genes are in ten clusters.

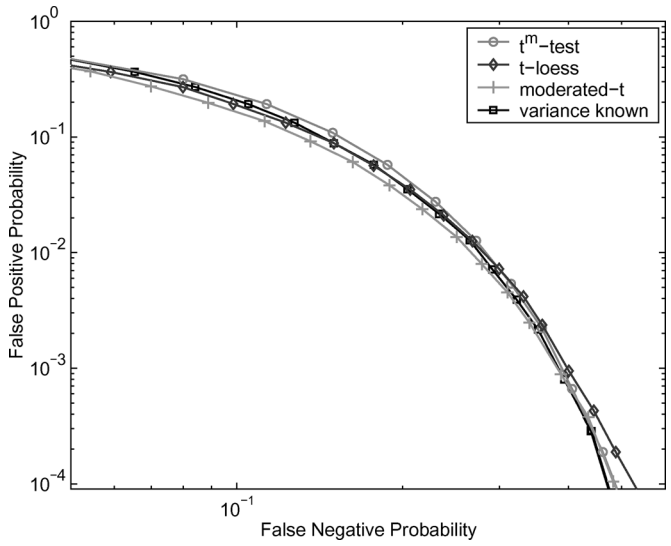

Fig. 7. Comparison of gene pooling methods ($m_c = m_t = 4$, $n_d = 50$); both differentially and nondifferentially expressed genes are in a single cluster.

nondifferentially expressed genes are in a single cluster. We plot the FPP versus FNP in Fig. 7 for this setting. Since data are in only two clusters, moderated t-test and loess variance smoothing method can obtain very good variance estimates. Since our cluster analysis identifies these two clusters correctly, our model-based variance estimation method also yields very good variance estimates. Therefore, as seen from Fig. 7, three methods all have similar performance close to the lower bound when the variance is known.

B. Colon Data

We used colon data of [28] and a cross-validation method similar to that of [29] to compare the performance of different tests. The colon data set is from Affymetrix oligonucleotide microarray. It contains 62 samples of 6500 genes: 40 tumor and 22 normal colon tissue samples. We normalized each array by subtracting the mean intensity and used the first 1000 genes with the highest minimal intensity. We first used the t-test, which uses all samples to rank genes. Since the number of samples is relatively large, this ranking is reasonably accurate. We then randomly take four samples from each of the tumor and normal samples. Based on these eight samples, we used eight different tests to rank genes. These eight tests are t-test, S-test, regularized t-test, t^m-test, S^m-test, regularized t^m-test, moderated t-test of [12], and the t-test that uses loess smoothed variance of [19]. Comparing the gene ranking of each test with the ranking yielded by the t-test using all samples, we obtain the number of common genes in the top 20, 100, and 200 genes. We ran this procedure 200 times and obtained the average and the standard deviation of the number of common genes in top genes, which are listed in Table I. If a test has a larger number of common genes with a smaller standard deviation, it offers better performance. From Table I, it is seen that the t-test has the worst performance as expected. The S-test, regularized t-test, t^m-test, S^m-test, and regularized t^m-test offer comparable performance. Compared with the S-test and regularized t-test, our model-based variance estimation does not considerably improve performance in this case. This is probably because there is no well-separated clusters in this colon data set. Compared with our method, the moderated t-test and the t-test with loess smoothed variance offer slightly worse performance when the top 100 or 200 genes are considered and comparable performance when top 20 genes are considered.

C. IHF Data

We applied the t- and t^m- tests, as well as the regularized t- and t^m-tests, to IHF data, which measure the gene expression profiles in integration host factor IHF$^+$ and IHF$^-$ strains of *Escherichia coli* [30]. The IHF data set contains $m = 8$ replicates for each of 4290 genes: of these eight, replicates 1–4 are from IHF$^+$ cells, and replicates 5–8 are from IHF$^-$ cells. As in [31], we only used 1973 genes whose all eight replicates of gene expression are above the background.

