Detecting Differential Expressions in GeneChip Microarray Studies: A Quantile Approach

Huixia WANG and Xuming HE

In this article we consider testing for differentially expressed genes in GeneChip studies by modeling and analyzing the quantiles of gene expression through probe level measurements. By developing a robust rank score test for linear quantile models with a random effect, we propose a reliable test for detecting differences in certain quantiles of the intensity distributions. By using a genomewide adjustment to the test statistic to account for within-array correlation, we demonstrate that the proposed rank score test is highly effective even when the number of arrays is small. Our empirical studies with real experimental data show that detecting differences in the quartiles for the probe level data is a valuable complement to the usual mixed model analysis based on Gaussian likelihood. The methodology proposed in this article is a first attempt to develop inferential tools for quantile regression in mixed models.

KEY WORDS: Gene expression; Microarray data; Quantile regression; Random effect; Rank score test.

1. INTRODUCTION

Recent advances in high-throughput technologies such as microarrays are playing a major role in our understanding of the molecular mechanisms that underlie normal and dysfunctional biological processes. Gene expression profiling on microarrays has enabled the measurement of thousands of genes in a single RNA sample. Gene expression data, however, are noisy, and require careful statistical analysis. One of the basic questions that interests the biologists is to identify differentially expressed genes across experimental conditions. In this article we focus on the GeneChip data from Affymetrix, a popular commercial oligonucleotide technology for studying gene expression. A GeneChip is a glass wafer synthesized with oligonucleotides, where each gene is represented by a probe set composed of 11–20 probe pairs, each of which consists of a perfect match (PM) probe and a mismatch (MM) probe. Each array contains a varying number of probe sets, usually in the thousands, depending on the experiments. The labeled mRNA samples are added to the arrays and hybridize to the probes. After hybridization, arrays are scanned and the scanned images are analyzed to obtain intensity measurements for each probe. These intensities measure how much hybridization has occurred in the corresponding probes.

A common approach to analyze GeneChip data is to summarize the probe intensities for each probe set, then use the summarized values for statistical inference. Among the most popular summarization methods are the Microarray Suite 5 (MAS5) of Affymetrix (2001), the model-based expression index of Li and Wong (2001), and the robust multiarray analysis (RMA) of Irizarry et al. (2003). Although the summarization methods are certainly useful, the probe level intensities contain more information and may give more power for statistical inference. Based on the probe level data, Chu, Weir, and Wolfinger (2002) compared the intensity measurements between different biological samples for each gene, using a mixed model that treats the array effect as a random effect,

$$y_{ijk} = \mu + T_i + P_k + a_{ij} + e_{ijk}, \qquad (1)$$

where y_{ijk} indicates the log base 2 transformed intensity measurement after some appropriate normalization, μ stands for the overall level, T_i is the ith treatment effect, P_k is the kth probe effect, a_{ij} is the effect of the jth array nested within the ith treatment, and the e_{ijk} is the random error. Both the a_{ij}'s and the e_{ijk}'s in their work are assumed to be iid normal random variables with mean 0, and variances σ_a^2 and σ_e^2, respectively, and the e_{ijk}'s are independent of the a_{ij}'s. By treating the array effect as a random effect, the model accounts for within-array correlation that is often seen in GeneChip data. Inference on T_i in model (1) can be made with the standard likelihood-based methods in linear mixed models; see, for example, Khuri, Mathew, and Sinha (1998) and Littell, Milliken, Stroup, and Wolfinger (1996). In essence, one compares the means of the log intensity distributions under different treatments. However, it is clear from microarray data that the normality assumptions are often severely violated for interesting genes. Unusual probes and outlying probe level measurements also occur frequently to upset normality. There are biological as well as technical reasons for those problems, as we shall discuss later in the article. To avoid the distributional assumptions, we propose a robust inferential method for model (1) by the means of quantile regression.

Instead of focusing on the changes in the mean, the quantile regression of Koenker and Bassett (1978) models the conditional quantiles, allowing us to test whether there is a change in the τth quantile of y for any given $\tau \in (0, 1)$. When the conditional distributions of y are non-Gaussian, the mean may not be the most appropriate summary, and a change in distributions may not be effectively detected through the means. We show that the *quantile* approach is a highly valuable complement to the *mean* approach for detecting differential gene expressions, especially in the presence of unusual probes or arrays in the microarray data.

Inference for linear quantile regression models has become a subject of intense study in recent years, and statistics and econometrics software (e.g., R, SAS, and Stata) has started to include inferential methods in their packages for quantile regression. However, existing inferential methods such as those discussed by Koenker (2005) and Kocherginsky, He, and Mu (2005) are developed for independent data. The primary objective of the present article is to develop a rank score test for

Huixia Wang is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *wang@stat.ncsu.edu*). Xuming He is Professor, Department of Statistics, University of Illinois, Champaign, IL 61820 (E-mail: *x-he@uiuc.edu*). The research was supported in part by NSF awards DMS-01-02411 and DMS-06-04229, and the NIH–NIDCD grant R01-DC3378. The authors thank Roger Koenker and Stephen Portnoy, and an associate editor for stimulating discussions and helpful suggestions, and the Bioinformatics Unit at the University of Illinois for making available several microarray datasets, including those not disclosed in this article.

Journal of the American Statistical Association
???? 0, Vol. 0, No. 00, Theory and Methods
DOI 10.1198/00

quantile regression models with a random effect as in model (1). We recognize that most other inferential methods require a modestly large sample size, but microarray experiments are often carried out with a small number of arrays. Therefore we propose to use a genomewide estimate of within-array correlation to preserve a valid significance level and the desired false positive rates at small sample sizes. The details of the proposed rank score test are given in Section 2, and Monte Carlo assessments of the test are provided in Section 3. By applying the rank score test to two GeneChip studies, we demonstrate in Section 4 how and why our proposed quantile approach adds value to GeneChip studies. Some concluding remarks are made in Section 5. Some technical details are given in the Appendix.

Although the present article focuses on Affymetrix GeneChip data, its primary goal is to develop inferential tools based on quantiles. In microarray studies, there are other important issues that are of interest or concern to statisticians. For example, Hu and He (2006) and Fan, Huang, and Peng (2005) proposed new methods of normalization, and Fan, Chen, Chan, Tam, and Ren (2005) used a semilinear in-slide model for Affymetrix arrays, all of which can aid detection of significant genes. For the sake of focus, we do not elaborate on these issues in this article.

