9-1-2012

Analyzing Multiple-Probe Microarray: Estimation and Application of Gene Expression Indexes

Mehdi Maadooliat
Marquette University, mehdi.maadooliat@marquette.edu

Jianhua Z. Huang
Texas A & M University - College Station

Jianhua Hu
University of Texas M.D. Anderson Cancer Center

Accepted version. *Biometrics,* Vol. 68, No. 3 (September 2012): 784-792. DOI.
Mehdi Maadooliat was affiliated with Texas A&M University at the time of publication.

Biometrics 000, 1–18 DOI: 000
000 2010

Analyzing multiple-probe microarray: estimation and application of gene expression indexes

Mehdi Maadooliat[1]

Jianhua Z. Huang[2]

Jianhua Hu[3,*]

[1]Graduate student, Department of Statistics, Texas A&M University, College Station, TX, USA,

[2]Professor, Department of Statistics, Texas A&M University, College Station, TX, USA,

[3]Associate Professor, Department of Biostatistics, Division of Quantitative Sciences,

University of Texas M.D. Anderson Cancer Center, Houston, TX, USA

email: jhu@mdanderson.org

Summary: Gene expression index estimation is an essential step in analyzing multiple probe microarray data. Various modeling methods have been proposed in this area. Amidst all, a popular method proposed in Li and Wong (2001) is based on a multiplicative model, which is similar to the additive model discussed in Irizarry et al. (2003a) at the logarithm scale. Along this line, Hu et al. (2006) proposed data transformation to improve expression index estimation based on an ad hoc entropy criteria and naive grid search approach. In this work, we re-examined this problem using a new profile likelihood-based transformation estimation approach that is more statistically elegant and computationally efficient. We demonstrate the applicability of the proposed method using a benchmark Affymetrix U95A spiked-in experiment. Moreover, We introduced a new multivariate expression index and used the empirical study to shows its promise in terms of improving model fitting and power of detecting differential expression over the commonly used univariate expression index. As the other important content of the work, we discussed two generally encountered practical issues in application of gene expression index: normalization and summary statistic used for detecting differential expression. Our empirical study shows somewhat different findings from the MAQC project (MAQC, 2006).

Key words: Differential expression detection; Fold change; Multivariate expression index; Normalization; Profile likelihood; Transformation model.

1. Introduction

Microarray technologies have been used to monitor expression intensities of thousands of genes simultaneously in a wide range of organisms. We focus on the popular Affymetrix short oligonucleotide array platform (Lockhart et al., 1996; Parmigiani et al., 2003). In this system, each gene is represented by multiple oligonucleotide probes, or a "probe set." A probe set contains 10-20 probe pairs whose expression intensities are measured via hybridization to the targeted sample cRNA. Each probe pair consists of the perfect match (PM) probe with the target mRNA sequence of 25 nucleotides and the counterpart mismatch (MM) probe that is identical to the PM probe except for a base change at the middle (13th) position.

The first part of this work concerns an important statistical problem in analyzing affymetrix array data, that is, estimation of gene expression based on the multiple-probe information. A number of statistical methods have been proposed to address this problem. GeneChip software (MAS version 5.0; Affymetrix 2004) computes a robust form of mean differences between PM and MM probes. Chu et al. (2002) applied mixed linear models to account for the dependence among logarithm-transformed probe intensities. Li and Wong (2001) proposed a multiplicative model-based expression index, which has been used and implemented in the dChip software (www.dchip.org). Its superiority over several existing methods has been shown in Lemon et al. (2002) via both analytic arguments and empirical data. The so-called Li-Wong Reduced (LWR) model was proposed based on the differences between PM and MM intensities. Since the MM probes are designed originally for measuring the background/nonspecific intensities, the differences are considered to be the signal intensities in LWR. The model is expressed as

$$PM_{ij} - MM_{ij} = \theta_i \phi_j + \epsilon_{ij} \tag{1}$$

where PM_{ij} and MM_{ij} are the PM and MM intensity values for the i^{th} $(i = 1, \cdots, I)$ array and the j^{th} $(j = 1, \cdots, J)$ probe pair for the gene, θ_i is the true expression index, ϕ_j

is the rate of response in the corresponding PM probe, and the residuals $\epsilon_{ij} \sim N(0, \sigma^2)$. The model identifiability is ensured by posing the constraint $\sum_j \phi_j^2 = J$. Later work showed that the underlying distribution assumptions of normality and constant variance across the probes do not hold and data transformation techniques can be used to resolve this problem (Geller et al., 2003; Hu et al., 2006).

Hu et al. (2006) established the connection between the Li-Wong model and the first characteristic mode of the singular value decomposition (SVD) of the probe intensity matrix, and proposed a parametric transformation model on PM intensities. More specifically, they proposed a grid search over a parametric transformation family (e.g., Box-Cox) and selected the optimal value of the transformation parameters by maximizing the normalized discrete Shannon entropy defined on the singular values of the residual matrix. Note that, the empirical results provided in their paper show a good level of improvement in homogeneity of variance of the residuals and efficiency of the expression index. Moreover, the transformation model does not require knowledge of the experimental design.

We propose a more statistically principled estimation method than the entropy-based procedure in Hu et al. (2006). The transformation model can be written as

$$f(y_{ij}|\boldsymbol{\eta}) = \theta_i \phi_j + \epsilon_{ij}, \tag{2}$$

where $f(\cdot|\boldsymbol{\eta})$ is a monotonic transformation, $\boldsymbol{\eta}$ is the vector of transformation parameters, and ϵ_{ij}'s are independent normal random errors with mean 0 and constant variance σ^2. The goal of this transformation model is to stabilize the variance of the residuals and thus achieve an efficient estimates for the gene expression index under normality assumption. This is similar in spirit to performing logarithm transformation often seen in analyzing microarray experiments, as discussed in Geller et al. (2003); Hu et al. (2006) and Irizarry et al. (2003b). We consider a wider class of transformations which contains the logarithmic transformation and LWR model (identity transformation) as a special case. We expect to achieve a better

performance based on the general model, since our empirical investigation shows that the logarithm transformation is not always optimal. The proposed likelihood based method and the entropy based procedure in Hu et al. (2006) are closely related since the normal distribution has the maximum entropy property. However, our simulation study as reported in Web Appendix A shows that the proposed method exhibits smaller variability of parameter estimation than the ad hoc entropy procedure. Application to a benchmark Affymetrix U95A spiked-in experiment (Irizarry et al., 2003a,b) also indicates that the proposed method has superior performance in terms of reflecting the true patterns in a controlled experiment, comparing to the Li-Wong model.