We first performed cluster analysis on this data set with the three models described in Section II. Based on the BIC, we found that the best model and the cluster number are model 1 and $K = 35$, respectively. This implies that expression levels of a gene in 8 arrays have the same variance. The quantity n_k defined in Section II can be viewed as the size of cluster k. From our cluster analysis, we found that cluster 32 has the minimum cluster size $n_{32} = 2$, and gene 1560 (rplY) and 1677 (b1031) belong to this cluster with a probability greater than 0.999. The sample variances for these two genes are $s^2_{1560,c} = 1.12\times10^{-6}$, $s^2_{1560,t} = 2.22 \times 10^{-6}$, $s^2_{1677,c} = 1.58 \times 10^{-6}$, and $s^2_{1677,t} = 2.05 \times 10^{-6}$. Indeed, these two genes have similar variances, and we used 16 replicates of these two genes to estimate the variance, which effectively doubles the number of replicates in variance estimation. The second and third smallest clusters have cluster sizes of 3 and 8, respectively. Hence, for most of genes, our model-based variance estimation considerably increases the number of replicates.

For the t-statistic, we used student's distribution with six degrees of freedom to calculate the p-value as [30]. For the regularized t- and t^m-tests, we chose $v_{0,c} = v_{0,t} = 6$ and used student's distribution with 16 degrees of freedom to

TABLE I
NUMBER OF COMMON GENES IN TOP 20, 100, AND 200 GENES

	t-test	S-test	Reg. t	t^m-test	S^m-test	Reg. t^m	Loess var.	Moderated t
Top 20	1.2 ± 1.3	2.3 ± 1.8	1.9 ± 1.5	2.0 ± 1.6	2.1 ± 1.6	2.1 ± 1.8	2.2 ± 1.7	2.22 ± 1.9
Top 100	17.0 ± 8.1	26.7 ± 9.1	25.9 ± 7.9	27.9 ± 10.3	27.8 ± 9.0	27.9 ± 10.8	25.9 ± 10.0	23.6 ± 10.5
Top 200	54.7 ± 18.3	73.5 ± 18.9	72.7 ± 16.8	73.6 ± 20.2	74.6 ± 18.9	73.0 ± 21.0	70.1 ± 20.4	68.5 ± 21.2

TABLE II
RANKING OF FIRST TEN GENES WITH ABSOLUTE FOLD CHANGE VALUES LESS THAN 2

Gene	Fold	t-test		t^m-test		Reg. t-test		Reg. t^m-test	
		Statistic	Rank	Statistic	Rank	Statistic	Rank	Statistic	Rank
b2226	1.67	13.50	4	2.58	158	1.91	202	2.95	95
ilvA	-1.48	-11.20	10	-2.52	171	-2.11	165	-2.87	100
rfaD	-1.95	-10.31	14	-3.80	49	-2.19	148	-4.25	38
ylem	1.75	8.87	23	2.21	282	2.95	95	2.40	173
b2295	1.76	8.01	27	2.54	165	2.03	183	2.86	103
b2255	-1.71	-7.83	29	-3.09	91	-2.65	90	-3.44	63
b1725	1.72	7.50	34	1.66	483	1.42	399	1.89	297
yedE	1.33	6.67	44	1.47	591	1.11	594	1.68	404
ilvE	-1.62	-5.99	54	-3.41	69	-2.06	170	-3.66	57
leuA	1.53	5.65	59	2.16	263	1.73	240	2.41	170

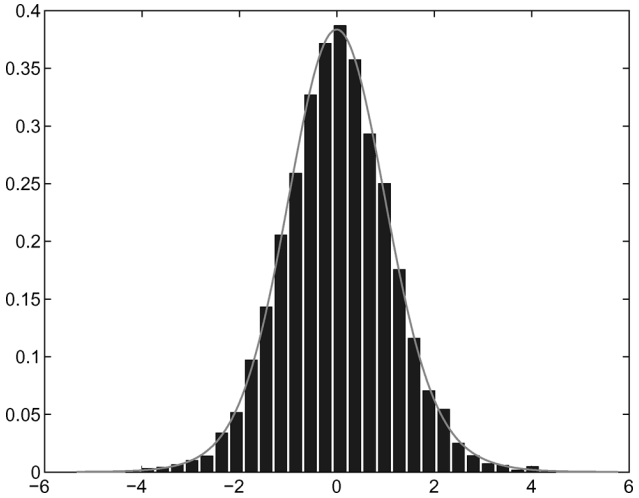

Fig. 8. Histogram of the null scores of the t^m-test and the null distribution estimated from the null scores with a normal mixture model.