2. QUANTILE REGRESSION AND PROPOSED TEST

In this section we consider a more general form

$$y_{ijk} = \mathbf{x}_{ijk}^T\boldsymbol{\alpha} + \mathbf{z}_{ijk}^T\boldsymbol{\beta} + u_{ijk}, \quad i = 1,\ldots,I, j = 1,\ldots,J_i, k = 1,\ldots,K_{ij}, \quad (2)$$

where y_{ijk} is defined in the same way as before, $\mathbf{x}_{ijk}$ and $\mathbf{z}_{ijk}$ are the $p \times 1$ and $q \times 1$ design vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the p- and q-dimensional parameters, which in model (1) represent the probe and treatment effects, respectively, and $u_{ijk} = a_{ij} + e_{ijk}$ is the composite error term. In addition to the treatment and probe effects, other covariates (e.g., log mismatch intensity or some phenotype measurements) are possible in model (2). Obviously, the model arises in other applications too, where an additive random effect represents subjects or clusters. Throughout this article, we assume that the first element of $\mathbf{x}_{ijk}$ is 1, corresponding to the intercept in $\boldsymbol{\alpha}$.

For any given τ, we consider the τth quantile of y given $(\mathbf{x}, \mathbf{z})$. For identifiability, we assume that the τth quantile of u is zero. Other than that, no distributional assumptions are made on u. We consider testing the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus the alternative hypothesis $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$.

Following Koenker and Bassett (1978), the quantile regression estimate of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is obtained by minimizing $\sum_{ijk} \rho_\tau(y_{ijk} - \mathbf{x}_{ijk}^T\boldsymbol{\alpha} - \mathbf{z}_{ijk}^T\boldsymbol{\beta})$, where $\rho_\tau(u) = u(\tau - I_{\{u<0\}})$ is the quantile loss function. Under mild conditions, the estimate is asymptotically normal, so the Wald-type tests can be carried out under the justification of large sample theory; see He, Zhu, and Fung (2002). Related work appears in Koenker (2004), where a shrinkage approach is used to predict a_{ij} in addition to the estimation of the fixed effects. In either approach, the large sample variance of the quantile estimate involves the unknown density of u and the large-sample inference based on the Wald-type test is generally unstable at small sample sizes. To avoid direct estimation of the nuisance parameter (i.e., density), we turn to the idea of rank score test for quantile regression as considered by Gutenbrunner, Jurêcková, Koenker, and Portnoy (1993), hereafter GJKP.

2.1 Quantile Rank Score

We focus on the quantile rank score, which was proposed in GJKP for iid error models and is now extended to model (2). When $\tau = .5$, it is a generalization of the sign test in the univariate sample. First, we fix the notation.

Let $n = \sum_i^I \sum_{j=1}^{J_i} K_{ij}$ and $L = \sum_i^I \sum_{j=1}^{J_i} K_{ij}(K_{ij} - 1)$. Upon rewriting model (2) in matrix form as $\mathbf{Y}_n = \mathbf{X}_n\boldsymbol{\alpha} + \mathbf{Z}_n\boldsymbol{\beta} + \mathbf{U}_n$, we have $\mathbf{Y}_n$ and $\mathbf{U}_n$ as n-dimensional vectors, and $\mathbf{X}_n$ and $\mathbf{Z}_n$ as $n \times p$ and $n \times q$ matrices, respectively. Furthermore, we let $\mathbf{H} = \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T$ and $\mathbf{Z}_n^* = (\mathbf{z}_{ijk}^*)_{n\times q} = (\mathbf{I} - \mathbf{H})\mathbf{Z}_n$.

The piecewise derivative of $\rho_\tau(u)$ is called the score function $\psi_\tau(u) = \tau - I_{\{u<0\}}$. In what follows, we write $\widetilde{\mathbf{X}}_n = (\mathbf{X}_n, \mathbf{Z}_n)$ and use F_1 to denote the common marginal distribution function of u_{ijk}, for any i, j, k.

The quantile rank score test is based on

$$\mathbf{S}_n = n^{-1/2} \sum_{ijk} \mathbf{z}_{ijk}^* \psi_\tau(\hat{u}_{ijk}(\tau)),$$

where $\mathbf{z}_{ijk}^*$ are elements of $\mathbf{Z}_n^*$ and $\hat{u}_{ijk}(\tau) = y_{ijk} - \mathbf{x}_{ijk}^T\hat{\boldsymbol{\alpha}}(\tau)$ are the residuals of the quantile estimate obtained under the null hypothesis. Letting

$$\mathbf{Q}_n(\delta) = n^{-1} \sum_{ijk} \mathbf{z}_{ijk}^* \mathbf{z}_{ijk}^{*T} \tau(1-\tau) + n^{-1} \sum_{ij} \sum_{k_1 \neq k_2} \mathbf{z}_{ijk_1}^* \mathbf{z}_{ijk_2}^{*T} (-\tau^2 + \delta) \quad (3)$$

with

$$\delta = P(u_{111} < 0, u_{112} < 0), \quad (4)$$

we define the quantile rank score test statistic as

$$T_n(\tau) = \mathbf{S}_n^T\{\mathbf{Q}_n(\hat{\delta})\}^{-1}\mathbf{S}_n, \quad (5)$$

where

$$\hat{\delta} = (L-p)^{-1} \sum_{ij} \sum_{k_1 \neq k_2} I_{\{\hat{u}_{ijk_1}(\tau)<0, \hat{u}_{ijk_2}(\tau)<0\}}. \quad (6)$$

The second term on the right side of (3) accounts for the dependence within arrays. If the u_{ijk}'s are iid, then this reduces to the quantile rank score test of GJKP (with $\delta = \tau^2$). If the number of arrays J_i tends to infinity, then, without surprise, we have the following asymptotic limiting distribution.

Theorem 1. Under Assumptions A.1–A.4 spelled out in the Appendix, the null distribution of $T_n(\tau)$ converges to χ_q^2 as $\min_i\{J_i\} \to \infty$.

The proof of Theorem 1 is outlined in the Appendix. In Section 3 we see through simulation that this δ adjustment to $\mathbf{Q}_n$ is critical for valid inference when the errors u_{ijk} are correlated.