We also introduce a new multivariate gene expression index obtained through the connection between the multiplicative model and SVD. It is noted that all the existing methods concern only the univariate expression index. Through extensive real data exploration, we show the benefit of using the multivariate index in terms of model fitting and applications, such as differential expression detection.

The second part of the work is devoted to discussion of two practical issues in analysis of microarray data using estimated gene expression index, namely, normalization and summary statistic used for detecting differential expression. These issues are important but still under debate. MAQC (2006) used empirical studies over some of the known techniques, and concluded that normalization has little impact on the result of detecting differentially expressed genes, and p-value has no gain over fold-change in terms of gene ranking of differential expression. We use the benchmark spiked-in data to re-investigate these two issues for the expression indexes based on our proposed model. Our empirical study seems telling a different story from MAQC project.

We describe the likelihood based estimation procedure and introduce the two-dimensional expression index model in Section 2. We use the well known benchmark human spiked-in dataset to demonstrate the applicability of the proposed expression index estimation method

in Section 3, and explore the two practical issues of normalization and summary statistic for detecting differential expression in Section 4. Some final remarks are given in Section 5.

2. Profile likelihood based expression index estimation

2.1 *Estimation procedure*

We consider the transformation model (2) in which y_{ij} takes the value of preprocessed and normalized multiple-probe index. We denote the parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\eta}^T, \boldsymbol{\theta}^T, \boldsymbol{\phi}^T, \sigma^2)^T$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)^T$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)^T$, and write out the log-likelihood function as

$$\ell(\boldsymbol{\Theta}) = -\frac{IJ}{2}\log(\sigma^2) \quad + \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\Big[\log|f'(y_{ij}|\boldsymbol{\eta})| - \frac{\big\{f(y_{ij}|\boldsymbol{\eta}) - \theta_i\phi_j\big\}^2}{2\sigma^2}\Big] \tag{3}$$

It is noticeable that maximization of (3) with respect to $\boldsymbol{\Theta}$ simultaneously is computationally expensive. Hence, we turn to the profile likelihood method which comprises of two phases.

In the first phase, we fix the transformation parameter vector $\boldsymbol{\eta}$ at $\boldsymbol{\eta_0}$ and estimate only $\boldsymbol{\theta}, \boldsymbol{\phi}$, and σ^2. The maximum likelihood estimates (MLEs) of the parameters can be viewed as functions of $\boldsymbol{\eta_0}$, and can be easily obtained using the connection between SVD and the least square estimates (or equivalently MLEs with normal residuals) established in Hu et al. (2006). The explicit forms of the parameter estimates are

$$\begin{aligned} \widehat{\boldsymbol{\theta}(\boldsymbol{\eta_0})} &= \boldsymbol{u_1}\frac{\sigma_1}{\sqrt{J}}, & \widehat{\boldsymbol{\phi}(\boldsymbol{\eta_0})} &= \boldsymbol{v_1}\sqrt{J}, \\ \widehat{\sigma^2(\boldsymbol{\eta_0})} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\big\{f(y_{ij}|\boldsymbol{\eta_0}) - \widehat{\theta(\boldsymbol{\eta_0})_i}\widehat{\phi(\boldsymbol{\eta_0})_j}\big\}^2, \end{aligned} \tag{4}$$

where, $\boldsymbol{u_1}$ and $\boldsymbol{v_1}$ are the left and right singular vectors, respectively, corresponding to the largest singular value σ_1 by performing SVD on the matrix $\{f(y_{ij}|\boldsymbol{\eta_0})\}$. We do not need the full SVD decomposition, and efficient algorithm for low-rank matrix approximation can be used to speed up the calculation of the leading vectors (e.g. Achlioptas and Mcsherry, 2007).

In the second phase, we aim to obtain the estimate of $\boldsymbol{\eta}$ via maximizing the profile log-likelihood function

$$\ell_p(\boldsymbol{\eta}) = -\frac{IJ}{2}\log(\sigma^2(\boldsymbol{\eta})) \quad + \quad \sum_{i=1}^{I}\sum_{j=1}^{J}\Big[\log|f'(y_{ij}|\boldsymbol{\eta})| - \frac{\big\{f(y_{ij}|\boldsymbol{\eta}) - \theta(\boldsymbol{\eta})_i\phi(\boldsymbol{\eta})_j\big\}^2}{2\sigma^2(\boldsymbol{\eta})}\Big]. \quad (5)$$

Notice that the profile log-likelihood function only contains the vector of parameters $\boldsymbol{\eta}$ with all other parameters being expressed as functions of $\boldsymbol{\eta}$. The variety of appropriate optimization techniques such as Downhill Simplex or gradient based algorithms (e.g. Avriel, 1976) can be used subject to the structure of the family of transformations. It is worth emphasizing that the obtained estimates based on the profile likelihood are the MLEs of the parameters, and thus have good theoretical properties. We abbreviate the whole profiling and singular value decomposition procedure as PSVD.

So far our development is only for a single gene. However, we encounter thousands of genes in a real microarray experiment and aim at estimating a common transformation for all these genes. In this general case, the log-likelihood and the profile log-likelihood are respectively the summation of (3) and (5) over all involved genes. Our algorithm can be easily extended to calculate the maximum likelihood estimate. Despite of the large dimensionality, the SVD technique enables the computational feasibility of handling all the genes simultaneously in our procedure.

Given the general framework, we focus on the popular Box-Cox transformation family as an example hereafter. Let $\boldsymbol{Y}$ denote the untransformed data. The Box-Cox transformation for each element of $\boldsymbol{Y}$ is defined as

$$f(y_{ij}|\beta) = \begin{cases} \dfrac{y_{ij}^{\beta} - 1}{\beta}\ , & \beta \neq 0, \\ \log(y_{ij})\ , & \beta = 0. \end{cases} \quad (6)$$

Since the transformation parameter $\boldsymbol{\eta} = \beta$ is one dimensional, the 3$^{\text{rd}}$ step of the PSVD procedure becomes a simple optimization problem for a univariate concave function. In our implementation, we adopt the popular L-BFGS-B optimization algorithm (Byrd et al., 1994).