calculate the p-value as [8] and [31]. For the t^m-statistic, we used the balanced permutation of [7] to form null scores and fit these null scores to a normal mixture model to estimate the null distribution [25]. For model fitting, we used the program EMMIX [32], which is available at http://www.maths.uq.oz.au/~gjm/emmix/emmix.html. Since the number of null scores produced by all 36 balanced permutations is too large for the EMMIX, we randomly chose four balanced permutations, which yields 7892 null scores. If y denotes the null score, the model fitting yields the following pdf for y

$$f(y) = 0.721 f_1(y| -0.0058, 0.8854) + 0.279 f_2(y|0.0792, 2.0458) \quad (20)$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the normal pdf's defined in (3). The histogram of the null scores and the estimated null distribution are depicted in Fig. 8; it is seen that they match each other very well.

For a p-value less than 0.0001 (not corrected for multiple tests), the t-test identifies 20 genes, while the t^m-test identifies 19 genes. Both the regularized t- and t^m-tests identify 16 genes. Among the 20 genes with the smallest p-values, the t-test has six, seven, and seven genes common to the t^m-test, and the regularized t- and t^m-tests, respectively; on the other hand, the t^m-test has 12 and 16 genes common to the regularized t- and t^m-tests, respectively. For a p-value less than 0.01, the t-test identifies 127 genes, while the t^m-test identifies 93 genes. The regularized t- and t^m-tests detect[2] 74 and 96 genes, respectively. Among the 127 genes with the smallest p-values, the t-test has 82, 85, and 92 genes common to the t^m-test, and the regularized t- and t^m-tests, respectively; the t^m-test has 98 and 116 genes common to the regularized t- and t^m-tests. It is observed that the t^m-test is relatively consistent with the regularized t- and t^m-tests, while the genes identified by the t-test have relatively large discrepancy with those identified by the other three tests.

As we mentioned in the Section I, due to the small sample size in variance estimation, the estimated variance may be too small to be true, which causes a very large absolute value for the t-statistic even when the fold change is small; this, in turn, increases FPP. To see the effect different tests have on the genes with small fold change, we list, in Table II, the values of test statistics and the rank of p-values for the first ten genes whose absolute fold change is less than 2. Note that the smaller the rank is, the smaller the p-value is, and the more significant the gene is. While the t-test gives these ten genes high significance, the other three tests reduce the significance of these genes. For example, gene b2226 in the t-test is ranked as the fourth significant gene, while the other three tests give it much higher ranking. If we check the raw data of this gene, the sample variances under two conditions are much smaller than other genes with similar mean expression levels, which implies that the sample variance may not be a reliable estimate of the true variance; and thus, the rank produced by the t-test for this particular gene may not be reliable either. Although the t^m-test and the regularized t- and

[2]This number is different from the 86 genes reported in [31], which may be due to the difference in implementing estimation of the prior variance.

t^m-tests are relatively consistent in identifying genes with top significance as we discussed before, the ranks for these ten genes given by these three tests seem not to be very much consistent as shown in Table II. This suggests that statistical inference for these genes may exhibit large uncertainty, and further biological analysis or experiments are needed to obtain more convincing results.

IV. Discussion

In microarray experiments, the number of replicates is often small due to high cost and intensive labor required. When one employs statistical tests to identify differentially expressed genes under different conditions from microarray data, the small sample size causes two major problems: 1) low signal-to-noise ratio (SNR) and 2) large errors in estimated variances used in statistical tests. Both problems increase the probability of false discovery and decrease the detection power. The SNR can only be increased by increasing the number of replicates, since the gene expression levels are determined by biological processes inherent to the cell. Our focus in this paper was to exploit relevant information contained in the microarray data to improve the estimation of variance.