2.2 Variations in Estimating δ

When the chi-squared distribution is used as the reference distribution, the small-sample performance of the test $T_n(\tau)$ depends on how δ is estimated in $\mathbf{Q}_n$. Our Monte Carlo simulation study shows that the particular estimate (6), based on the signs of the residuals obtained under H_0, does well in preserving the significance level of the test even for J_i as small as 3. This quantile rank score test will be referred to as QRS_0. Using an alternative estimate of δ based on the residuals obtained under

the alternative hypothesis leads to an asymptotically equivalent test, later referred to as QRS_1. Monte Carlo comparisons (not reported in this article) indicate that QRS_1 has higher power at the cost of slightly inflated level. After calibrating for type I errors, these two variants of the test have about the same performance.

When the number of arrays J_i is small (in the range of 3–10), we propose to take advantage of the large number of genes available on those arrays to obtain a more stable estimate of δ. If we average the $\hat{\delta}$'s used in QRS_0 across genes to obtain a combined estimate $\bar{\delta}$, we then conduct the quantile rank score test by rejecting H_0 if $T_c(\tau) = \mathbf{S}_n^T\{\mathbf{Q}_n(\bar{\delta})\}^{-1}\mathbf{S}_n > \chi_q^2(\alpha)$, the upper αth quantile of the chi-squared distribution with q degrees of freedom. The resulting test will be denoted QRS_c.

It is not unreasonable to assume that the sign correlations of the error terms in the probe intensity measures are about the same across the genes; these correlations are all due to the probe sets living on the same arrays. The common δ approach is a simplistic form of shrinkage. As demonstrated in Section 3, our proposed test QRS_c tends to outperform QRS_0 and QRS_1 when the J_i's are small, even when the true δ values are not exactly shared across the genes. However, more sophisticated shrinkage methods toward information-sharing tests across genes may be developed in future work using empirical Bayes ideas such as those behind the refined t test given in Lönnstedt and Speed (2002).

2.3 Heteroscedastic Errors

The assumption of homoscedastic errors in u_{ijk} simplifies the derivation of $\mathbf{Q}_n$ in the quantile rank score test, but in reality, the quantile rank score test for the hypothesis on treatment effects is highly robust against heteroscedastic errors. Consider the case where the u_{ijk}'s are not exchangeable, but $\delta_{ij,k_1,k_2} = P(u_{ij,k_1} < 0, u_{ijk_2} < 0) = \delta_{i,k_1,k_2}$ are common across j for given i, k. Theorem 1 remains valid if $\mathbf{Q}_n$ is replaced by

$$\mathbf{Q}_{n,2} = n^{-1}\sum_{ijk}\mathbf{z}_{ijk}^*\mathbf{z}_{ijk}^{*T}\tau(1-\tau) + n^{-1}\sum_{ij}\sum_{k_1\neq k_2}\mathbf{z}_{ijk_1}^*\mathbf{z}_{ijk_2}^{*T}\left(-\tau^2+\delta_{i,k_1,k_2}\right).$$

For testing the treatment effect with $z_{ijk} = z_{ij}$ (=1 or −1), this reduces to (3) with δ as the average of δ_{i,k_1,k_2} over i and k, which can be consistently estimated by (6). Therefore, the quantile rank score test given in Section 2.1 remains valid in this rather realistic heteroscedastic error models for GeneChip studies, where the replicate arrays are taken to be exchangeable. For more general forms of heteroscedasticity, we refer to the doctoral dissertation of the first author (Wang 2006).

2.4 Combined Tests on Multiple Quantiles

A change in the gene expression measures might be most distinguishable in the upper or lower quantiles, depending on the distribution of the probe level measurements. It might be useful to formulate joint hypotheses about the relevance of certain groups of covariates at several quantiles, but here we restrict ourselves to the partitioned mixed model (2) and propose a combined rank score test.

The null hypothesis to be tested is

$$H_0 : \beta(\tau_1) = \beta(\tau_2) = \cdots = \beta(\tau_l) = 0,$$

where $\boldsymbol{\beta}(\tau_a)$ is the coefficient vector of the quantile regression model at the quantile level τ_a for a set of $0 < \tau_1 < \tau_2 < \cdots < \tau_l < 1$.

Let

$$\mathbf{D}_{1n} = n^{-1}\sum_{ijk}\mathbf{z}_{ijk}^*\mathbf{z}_{ijk}^{*T} \quad \text{and}$$

$$\mathbf{D}_{2n} = n^{-1}\sum_{ij}\sum_{k_1\neq k_2}\mathbf{z}_{ijk_1}^*\mathbf{z}_{ijk_2}^{*T},$$

and define

$$\mathbf{W}_n = \left(\mathbf{S}_n^{(1)}, \mathbf{S}_n^{(2)}, \ldots, \mathbf{S}_n^{(l)}\right),$$

where $\mathbf{S}_n^{(a)} = n^{-1/2}\sum_{ijk}\mathbf{z}_{ijk}^*\psi_{\tau_a}(\hat{u}_{ijk}(\tau_a))$ is the quantile regression score at the quantile level τ_a, $a = 1, \ldots, l$. The asymptotic covariance matrix of $\mathbf{W}_n$, $\mathbf{V}_n^* = (\mathbf{v}_n^{(ab)*})$, $1 \le a, b \le l$, is given by

$$\mathbf{v}_n^{(ab)*} = (\tau_a - \tau_a\tau_b)\mathbf{D}_{1n} + (-\tau_a\tau_b + \delta(\tau_a, \tau_b))\mathbf{D}_{2n},$$

where $\delta(\tau_a, \tau_b) = P(u_{111}(\tau_a) < 0,\ u_{112}(\tau_b) < 0)$, $u_{111}(\tau_a) = u_{111} - F_1^{-1}(\tau_a)$, and $u_{112}(\tau_b) = u_{112} - F_1^{-1}(\tau_b)$.

For the sake of clearer presentation, we denote $\hat{u}_{ijk}(\tau_a)$ as the estimated residuals from the quantile regression model at the quantile level τ_a. Then we can estimate the δ's by

$$\hat{\delta}(\tau_a, \tau_b) = (L-2p)^{-1}\sum_{ij}\sum_{k_1\neq k_2} I_{\{\hat{u}_{ijk_1}(\tau_a)<0, \hat{u}_{ijk_2}(\tau_b)<0\}},$$

and foregoing term $(L-2p)^{-1}$ should be replaced by $(L-p)^{-1}$ when $\tau_a = \tau_b$, consistent with (6). Furthermore, as with QRS_c in the preceding section, we may carry out the combined test by assuming common δ's across genes, and we refer to this combined rank score test as CRS_c. By replacing the nuisance parameters with the corresponding estimates, we obtain a consistent estimator of $\mathbf{V}_n^*$ and denote it as $\mathbf{V}_n$.