We conducted simulation studies to make comparisons between the PSVD method and

the entropy-based procedure in Hu et al. (2006). We studied both cases of normally and non-normally distributed data. The observation is that the PSVD method yields more efficient parameter estimate than the entropy-based procedure. The simulation study also demonstrates the advantage of data transformation in improving the model fit. We include the detailed description of the simulation studies in the Web Appendix A.

2.2 *Multivariate expression index*

It is aforementioned that the MLE of $\boldsymbol{\theta}, \boldsymbol{\phi}$ in model (2) are the singular vectors corresponding to the largest singular value of data matrix $\{f(y_{ij}|\boldsymbol{\eta})\}$. It is also known that the data matrix can be represented as the sum of rank-one matrices based on SVD. The most important rank-one matrix is associated with the largest singular value and it captures the highest "energy" in the data, where "energy" is defined by either the 2-norm or Frobenius norm(e.g. Trefethen and Bau, 1997) of the matrix. An intuitive question is whether this rank-one matrix is sufficient to capture the majority of energy in the data and it is interesting to investigate if the rank-one matrix corresponding to the second largest singular value contains nontrivial information about gene expression index.

One of the typical way of assessing the relative importance of the low-rank matrices generated by the SVD is to look at the ratio of the corresponding singular values. Following this practice, we include the rank-one matrix associated with the second largest singular value in expression index estimation only if $\dfrac{d_2}{d_3} > c$, where d_i denotes the i^{th} largest singular value of the transformed data $f(y_{ij}|\hat{\beta})$ in (2). In our empirical investigation, we take the relatively conservative threshold $c = 3$. This idea also has been used in Hu et al. (2009). The model can be explicitly written as

$$f(y_{ij}|\hat{\beta}) = \begin{cases} \theta_i\phi_j + \epsilon_{ij} \ , & \dfrac{d_2}{d_3} \leqslant c \\ \theta_i^{(1)}\phi_j^{(1)} + \theta_i^{(2)}\phi_j^{(2)} + \varpi_{ij} \ , & \dfrac{d_2}{d_3} > c \end{cases} \tag{7}$$

When a gene can be represented using one component, we take θ_i as its univariate expression

index for array i; Otherwise, we take the vector $(\theta_i^{(1)}, \theta_i^{(2)})$ as its two-dimensional expression index.

Note that more components can be considered in the similar way. Our empirical investigation suggests that the third and higher-order singular vectors mainly contain noises for short oligonucleotide array experiments and thus they will not be considered. More detailed description is deferred to section 3.2.

3. Application to a benchmark spiked-in data

We used a benchmark Affymetrix U95A spiked-in experiment as an example. It consists of $12,610$ genes non-differentially expressed across all the arrays, and 16 genes spiked-in at 14 known concentration levels ranging from 0 to $1,024$ picomolars, each of which has at least 3 replicate arrays. A Latin square design was used for the arrangement of 16 genes at different concentration levels in 59 arrays. The detailed description of the experimental design can be found in Irizarry et al. (2003a,b).

In real data analysis, we need to first perform data preprocessing procedure. Irizarry et al. (2003a,b) introduced quantile normalization and an alternative background intensity estimation method rather than directly using MM intensities as the Li-Wong model. Because of its good empirical performance, this preprocessing procedure has been widely used. Therefore, we adopted this approach and took the background adjusted and normalized PM probe intensities PM_{ij}^* as the data value prior to transformation in our development.

3.1 *Comparisons of expression index estimates*

We estimated the expression indexes of all the genes using the Li-Wong and PSVD methods, respectively. The PSVD method applied to all the genes simultaneously yields the estimate of the transformation parameter $\hat{\beta} = 0.177$. We notice there is a difference of the data scale among the two methods: the Li-Wong method is conducted at the original scale, and the

PSVD procedure is conducted at the nonlinear scale of the specific Box-Cox transformation. Because of the different scales in the Li-Wong and PSVD models, we transform the estimated expression index using the PSVD method back to the original scale for fair comparison. This can be performed using $\hat{\theta}_i^{\text{final}} = (\hat{\theta}_i + 1)^{\frac{1}{\beta}}$ for $i = 1, \ldots, I$, where $\hat{\theta}_i^{\text{final}}$ is a reasonable approximation of the expression index on the original scale.

It is known that all the genes but the spiked-in ones should be constantly expressed. So a good method should yield small variations in estimated expression indexes of the non-spiked-in genes across all the arrays. The left panel in Figure 1 shows the box plot of the standard errors of the estimated expression indexes of the nonspiked-in genes for both of the methods at logarithmic scale. The PSVD method is clearly superior to the Li-Wong method with overall smaller and less variable standard errors. The average (sample standard deviation) of the standard errors of all the genes are 40.49(77.53), and 11.05(25.41), respectively, for the Li-Wong and PSVD methods.

[Figure 1 about here.]

We also interrogated the 16 spiked-in genes spotted to the arrays at 14 concentration levels of $0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512$ and 1024 picomolars. It is intuitive to assess the correlation between the estimated expression indexes and the corresponding true concentration levels for each gene. Because the concentration levels are designed in the two-fold fashion, we examined the Pearson's correlation coefficient posterior to base 2 logarithm transformation of both the concentration levels and $\hat{\boldsymbol{\theta}}_{final}$. The box plots of the correlation coefficients of the 16 genes are displayed in the right panel of Figure 1. We can see that the PSVD method outperforms the Li-Wong model. The average (sample standard deviation) of the correlation coefficients for the Li-Wong and PSVD methods are, respectively, 0.98(0.02) and 0.99(0.01). In summary, the PSVD method empirically performs better than the Li-Wong method.

3.2 *Value of multivariate expression index*

We used the same dataset to explore applicability of the two-dimensional expression index. First, we observed that the second component is required by none of the 16 spiked-in genes without the transformation (or under the Li-Wong model) but required by 6 genes with the transformation (Box-Cox with $\hat{\beta} = 0.177$), based on the criterion defined in Section 2.2. On the other hand, the second component is required by $1,044$ nonspiked-in genes without the transformation, but only by 38 with the transformation.

We focus on the 6 spiked-in genes that require the two dimensional expression index for further investigation. Figure 2 shows an example gene *546_at*, in which the left upper panel corresponds to the case of using only the first singular vector and the left lower panel corresponds to the case of using the first two singular vectors. In both panels, the probes having the largest and the second largest residual variance upon fitting model (2) are highlighted with the symbols of the cross and the black circle, respectively. The plot of the residuals ϵ_{ij} upon using only the first singular vector and the residuals ϖ_{ij} upon using the first two singular vectors against the transformed data are contained in the left column. It is clearly seen that the residuals of the most variable probes using just one singular vector become much more homogeneous by using two singular vectors. We also examined the normal quantile-quantile plots of the residuals and observed that the residual distribution is closer to the normal distribution using two singular vectors.