While clustering algorithms have been applied to identify co-expressed genes [2], we used model-based cluster analysis to identify genes with similar variance to improve variance estimation. Unlike hierarchical clustering [18], where the correlation in gene expression is used as a metric to cluster genes, the clustering analysis based on a Normal mixture exploits both mean expression level and variance, which renders it particularly suitable for variance estimation. The question one may raise is whether the normal mixture model holds true for microarray data. First, it was shown in [15] that some raw or transformed microarray data are nearly normal. Second, there is no need to perform hard clustering in our variance estimation, i.e., it is not necessary to assign a gene to a particular cluster. Instead, we only need the probability that a gene belongs to a cluster; as a result, we do not require the microarray data to obey a normal distribution. Depending on the model parameters, the normal mixture models we used here are suitable for a wide range of probability distributions. We assumed that different genes are independent, which makes the problem mathematically tractable. If some genes are correlated, one may use a normal mixture model that takes into account this correlation—an approach may worth further investigation.

The variance estimated from our model-based method can be used in any statistical tests requiring variance information. We applied our variance estimation to t- and regularized t-tests, as well as the S-statistic used in the empirical Bayesian approach of [7]. Using simulated data, we illustrated that with our model-based variance estimation, the detection power of these tests is increased for a fixed FPP, or equivalently, the FPP is decreased for a fixed detection power. Using the colon microarray data of [28] and a cross-validation method, we demonstrated that our model-based variance estimation method improves the performance of t-test and offers slightly better performance than other two gene pooling methods: the moderated t-test of [12] and the t-test with smoothed sample variance of [19]. Compared with the S-test and regularized t-test, our method offers comparable performance in this case, probably because there is no well-separated clusters in this data set. When applying these tests to the IHF microarray data of [30], the t- and regularized t-tests with our model-based variance estimation and the regularized t-test of [8] reduce the significance of those genes that have a low fold change but are assigned a high significant score by the t-test with sample variance, which potentially reduces the FPP.

References

[1] G. A. Churchill, "Fundamentals of experimenal design for cDNA microarrays," *Nat. Genet.*, vol. 32, pp. 490–495, Dec. 2002.

[2] J. Quackenbush, "Computational analysis of microarray data," *Nat. Rev. Genet.*, vol. 2, pp. 418–427, Jun. 2001.

[3] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc. Natl. Acad. Sci. USA*, vol. 93, pp. 10 614–10 619, Oct. 1996.

[4] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol.*, vol. 4, no. 4, pp. 210.1–210.10, Mar. 2003.

[5] M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin, "Microarray expression profiling identifies genes with altered expression in HDL-defeficient mice," *Genome Res.*, no. 10, pp. 2022–2029, 2000.

[6] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 9, pp. 5116–5121, Apr. 2001.

[7] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayers analysis of a microarray experiment," *J. Amer. Stat. Assoc.*, vol. 96, no. 456, pp. 1151–1160, Dec. 2001.

[8] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[9] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, pp. 31–46, 2002.

[10] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui, "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data," *J. Comput. Biol.*, vol. 8, no. 1, pp. 37–52, 2001.

[11] X. Cui, J. T. Hwang, J. Qiu, N. J. Blades, and G. A. Chruchill, "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, vol. 6, no. 1, pp. 59–75, Jan. 2005.

[12] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genetics Mol. Biol.*, vol. 3, no. 1, Article 3A, Feb. 2004.

[13] G. J. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[14] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *Comput. J.*, vol. 41, no. 8, pp. 578–588, 1998.

[15] K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo, "Model-based clustering and data transformations for gene expression data," *Bioinformatics*, vol. 17, no. 10, pp. 977–987, 2001.

[16] D. Ghosh and A. M. Chinnaiyan, "Mixture modeling of gene expression data from microarray experiments," *Bioinformatics*, vol. 18, no. 2, pp. 275–286, 2002.

[17] G. J. McLachlan, R. W. Bean, and D. Peel, "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, vol. 18, no. 3, pp. 413–422, 2002.

[18] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 14 863–14 868, Dec. 1998.

[19] X. Huang and W. Pan, "Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays," *Funct. Integr. Genomics*, vol. 2, pp. 126–133, 2002.

[20] N. Jain, J. Thatte, T. Braciale, K. Ley, M. O'Connell, and J. K. Lee, "Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays," *Bioinformatics*, vol. 15, no. 15, pp. 1945–1951, 2003.