Theorem 2. Under Assumptions A.1–A.4 in the Appendix, we have, under H_0,

$$T_n = \mathbf{W}_n^T\mathbf{V}_n^{-1}\mathbf{W}_n \xrightarrow{D} \chi_{lq}^2 \quad \text{as } \min_i\{J_i\} \to \infty. \tag{7}$$

For GeneChip data, we propose testing the hypotheses at the three quartiles. An alternative to the combined test is to use the Bonferroni adjustment to the p values of these three individual tests. The relative performances of these two approaches depend on the correlations among the individual scores.

3. MONTE CARLO SIMULATIONS

This section summarizes our findings on the performance of the quantile rank score tests through Monte Carlo simulations. We generate data from models that mimic those we encountered in GeneChip experiments.

3.1 Model Descriptions

The simulation study is based on model (1). We assume that the number of treatments $I = 2$, the number of probes $K = 16$, and each treatment has J replicate arrays. The parameter μ is chosen to be 8 and the probe effects P_i are generated from $N(0, 4)$. The parameter $\beta = T_1 - T_2 = T_1$ (with $T_2 = 0$) measures the treatment effect. The following four different cases are considered in this study.

- *Case 1*. Fixed-effect linear model. The a_{ij}'s are set to 0 and the e_{ijk}'s are generated randomly from the distribution $N(0, .4^2)$.
- *Case 2*. The a_{ij}'s are generated randomly from the distribution $N(0, .2^2)$ and the e_{ijk}'s are generated from $N(0, .4^2)$.
- *Case 3*. The a_{ij}'s are generated from $N(0, .2^2)$ and the e_{ijk}'s are set at $.4z$, where z follows the mixture distribution $.9N(0, 1) + .1t_1$.
- *Case 4*. The a_{ij}'s and e_{ijk}'s are generated from $N(0, \sigma_a^2)$ and $N(0, \sigma_e^2)$, respectively, where σ_a is chosen to be .2 and is kept unchanged, σ_e^2 is set to be $\sigma_a^2(1-\gamma)/\gamma$ varying from gene to gene, and $\gamma = \sigma_a^2(\sigma_a^2 + \sigma_e^2)^{-1}$ is the intra-array correlation coefficient. The γ's are generated by converting Fisher's z, which is randomly chosen from $N(.2, .1^2)$, back to the correlation scale. For the particular z generated in our study, δ ranges between .25 and .32.

In each case, we generate 100 datasets, each consisting of 120 genes, of which 100 genes are nondifferentially expressed ($\beta = 0$) and 20 genes are differentially expressed ($\beta \neq 0$). This is a simplistic case where those 20 genes can be viewed as a cluster. The fixed-effect parameter values are held constant across all simulations. We also vary the treatment effect β under H_1 from .1 to 1 (in increments of .1) in the study. For the rank score tests, we concentrate on the median regression here.

3.2 Simulation Results

For each simulated dataset, we test the null hypothesis on 120 genes simultaneously, with the false discovery rate (FDR) adjustment following Benjamini and Hochberg (1995). The genes with FDR adjusted p values smaller than .05 are identified as significant. The Benjamini and Hochberg adjustment is chosen for convenience, but other FDR adjustments can be expected to yield similar comparisons.

Table 1 summarizes the results in terms of true positives (TP) and false positives, where TP is the number of identified genes that are truly differentially expressed (with an ideal value of 20) and FDR is the empirical FDR, averaged over the 100 datasets. The FDR for a given dataset is taken as 0 when no gene is detected. We report the results for three values of J and several values of β. The test QRS denotes the quantile rank score test without any δ adjustment; MIX denotes the test carried out using PROC MIXED from SAS Systems.

Some inflated FDRs are shown in bold in the table. It is evident that QRS without the δ adjustment loses complete control of FDR in Cases 2–4, where the measurements y_{ijk} are dependent. This affirms that the δ adjustment to Q_n is critical when the errors are correlated.

When J is small (3 and 5), the test QRS_1 has much higher FDR than the nominal level, which indicates that the chi-squared approximation for QRS_1 deteriorates at small samples.

Table 1. The Number of True Positives (TP) and the Estimated False Discovery Rate (FDR) in Cases 1–4

	QRS		QRS_0		QRS_1		QRS_c		MIX	
Experiment	*TP*	*FDR*	*TP*	*FDR*	*TP*	*FDR*	*TP*	*FDR*	*TP*	*FDR*
Case 1										
$J = 3, \beta = .3$	13	**.106**	0	.000	14	**.268**	7	.020	0	.000
$J = 5, \beta = .3$	17	.062	0	.000	17	**.152**	15	.026	15	.000
$J = 25, \beta = .3$	20	.044	20	.038	20	.060	20	.020	20	.031
Case 2										
$J = 3, \beta = 1.0$	20	**.523**	0	.000	20	**.282**	19	.011	11	.011
$J = 5, \beta = .7$	20	**.468**	8	.002	19	**.161**	19	.023	18	.043
$J = 25, \beta = .3$	20	**.443**	19	.034	19	.051	19	.030	19	.033
Case 3										
$J = 3, \beta = 1.0$	17	**.160**	0	.000	15	**.268**	10	.022	0	.000
$J = 5, \beta = .7$	14	**.149**	0	.000	12	**.191**	9	.033	0	.000
$J = 25, \beta = .3$	13	**.146**	7	.031	10	.075	9	.043	1	.003
Case 4										
$J = 3, \beta = 1.0$	20	**.521**	0	.000	19	**.280**	18	.017	10	.010
$J = 5, \beta = .7$	20	**.473**	6	.001	19	**.163**	17	.030	17	.040
$J = 25, \beta = .3$	20	**.449**	18	.032	19	.056	18	.039	19	.032

NOTE: The desired FDR is .05. Some inflated FDRs are shown in bold.

This problem is mostly rectified by QRS_c, which controls FDR reasonably well without losing much power when the true δ are constant (Cases 1–3), as well as when they differ slightly from gene to gene (Case 4), supporting our preference for a genomewide adjustment of δ. For $J = 25$, the three variations QRS_0, QRS_1, and QRS_c perform very similarly.