[Figure 2 about here.]

The right column of Figure 2 indicates that the fit of the most variable probe 15, indicated by the crosses, is unsatisfying using the rank-one matrix approximation. It is intriguing to interrogate the benefit of using the two singular vectors for this probe. Model (2) tells that a linear relation between the transformed data $f(PM^*_{ij}|\hat{\beta})$ and the univariate expression index $\hat{\theta}_i$ $(i = 1, \cdots, I)$ should be observed for any probe j if the probes behave consistently. Such a plot for probe 15 is displayed in the right upper panel of Figure 2 and a random pattern is

shown. We examined the normally behaved probes and observed a clear linear pattern for all of them. In the right lower panel, we presented the plots of $\widehat{\theta_i^{(2)}}$ versus $f(PM_{ij}^*|\hat{\beta}) - \widehat{\theta_i^{(1)}}\widehat{\phi_j^{(1)}}$ obtained using the two-component model. We can observe a clear linear pattern for probe 15, implying that the second singular vector contains important information about this probe and makes substantial contribution to $\widehat{\theta_i^{(2)}}$.

The discussion above demonstrates the merit of using the second singular vector from the perspective of model fitting. Next, we investigate its empirical value using the known feature of the spike-in experiment. We focus on two spiked-in genes: *1091_at* and the earlier discussed *546_at*. With the known concentration level of each array for a spiked-in gene, we can examine the capability of distinguishing different concentration levels by using the two-dimensional expression index. For this purpose, we adopt K-means clustering procedure implemented using R function "pam" (Theodoridis and Koutroumbas, 2006). For gene *546_at*, we focused on the arrays at the concentration levels of $32, 64, 128, 256, 512$, and 1024 picomolars. The other gene *1091_at* uses the arrays with the lowest concentration levels of $0, 0.25, 0.5$, and 1 picomolars. This represents the most difficult case for discrimination because the concentration levels are the most close to each other. We apply the K-means procedure to cluster the arrays into three classes, using separately the univariate and two-dimensional expression indexes as the input. Performance of the two expression indexes will be evaluated based on the output of the K-means procedure.

An expression index is considered to be good if the corresponding clustering succeeds in assigning the same group membership to the arrays at the same concentration level and cluster together the arrays with the concentration levels at the similar magnitude. Figure 3 contains the plots of the concentration levels of the arrays versus their group membership produced by clustering. The left and right panels correspond to the univariate and two-dimensional expression index, respectively. Genes *546_at* and *1091_at* are shown respectively

in the upper and lower regions, separated by the middle line in black, of the two panels. For gene *546_at* and based on univariate expression index, we observe that an array is assigned to cluster B while two others are assigned to cluster A at the concentration level of 64 picomolar, and similarly an array is assigned to cluster C while two others are assigned to cluster B at the concentration level of 256 picomolar. In contract, the two-dimensional expression index succeeds in clustering all the replicate arrays together and grouping the arrays at different concentration levels correctly according to the order of the concentration magnitudes. For gene *1091_at*, the two-dimensional expression index again shows its promise in differentiating groups of different concentration levels.

[Figure 3 about here.]

The proposed procedure yields either an univariate or a multivariate index for a gene. For both testing of differential expression across the groups for a single gene and ranking among the genes, we can use the p-values derived from the ordinary two-sample t-test for an univariate index and the p-values derived from the Hotelling's T-square test (Mardia et al., 1979) for a multivariate index. For gene *1091_at* we performed Hotelling's T-square test between two groups of arrays composed of low concentration levels $\ell 1 = \{0, 0.25\}$, and $\ell 2 = 0.5$. The first group includes two concentration levels to ensure enough array replicates in the group. For the test of equal mean expression intensities between the two groups, using the univariate expression index obtains the p-value of 0.2 and the 95% confidence interval of $(-0.14, 0.51)$, while using the two-dimensional expression index obtains the p-value of 0.02 and the simultaneous 95% confidence interval of $(-0.29, 0.66)$ and $(-0.82, -0.07)$ for the two dimensions of the expression index. Thus, in this example, the two-dimensional expression index is more powerful to detect differential expression than the univariate expression index.

In summary, our empirical investigation shows that usage of the second singular vector has value in terms of improvement of both the model fit and capability of detecting differential expression.

4. Discussion of two practical issues

There is much discussion of array normalization in the literature. Many researchers view normalization as a way to make intensities of arrays comparable. For example, Bolstad et al. (2003) studied various normalization methods and demonstrated the good performance of quantile normalization. Quantile normalization intends to make a set of quantiles identical across the arrays and was incorporated in the robust multi-array analysis (RMA) method described in Irizarry et al. (2003b). The detailed discussion of normalization techniques is referred to Bolstad et al. (2003). The data preprocessing steps including normalization are conducted to perform more meaningful downstream analysis, in which detection of differential expression is usually a primary task. This problem is typically tackled using some classical test statistics or equivalently the associated p-value. For example, ordinary two-sample t test statistic is usually used for two-group comparison and preferred over the naive summary statistic, fold-change (FC, ratio of mean group intensity) rule, which does not account for the variation among the samples.

On the other hand, MAQC (2006) drew rather different conclusions via some real data studies. In particular, they showed that the impact of array normalization on detection of differential expression is trivial for various existing techniques and the classical test statistics or the corresponding p-values do not show the advantage over the simple fold change rule in terms of gene ranking.

Herein we re-examine these two issues using the same Affymetrix U95A spiked-in data set. The known truth associated with the specific design of this data set allows assessing various criteria including the performance of gene ranking. Our studies are based on gene expression index estimates obtained from both our PSVD and the Li-Wong model.

In this data set, all except for 16 spiked-in genes have constant expression across all the 59 arrays. The Latin square design creates multiple sets of 12 arrays where all the arrays in a

set have the common concentration levels for each of the 16 spiked-in genes. Therefore, such 12 arrays in a set can be treated as array replicates. Also among the sets, a spiked-in gene has different concentration levels. In our study, we focus on two such sets of 12 arrays and are interested in comparison of the arrays between these two groups.