[21] C. Kooperberg, A. Aragaki, A. D. Strand, and J. M. Olson, "Significance testing for small microarray experiments," *Stat. Med.*, vol. 24, no. 15, pp. 2281–2298, May 2005.
[22] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistic Sinica*, no. 12, pp. 111–139, 2002.
[23] A. P. Dempster, N. M. Larid, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, vol. 39, pp. 1–38, 1977.
[24] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern Recognit.*, vol. 28, no. 5, pp. 781–793, 1995.
[25] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.
[26] G. W. Wright and R. M. Simon, "A random variance model for detection of differential gene expression in small microarray experiments," *Bioinformatics*, vol. 19, no. 18, pp. 2448–2455, 2003.
[27] A. R. Martinez and W. L. Martinez. (2004) Model-based clustering toolbox for MATLAB. [Online]. Available: http://www.stat.washington.edu/fraley/mclust/soft.shtml
[28] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, pp. 6745–6750, Jun. 1999.
[29] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, "Is cross-validation better than resubstitution for ranking genes?," *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.
[30] S. M. Arfin, A. D. Long, E. T. Ito, L. Tolleri, M. M. Riehle, E. S. Paegle, and G. W. Hatfield, "Global gene expression profiling in Escherichia coli K12," *J. Biol. Chem.*, vol. 275, no. 38, pp. 19672–29 684, Sep. 2000.
[31] A. D. Long, H. J. Mangalam, B. Y. P. Chan, L. Tolleri, G. W. Hatfield, and P. Baldi, "Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework," *J. Biol. Chem.*, vol. 276, no. 23, pp. 19 937–19 944, June 2001.
[32] G. L. McLachlan, D. Peel, K. E. Basford, and P. Adams, "Fitting of mixtures of normal and t-components," *J. Stat. Softw.*, vol. 4, 1999.

Xiaodong Cai (S'00–M'01–SM'05) received the B.S. degree from Zhejiang University, China, the M.Eng. degree from the National University of Singapore, Singapore, and the Ph.D. degree from the New Jersey Institute of Technology, Newark, in 2001, all in electrical engineering.

From February 2001 to June 2001, he was a Member of Technical Staff at Lucent Technologies, NJ. From July 2001 to October 2001, he was a Senior System Engineer at Sony technology center, San Diego, CA. From November 2001 to July 2004, he was a Postdoctoral Research Associate in the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis. Since August 2004, he has been an Assistant Professor in the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL. His main research interests lie in the areas of bioinformatics, computational system biology, statistical signal processing, communications, and networking.

Georgios B. Giannakis (F'97) received the Diploma degree in electrical engineering from the National Technical University of Athens, Greece, 1981 and the MSc. degree in electrical engineering, the MSc. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 1983, 1986, and 1986, respectively.

After lecturing for one year at USC, he joined the University of Virginia, Charlottesville, in 1987, where he became a Professor of electrical engineering in 1997. Since 1999, he has been a Professor with the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis, where he now holds an ADC Chair in Wireless Telecommunications. His general interests span the areas of communications and signal processing, estimation and detection theory, time-series analysis, and system identification—subjects on which he has published more than 250 journal papers, 400 conference papers, and two edited books. Current research focuses on diversity techniques for fading channels, complex-field and space–time coding, multicarrier, ultra-wideband wireless communication systems, cross-layer designs, and sensor networks.

Dr. Giannakis is the (co)recipient of six paper awards from the IEEE Signal Processing (SP) and Communications Societies in 1992, 1998, 2000, 2001, 2003, and 2004. He also received Technical Achievement Awards from the IEEE SP Society in 2000 and from EURASIP in 2005. He served as Editor-in-Chief for the IEEE SIGNAL PROCESSING LETTERS, as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE SIGNAL PROCESSING LETTERS, as Secretary of the SP Conference Board, as member of the SP Publications Board, as member and Vice-Chair of the Statistical Signal and Array Processing Technical Committee, as Chair of SP for the Communications Technical Committee, and as a member of the IEEE Fellows Election Committee. He has also served as a member of the IEEE SP Society's Board of Governors, the Editorial Board for the PROCEEDINGS OF THE IEEE, and the steering committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.