Generally speaking, the proposed QRS_c is very competitive to MIX even in the Gaussian cases, where the latter loses some power because the FDR adjustment is too conservative. In Case 3, where e_{ijk} follows a mixture distribution of the standard normal and t_1, QRS_c is clearly more powerful than MIX. It is interesting to note that the efficiency of median regression for Gaussian data is clearly higher than 64%, the well-known asymptotic efficiency of the univariate sample median relative to the mean. In fact, the relative efficiency of the median increases for correlated data, which can be verified both mathematically and empirically.

3.3 Chi-Squared Approximation versus Resampling

When J is small, we naturally ask whether we can better approximate the p values than the (limiting) chi-squared distribution for the quantile rank score statistic. We considered permutation, arraywise bootstrap, and blocked wild bootstrap to generate reference distributions for the quantile rank scores, and we compared their performances with that of QRS_0. The details of our Monte Carlo study are skipped, but the results show that use of the chi-squared distribution for the quantile rank score test is hard to beat by those methods, even for small J.

4. EMPIRICAL DATA ANALYSIS

In this section we apply the proposed rank score tests to identify differentially expressed genes across two experimental conditions to two GeneChip studies. These examples demonstrate the value of the quantile approach in the analysis of GeneChip data.

4.1 Spiked-in Study

The spiked-in dataset from GeneLogic (*http://qolotus02.genelogic.com/datasets.nsf/*) consists of 32 human GeneChip

Table 2. Expected and Observed Ranks of the Spiked-in Probe Sets

Probe set	Concentration T1	Concentration T2	Expected rank	Observed rank RMA	Q_1	Q_2	Q_3	MIX	CRS_c
BioB-5_at	100.0	.5	1	1	1	1	1	2	1
BioB-3_at	.5	25.0	2	2	1	1	1	3	1
BioC-5_at	2.0	75.0	3	3	8	7	1	73	7
BioB-M_at	1.0	37.5	3	5	1	1	1	1	1
BioDn-3_at	1.5	50.0	5	4	1	1	1	5	1
DapX-3_at	37.5	3.0	6	7	1	1	1	6	1
CreX-3_at	50.0	5.0	7	9	7	7	1	7	8
CreX-5_at	12.5	2.0	8	8	10	118	1,858	2,179	10
BioC-3_at	25.0	10.0	9	6,431	1,096	114	9	12,319	11
DapX-5_at	5.0	1.5	10	10	9	118	1,266	3,433	9
DapX-M_at	3.0	1.0	11	6	1	1	1	4	1

arrays (HG-U95A) and all the arrays have a common background cRNA derived from an acute myeloid leukemia tumor cell line. In the experiment, 11 different cRNA fragments were hybridized at different picomolar concentrations to each array (apart from replicates). In this example, we choose six arrays from a pair of triplicates ($T1$ and $T2$), and the spiked-in concentrations for each of the 11 control cRNAs are shown in Table 2. Irizarry et al. (2003) provided some details on the experimental design.

The total number of probe sets in this study is 12,626. The data are normalized with the quantile normalization method from Bioconductor's affy package, using the default setting as of March 2006. We consider the log base 2 transformed perfect match (PM) intensity as dependent variable y. In this study we expect only those 11 spiked-in probe sets to be differentially expressed.

Considering the small number of replicates in the study, we choose to use the quantile rank score test QRS_c at three quartiles denoted Q_1, Q_2, and Q_3. The method MIX on the means is also included for comparison. To study the genes that have a difference between two concentration groups in one of the three quartiles, we also performed the combined rank score test CRS_c and the Bonferroni correction (Bonf) to the p values of the four individual tests (Q_1, Q_2, Q_3, and MIX) for each gene. A .05 cutoff is used thereafter for the FDR adjusted p values to identify differentially expressed genes. Table 3 summarizes the total number of significant probe sets (total positive), the number of correctly identified spiked-in probe sets (true positive), and the number of misidentified probe sets (false positive) for each method used. The Q_2 and Bonf perform similarly, both detecting 10 probe sets, 8 of which are the spiked-ins. The combined rank score test detects all 11 spiked-in probe sets successfully without a false positive.

Using the same dataset, Irizarry et al. (2003) calculated the observed ratios or "fold changes" between the two averages over the triplicate RMA summarized measures and obtained the ranks of the ratios. Here, we provide ranks based on the test statistics proposed in the present article as compared to the ranks obtained under RMA. For the probe sets that represent spiked-in cRNAs, the observed ranks of either the fold changes or the test statistics should, ideally, coincide with those of the true fold changes of the spiked-in concentrations. Table 2 summarizes the spiked-in concentrations for each of the 11 control cRNAs on the two sets of triplicates ($T1$ and $T2$), the expected ranks of the true fold changes, the observed ranks of RMA, QRS_c at three quartiles (Q_1, Q_2, and Q_3), MIX, and the combined rank score test (CRS_c). The quantile score test statistics often have ties; that is why, for example, the rank 1 is given to eight probe sets under Q_3 and, in this case, the next rank is 9.

Table 3. Summary Statistics for the Spiked-in Study

Tests	Total positive	True positive	False positive
Q_1	7	7	0
Q_2	10	8	2
Q_3	0	0	0
MIX	0	0	0
CRS_c	11	11	0
Bonf	10	8	2

From Table 2, we notice that 8 out of the 11 spiked-in genes are easy to detect, but 3 of them, namely CreX-5_at, BioC-3_at, and DapX-5_at show up in the top only by some measures. The combined test statistic CRS_c is a lucky winner, because it separates all the 11 spiked-in probe sets from the rest. Due to the small number of replicates, significance tests such as the two-sample t test applied to the RMA summarized expression indices fail to detect any spiked-in genes after the FDR adjustment, even though the ranks based on the RMA summarized values reflect the expected ranks quite well (except for one probe set). The use of probe level data results in a more powerful statistical inference here.

To explain some of the discrepancies among various tests used, we focus on DapX-5_at, which is singled out by the quantile rank score test statistic at Q_1, but not by any test statistic (say, MIX) on the mean. Figure 1 gives the expression profile of this probe set, where the x axis is the probe number and the y axis is the normalized $\log_2$(PM) centered by the median of six arrays within each probe. It is clear from the plot that array 4 shows the opposite direction of regulation from arrays 5 and 6. Earlier studies of this dataset did not identify this phenomenon and the reason for this outlying array is unknown. Possible reasons are image scratches and dust, but if one analyzes the probe level data with no awareness of outliers of this nature, inference based on averages may take its toll. In this case, array 4 leads to a large variance estimate under the Gaussian model, but the first quartiles are still clearly distinguishable for the two groups. If array 4 is removed, all the tests give consistent results.