We obtained the expression index estimates without normalization and with quantile normalization, respectively. Then we consider both the p-value obtained from the two-sample test and the simple absolute value of logarithm-transformed fold change (FC score) as the score for each gene. Negative p-value is used as the p-value score to be concordant with the order of the FC score. Intuitively, we expect the spiked-in genes to have smaller p-values and larger FC scores than all the other non-spiked-in genes. So it is sensible to assess the performance of gene summary score in terms of the power to distinguish between the two groups of non-spiked-in and spiked-in genes. For this purpose, we take a look at the receiver operating characteristic (ROC) curve and its commonly used summary statistic, the area under the curve (AUC) measurement. The results based on PSVD expression index estimates are shown in Figure 4. The ROC curves without normalization and using quantile normalization are plotted in the left upper and lower panels, respectively. In each panel, the p-value score and the FC score are plotted in solid and dotted lines, respectively, and the AUC measurements of the two scores are shown. We observe that quantile normalization clearly improves the discriminatory power of both scores over no normalization and the p-value score also has better performance than the FC score.

We also study the performance of gene ranking from another perspective. We plot the ranks of all the genes based on the p-value score along with those based on the FC score in the right panels of Figure 4. The line indicates the identical ranks between the p-value score and the FC score. The 16 spiked-in genes are indicated as the bold dots. In truth, the 16 genes should have higher ranks than all the other genes and, thus, they should locate

in the extreme right upper corner. Clearly, the normalization performs better than the no normalization case. The p-value score also outperforms the FC score, whereas the latter has more non-spiked-in genes ranking higher than the spiked-in ones. For example, in the normalization case the FC score obtains higher ranks for 5 out of the 16 genes with the largest rank difference of 13 while the p-value score outperforms by ranking the other 11 genes higher with the largest rank difference of 451. It is clear that the advantage of the p-value score over the FC score is more prominent in the no normalization case.

Next, we focus on study of the 16 spiked-in genes since their true concentration levels are known on all the arrays. Figure 5 shows the plot of ranks of the 16 genes based on each score along with their ranks based on the absolute difference of concentration levels between the two groups. We expect a score and the absolute concentration difference to be positively correlated. The cases without normalization and with quantile normalization are displayed in the upper and lower rows, respectively; and the FC score and the p-value score are displayed in the left and right columns, respectively. The corresponding Spearman's correlation coefficient is displayed in each panel. We notice that no normalization obtains correlation coefficients closer to zero than quantile normalization. The advantage of the p-value score is obvious over the FC score when the quantile normalization is applied, with the correlation coefficients of 0.202 and -0.141, respectively. The FC score yields the negative correlation coefficient and shows group-splitting pattern where some genes are positively correlated while the others show the opposite pattern. In the contrast, the p-value score in the right lower panel shows the highest correlation in the correct direction. This panel shows clearly that most genes are positively correlated, except for the three outliers in the right low corner. It appears that all the three genes have the highest concentration levels in at least a group, and thus this phenomenon could be caused by the saturation problem.

We obtained the similar results based on the expression index estimates of the Li-Wong model. The results are included in Web Appendix B.

In summary, our empirical investigations provide some evidence that: (1) normalization has non-negligible impact on detection of differentially expressed genes; (2) p-value has favorable performance in terms of ranking over the naive fold change score. We believe the difference between the p-value score and the FC score is mainly caused by that the former one takes into account the within-group variability while the latter one does not.

[Figure 4 about here.]

[Figure 5 about here.]

5. Discussion

We proposed a profile likelihood approach to estimate a transformation model for affymetrix short oligonucleotide array data. The proposed method is more statistically principled and efficient than the ad hoc entropy based method introduced in Hu et al. (2006) that is evident by our simulation studies. Its computationally efficiency comes with the simple SVD method used for parameter estimation. The real data example shows that it has some advantage over the popular Li-Wong model in terms of empirical performance. We also introduced a multivariate expression index which utilizes the first two singular vectors. Our empirical investigation shows the promise of using multivariate index in terms of both model fitting and differential expression detection. In addition, we re-examined two important practical issues in gene expression analysis, the value of normalization and statistical p-values. Our study shows that, when using the proposed method for generating expression indexes, normalization has impact on differential expression detection and statistical p-values have better performance than the simple fold change criterion in terms of gene ranking.

6. Supplementary Materials

Web Appendices referenced in Sections 1, 2, and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

7. Acknowledgement

We would like to thank an AE and a referee for their constructive comments. Mehdi Maadooliat and Jianhua Hu were supported by the National Science Foundation Grant DMS-0706818, National Institutes of Health Grants R01GM080503-01A1, R21CA129671, Cancer Center Support Grant P30 CA016672, and National Cancer Institute CA97007. Jianhua Z. Huang was supported by NSF (DMS-0606580, DMS-0907170), NCI (CA57030), and King Abdullah University of Science and Technology (KUS-CI-016-04). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

Achlioptas, D. and Mcsherry, F. (2007). Fast computation of low-rank matrix approximations. *Journal of the ACM* **54**.

Avriel, M. (1976). *Nonlinear Programming: Analysis and Methods*. Prentice-Hall, Englewood Cliffs, NJ.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1994). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16,** 1190–1208.

Chu, T.-M., Weir, B., and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Mathematical Biosciences* **176,** 35–51.

Geller, S. C., Gregg, J. P., Hagerman, P., and Rocke, D. M. (2003). Transformation and normalization of oligonucleotide microarray data. *Bioinformatics* **19,** 1817–1823.

Hu, J., He, X., Cote, G., and Krahe, R. (2009). Singular value decomposition-based alternative splicing detection. *Journal of the American Statistical Association* **104,** 944-953.

Hu, J., Wright, F. A., and Zou, F. (2006). Estimation of expression indexes for oligonucleotide arrays using the singular value decomposition. *Journal of the American Statistical Association* **101,** 41–50.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003a). Summaries of affymetrix genechip probe level data. *Nucleic Acids Res* **31,**.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264.

Lemon, W. J., Palatini, J. J., Krahe, R., and Wright, F. A. (2002). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. *Bioinformatics* **18,** 1470–1476.

Li, C. and Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *PNAS* **98,** 31–36.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14,** 1675–1680.

MAQC (2006). The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24,** 1151–1161.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L. (2003). *The analysis of gene expression data: Methods and Software*. Springer, New York.

Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition. 3rd edition*. Academic Press, London.

Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra.* SIAM: Society for Industrial and Applied Mathematics, Philadelphia, PA.

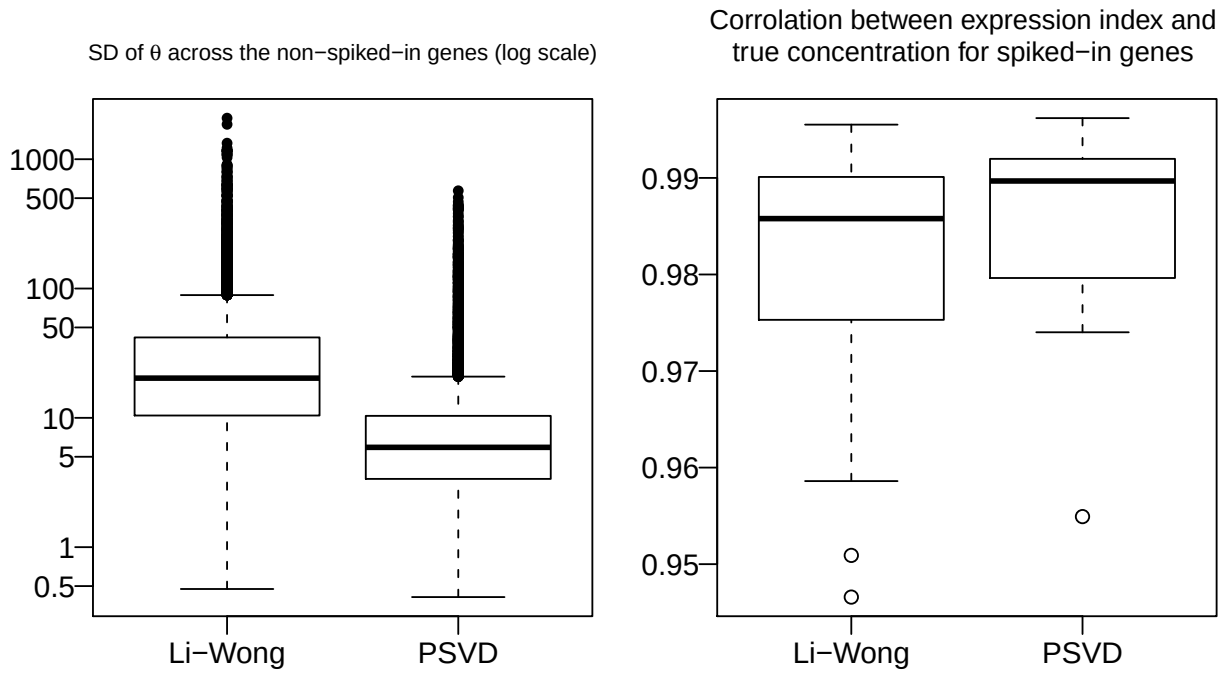

Figure 1. Left panel: Standard errors of estimated expression indexes of the nonspiked-in genes using the Li-Wong and PSVD methods. Right panel: Pearson's correlation coefficients between expression index($\hat{\boldsymbol{\theta}}_{final}$) and true concentration in base 2 logarithm of the spiked-in genes using the Li-Wong and PSVD methods.