4.2 Smoking Study

The smoking study was conducted by Boston University School of Medicine. The raw dataset was downloaded from

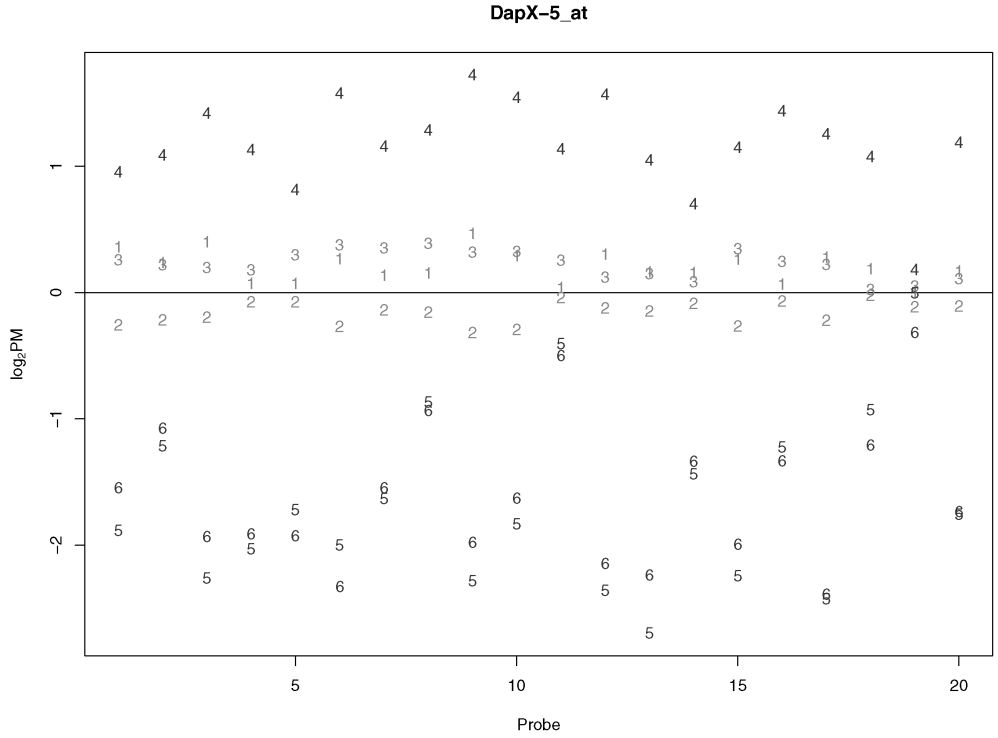

Figure 1. Expression Profiles of Gene DapX-5_at in the Spiked-in Study. The y axis is the normalized log$_2$(PM) centered by the median of six arrays within each probe. The symbols 1–3 represent the three replicate arrays from T1 and the symbols 4–6 represent the three arrays from T2.

the National Center for Biotechnology Information (accession number GSE994). The experiment was designed to study the effects of cigarette smoking on the human airway epithelial cell transcriptome. The dataset consists of 75 Affymetrix U133A human gene expression arrays, including 34 arrays with RNA samples from current smokers, 23 arrays from healthy nonsmokers, and 18 arrays from former smokers. More information on the dataset can be found in Spira et al. (2004). The number of genes analyzed is 22,204 and the number of probes is 11 for most of the genes. The data are preprocessed with the background correction and quantile normalization methods from Bioconductor's package affy. In this analysis, we focus on the genes that are found to be differentially expressed by one test, but not the other.

There are 217 genes detected by MIX but not by Q_2 and 378 genes detected by Q_2 but not by MIX. For those "controversial" genes, we examined their expression profiles and verified the probe sequences using the BLAST program from NCBI. We found that many of these cases have outlying probes or observations. Four such genes (210384_at, 201147_s_at, 208725_at, and 200654_at) are shown here for illustration. Figure 2 shows the boxplots of expression intensities with regard to different probes. The x axis represents the probes. The y axis is the $\log_2$ PM after quantile normalization. The grey boxes denote current smokers and the white boxes denote nonsmokers. The genes 210384_at and 201147_s_at are detected to be significant in the mean but not in the median, while genes 208725_at and 200654_at are found to be significant in the median but not in the mean. Table 4 summarizes the raw p values from MIX and Q_2 for these four genes before and after excluding the outlying probes or observations. The p values from the two-sample t test based on the RMA summarized values are included for reference. Note that RMA is not robust against outlying arrays, but rather is robust against a small number of outlying probes, so it is quite clear from the table that the results from RMA are more in line with those from Q_2, but the test is less powerful.

For gene 210384_at, the BLAST results show that probes 3, 6, 8, and 10 have no match to the human genome, while the rest of the probes match sequences on chromosome 21. From Figure 2, we notice the low intensities of these four probes. In addition, probes 3, 8, and 10 show smaller medians in the nonsmoker group, which is opposite to the comparisons from the other probes. This leads the test on the median to show no significance. After excluding these four suspect probes, both mean and median tests show that this gene has significantly higher intensities in nonsmokers than in smokers (see Table 4).

For gene 201147_s_at, the test MIX gives a p value .005, while the median test Q_2 gives a large p value .676. Figure 2 suggests that the significance in the mean is mostly due to quite a few outlying observations, which drive up the mean intensities of current smokers. After excluding the outlying observations (marked as circles in the boxplots), neither mean nor median tests shows significance. In this case it is unclear whether any conclusion can be made without further study.

For gene 208725_at, the BLAST results show that probes 1 and 6 have no match to any sequence on the human genome, while the other probes match sequences on chromosome 20. The two suspect probes, especially probe 6, inflate the variances to make it harder to detect changes in the mean. With these two probes removed, the p value from MIX falls from .088 to .005.

For gene 200654_at, we see an interesting phenomenon. Probes 8–11 have much lower intensities than the other probes.