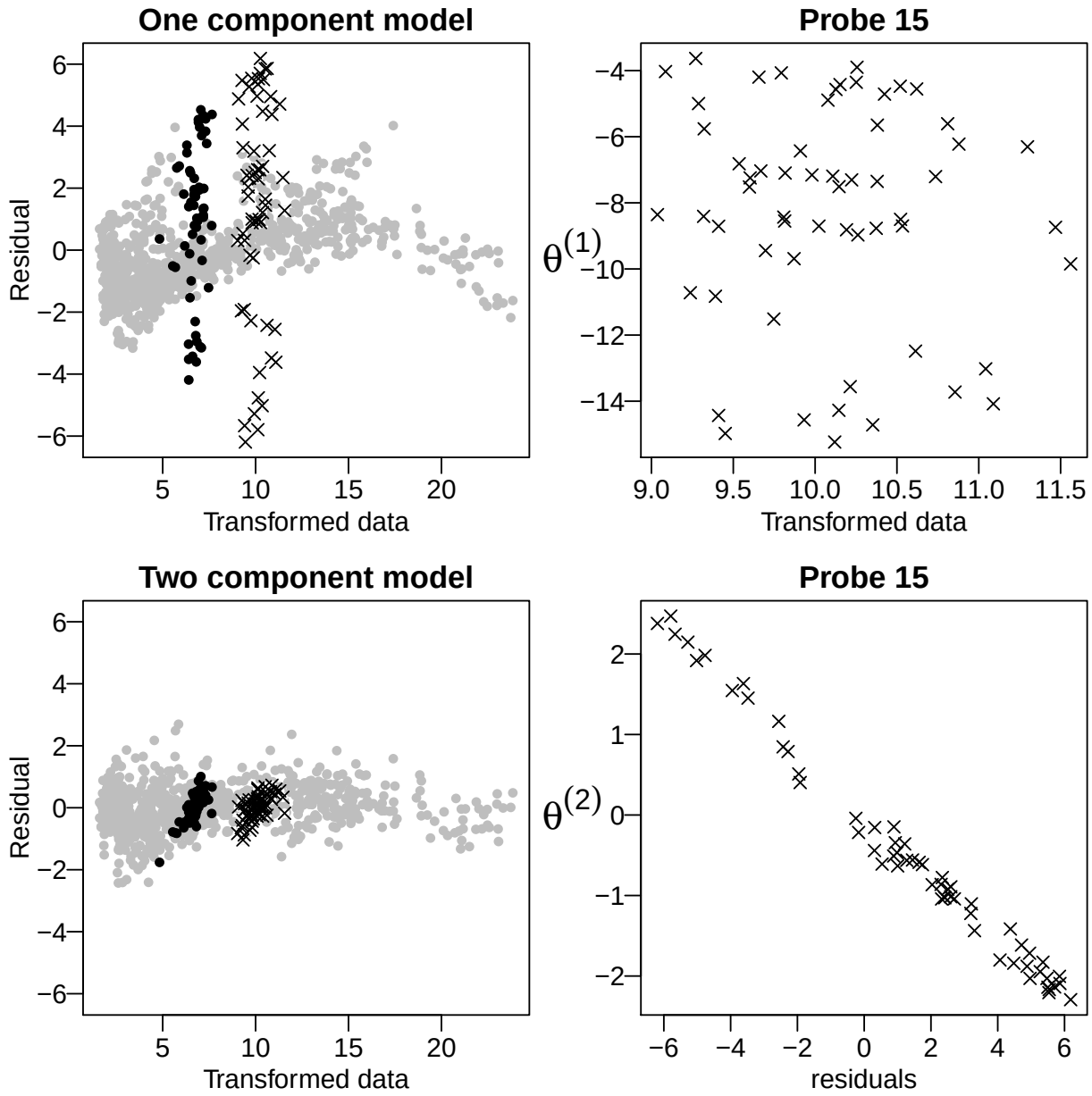

Figure 2. Model fitting using only one singular vector (upper panels) and two singular vectors (lower panels)‘. The left column contains the plot of residuals versus the transformed data. The right upper panel contains the plots of the transformed data $f(PM_{ij}^*|\hat{\beta})$ versus the estimated expression index $\widehat{\theta_i^{(1)}}$ for probe 15 of gene *546_at*; The right lower panel contains the plots of residuals $f(PM_{ij}^*|\hat{\beta}) - \widehat{\theta_i^{(1)}}\widehat{\phi_j^{(1)}}$ versus $\widehat{\theta_i^{(2)}}$ for the same probe.

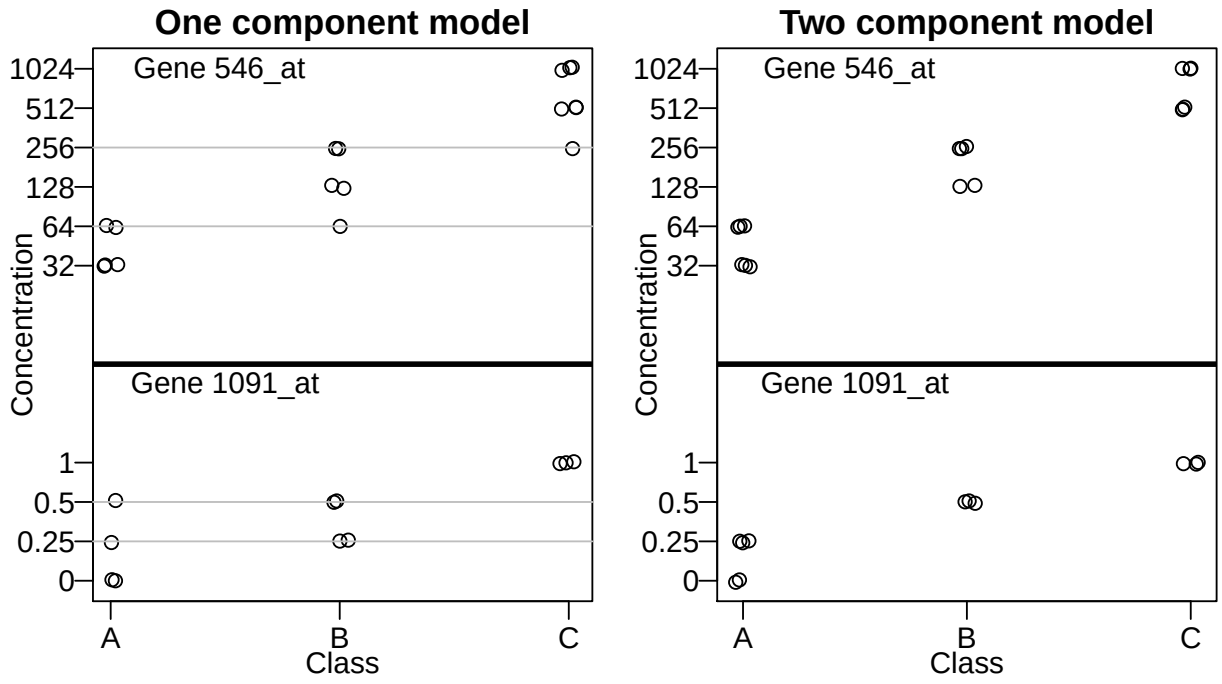

Figure 3. K-means clustering using univariate(left column) and two-dimensional(right column) expression index. The upper region corresponds to gene *546_at* at the concentration levels of $32, 64, 128, 256, 512$, and 1024 picomolars; The lower region corresponds to gene *1091_at* at the concentration levels of $0, 0.25, 0.5$, and 1 picomolars. Each panel contains the plot of the concentration levels of the arrays versus their group memberships produced by K-means clustering.