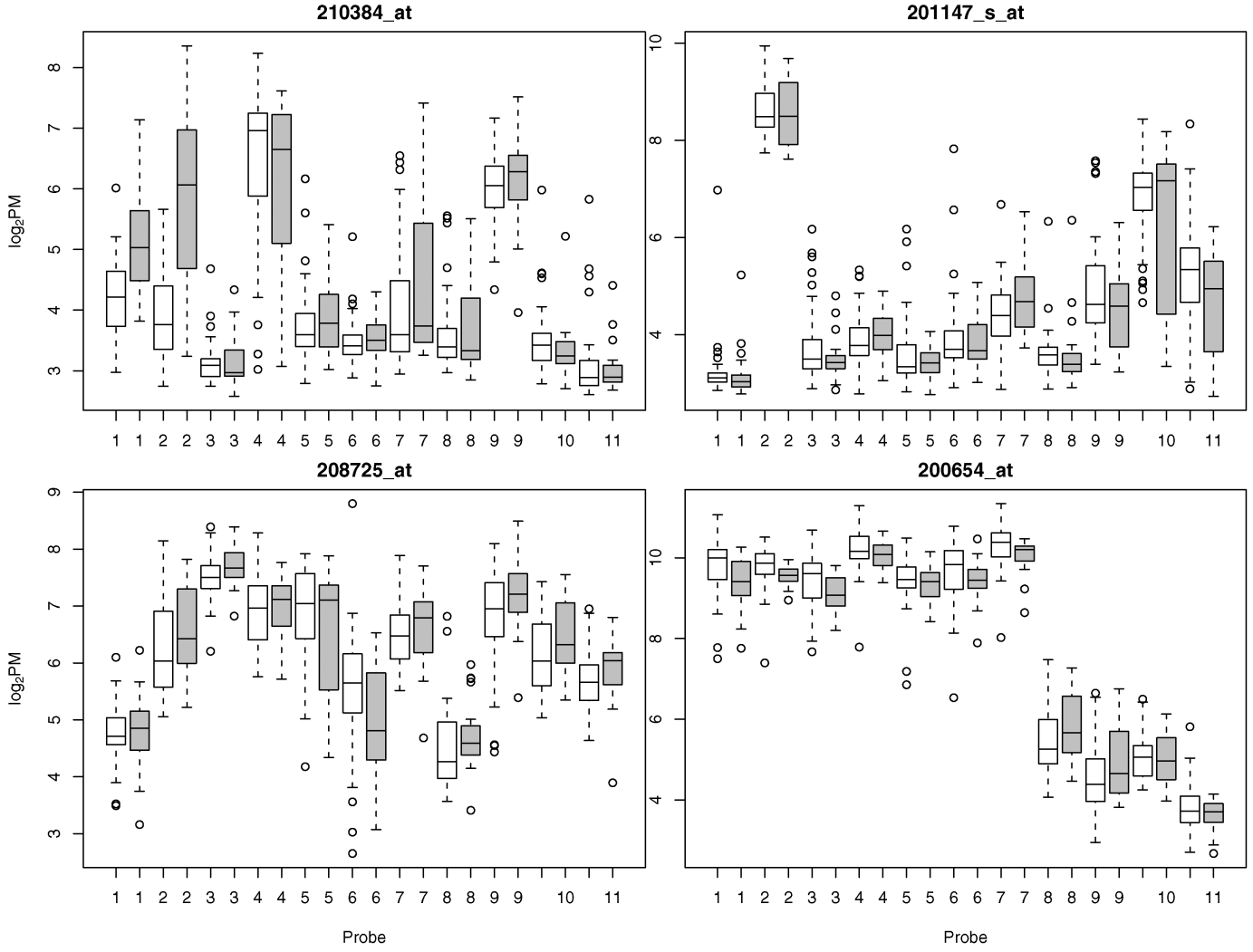

Figure 2. Expression Profiles of Four Example Genes in the Smoking Study. The x axis represents the probes. The y axis is the $\log_2$PM after quantile normalization. The grey boxes denote current smokers and the white boxes denote nonsmokers.

Furthermore, the first seven probes show higher median intensities for smokers, leading to significance in the quantile rank score test at Q_2. The differences in the means are found to be insignificant due to high between-probe variability. For this gene, alternative splicing might have occurred (see, e.g., Wu et al. 2005), but further studies are needed to explain the different gene expression profiles in probes 8–11 definitively.

The upshot of this example is that examining the differences in the quantiles in addition to the mean can help identify individual genes that cannot be summarized rightly by the means. As we know, different probes have different sensitivity and specificity, and, more importantly, problems in probe selection, cross-hybridization, probe-to-gene mapping, and alternative splicing all point to one thing, that is, it would be naive to trust the results from any test on the mean. Because of the large number of genes in microarray studies, human inspection of every single gene might, in practice, be prohibitive. The quantile approach considered in this article can be used to cross-check the results we normally obtain on the mean intensities. We have found that significant differences in the quartiles can often be detected more reliably than differences in the means. Discrepancies in the test results may flag anomaly in the probe sets used, suggesting focused validation studies on a smaller number of genes.

Table 4. The p Values Before and After Excluding the Outlying Cases for 4 Genes in the Smoking Study

		Before			After	
Gene ID	*Probes excluded*	Q_2	*MIX*	*RMA*	Q_2	*MIX*
210384_at	3, 6, 8, and 10	.033	.001	.050	.000	.000
201147_s_at	Outlying observations	.676	.005	.425	.873	.200
208725_at	1 and 6	.001	.088	.010	.000	.005
200654_at	8–11	.002	.249	.051	.003	.084

Some empirical studies of the current and earlier GeneChip data suggest that up to 20% of the probe level data might be compromised for certain genes; this supports our preference for hypothesis testing on the three quartiles. In one study, Mecham et al. (2004) pointed out that for mammalian Affymetrix microarrays, the proportion of inaccurate probes is in the 20% range. Quality improvement is being made constantly in microarray technology, but the quantile approach can be valuable even if the probe level measurements are free of technical problems. The genes of interest are often those with non-Gaussian

distributions in their intensities (on log scale), and the quantiles give more comprehensive summaries than the mean in those cases.

5. CONCLUSIONS

The quantile rank score tests are very useful in the analysis of probe level genomic data. The array effect, which is a random effect in linear models, requires an adjustment of the rank score test statistic as discussed in the present article. General inferential tools for linear quantile regression with correlated measurements are little discussed in the literature yet. In this article, we propose a simple genomewide δ adjustment to the quantile rank score test under the working assumption of an additive homogeneous random effect to the error. The resulting quantile rank score test is easy to implement and robust in performance. Our empirical studies of GeneChip data show that inference on the quartiles of the gene expression distribution often lead to more reliable results than inference on the mean. We also demonstrate that discrepancies in the significance results on the quartiles and on the mean often flag anomaly in the probe sets being considered. It is our hope that by taking the proposed quantile approach in the analysis of probe level microarray data, one can improve reliability of and confidence in the results.