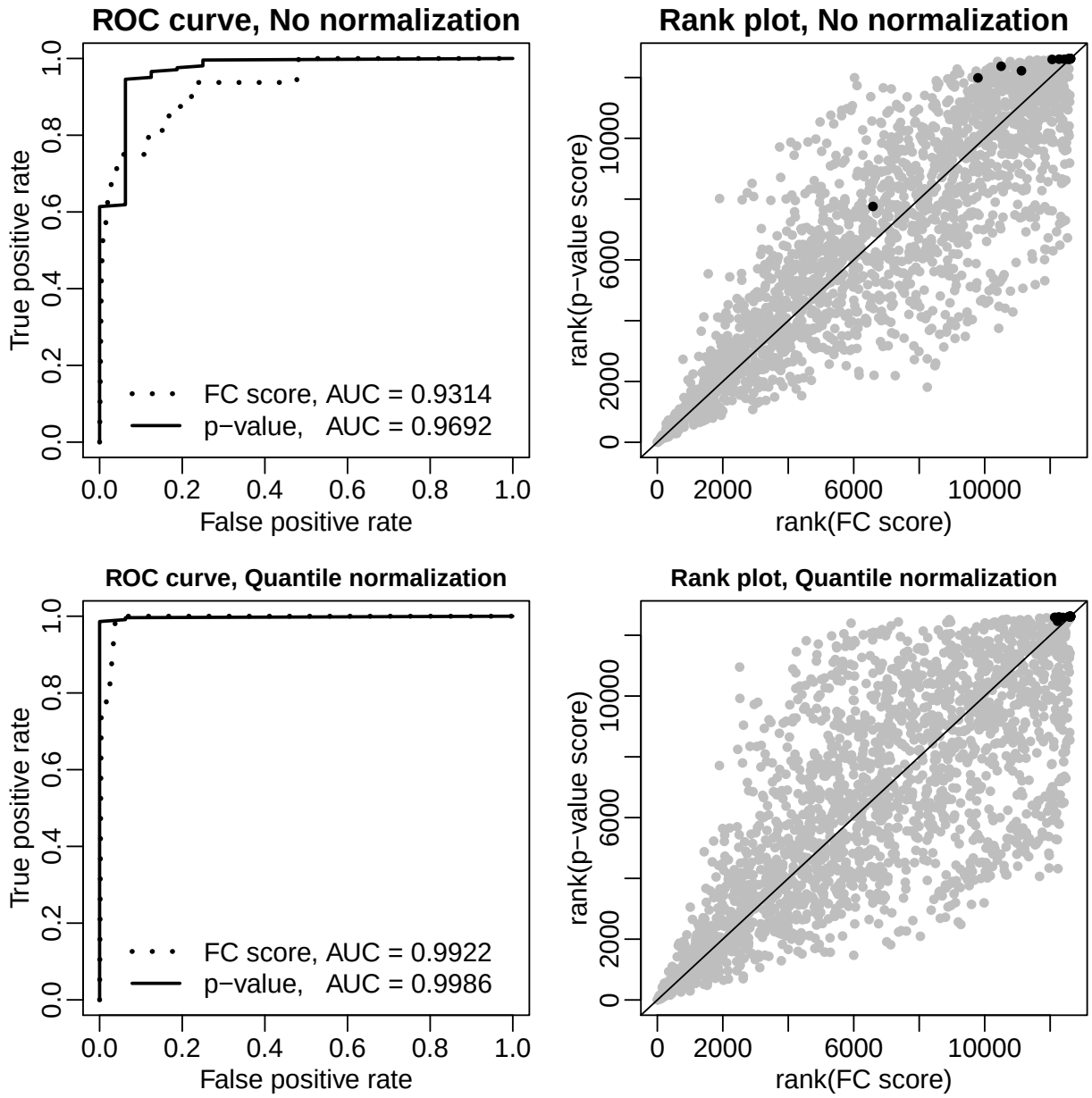

Figure 4. ROC curve and the rank plot for the PSVD Model. ROC curve for comparing the p-value and FC scores is given in the left column, and the rank plot for the rank(p-value score) vs. rank(FC score) is given in the right column. Upper row represents no normalization and the lower row represents the quantile normalization technique.

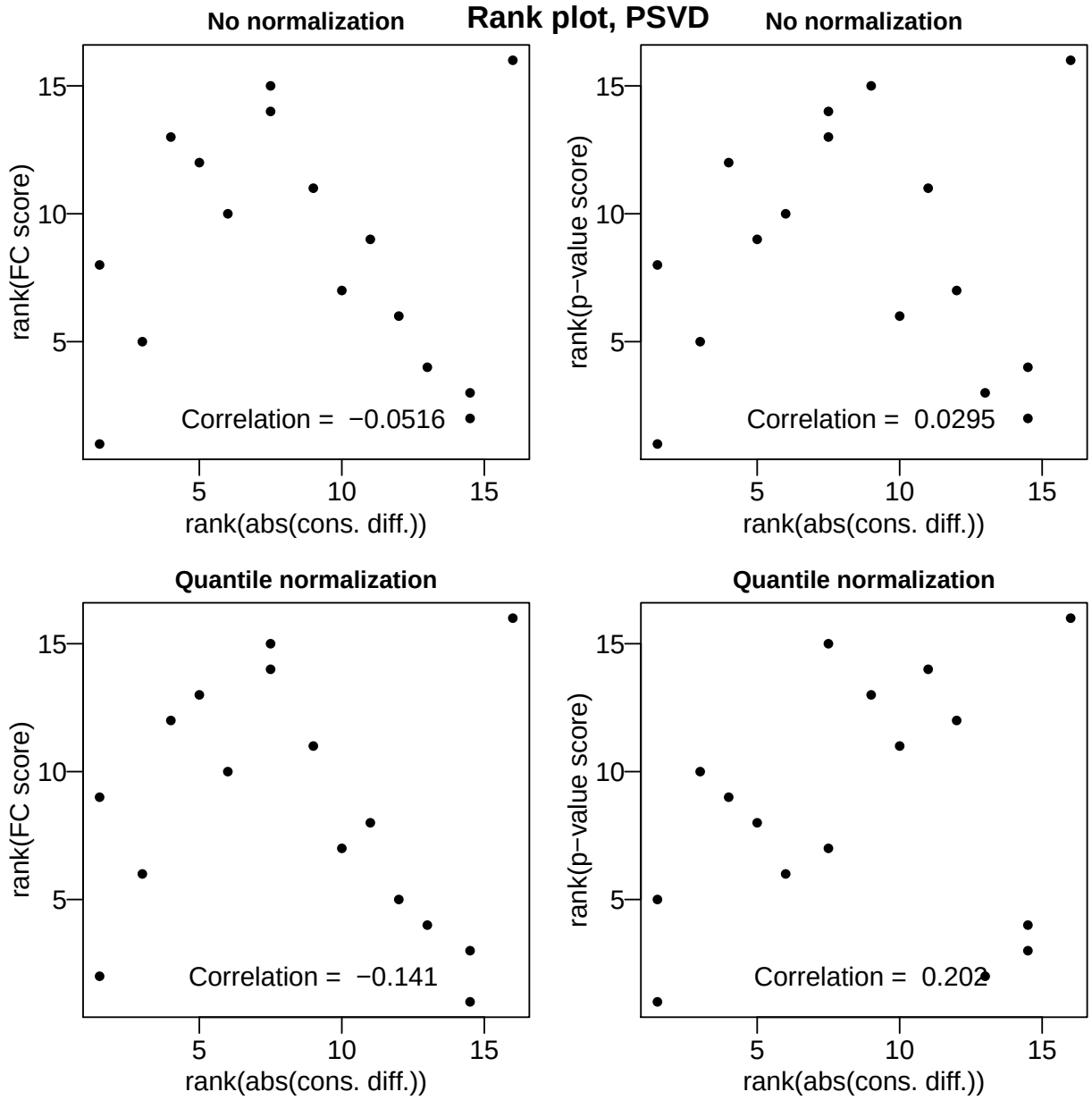

Figure 5. Rank plot for the 16 spiked-in genes based on the PSVD Model. Rank plot for the rank(FC score) vs. rank(abs(cons. diff.)) is given in the left column, and the rank plot for the rank(p-value score) vs. rank(abs(cons. diff.)) is given in the right column. Upper row represents no normalization and the lower row represents the quantile normalization technique.