In the GeneChip studies given in this article, we take the position that any apparent probe–treatment interaction in the data is a nuisance against which our comparisons of treatment should be robust. In other studies however, the interaction terms may be of interest. For example, the quantile approach will also be useful for analyzing the more recent release of the exon tiling arrays where inference needs to be made on the exon sites by treatment interaction. The proposed rank score test applies readily to those applications, but no empirical work has been done so far.

APPENDIX: ????

Following Section 2.1, we start with technical assumptions.

Assumption A.1. The function F_1 has a Lebesgue density $f_1 > 0$ with a bounded first-order derivative. The joint distribution function $F_{1,2}$ of u_{ijk_1} and u_{ijk_2} for any i, j and $k_1 \neq k_2$ is Lipschitz in a neighborhood of $(0, 0)$.

Assumption A.2. There exist $\max_{ijk} \|\widetilde{\mathbf{x}}_{ijk}\| = O(n^{1/4}/\sqrt{\log n})$ and $n^{-1}\sum_{ijk} \|\widetilde{\mathbf{x}}_{ijk}\|^3 = O(1)$ as $n \to \infty$, where $\|\widetilde{\mathbf{x}}_{ijk}\|$ denotes the Euclidean norm of $\widetilde{\mathbf{x}}_{ijk}$.

Assumption A.3. The minimum eigenvalues of $\mathbf{D}_{1n} = n^{-1}\mathbf{Z}_n^{*T}\mathbf{Z}_n^*$, $\mathbf{D}_{2n} = n^{-1}\sum_{ij}\sum_{k_1\neq k_2} \mathbf{z}_{ijk_1}^{*T}\mathbf{z}_{ijk_2}^*$, and $\mathbf{D}_{3n} = n^{-1}\mathbf{X}_n^T\mathbf{X}_n$ are bounded away from 0 as $n \to \infty$.

Assumption A.4. The set $\{K_{ij}, i = 1, \ldots, I, j = 1, \ldots, J_i\}$ is an uniformly bounded sequence of positive integers.

Note that Assumption A.2 implies that the maximum eigenvalues of $\mathbf{D}_{1n}$, $\mathbf{D}_{2n}$, and $\mathbf{D}_{3n}$ are bounded away from infinity.

The proof of Theorem 1 relies on the following three lemmas. Lemma A.1 follows easily from theorem 1 of He et al. (2002), so the proof is skipped.

Lemma A.1. Let $\hat{\boldsymbol{\alpha}}(\tau) = \arg\min_{\mathbf{a}\in R^p} \sum_{ijk} \rho_\tau(y_{ijk} - \mathbf{x}_{ijk}^T\mathbf{a})$ be a quantile estimate of $\boldsymbol{\alpha}(\tau)$. Then, under Assumptions A.2–A.4 and H_0, we have $\hat{\boldsymbol{\alpha}}(\tau) - \boldsymbol{\alpha}(\tau) = O_p(n^{-1/2})$.

Lemma A.2. Let $\mathbf{S}_n^* = n^{-1/2}\sum_{ijk} \mathbf{z}_{ijk}^*\psi_\tau(y_{ijk} - \mathbf{x}_{ijk}^T\boldsymbol{\alpha}(\tau))$. Then under Assumptions A.1–A.4 and under H_0, we have $\mathbf{S}_n = \mathbf{S}_n^* + o_p(1)$.

Proof. Consider any $\mathbf{t}$ such that $\|\mathbf{t}\| \le C(\log n)^{1/2}$ for some constant C. Let $u_{ijk}(\tau) = u_{ijk} - F_1^{-1}(\tau)$,

$$\mathbf{R}_{ij}(\mathbf{t}) = \sum_{k=1}^{K_{ij}} \mathbf{z}_{ijk}^*\big[\psi\big(u_{ijk}(\tau) - n^{-1/2}(\mathbf{x}_{ijk}^T\mathbf{t})\big) - \psi(u_{ijk}(\tau))\big], \quad \text{(A.1)}$$

and $\mathbf{r}_n(\mathbf{t}) = \sum_{ij}\mathbf{R}_{ij}(\mathbf{t})$. For each (i, j), we have

$$\begin{aligned}\operatorname{var}(\mathbf{R}_{ij}(\mathbf{t})) &\le K_{ij}\sum_{k=1}^{K_{ij}} \|\mathbf{z}_{ijk}^*\|^2 \cdot \big|P\big(u_{ijk}(\tau) < n^{-1/2}(\mathbf{x}_{ijk}^T\mathbf{t})\big) \\ &\quad - P(u_{ijk}(\tau) < 0)\big| \\ &= K_{ij}\sum_{k=1}^{K_{ij}} \|\mathbf{z}_{ijk}^*\|^2 f_1(F_1^{-1}(\widetilde{\tau}))\cdot n^{-1/2}|\mathbf{x}_{ijk}^T\mathbf{t}|, \quad \text{(A.2)}\end{aligned}$$

where $F_1^{-1}(\widetilde{\tau})$ is between $F_1^{-1}(\tau)$ and $F_1^{-1}(\tau) + n^{-1/2}(\mathbf{x}_{ijk}^T\mathbf{t})$, and the first part of Assumption A.1 is used in the last step.

Therefore, by Assumptions A.2 and A.4, we know that

$$\sum_{ij}\operatorname{var}(\mathbf{R}_{ij}(\mathbf{t})) \le c_1(n\log n)^{1/2} \quad \text{(A.3)}$$

and

$$\max_{ijk}\mathbf{R}_{ij}(\mathbf{t}) \le c_2\max_{ijk}\|\mathbf{z}_{ijk}^*\| \le c_3(n\log n)^{1/4}/\sqrt{\log n}, \quad \text{(A.4)}$$

where c_1, c_2, and c_3 are positive constants.

Thus, applying the well-known Hoeffding inequality, there is c_0 such that for $\lambda > 0$ and n large enough,

$$P\big\{\big|\mathbf{r}_n(\mathbf{t}) - E(\mathbf{r}_n(\mathbf{t}))\big| \ge c_0\lambda n^{1/4}(\log n)^{3/4}\big\} \le 2n^{-\lambda}. \quad \text{(A.5)}$$

Following a chaining argument similar to that used in the proof of Lemma A.2 in Koenker and Portnoy (1987), we can extend (A.5) uniformly in $\{\mathbf{t}: \|\mathbf{t}\| \le C(\log n)^{1/2}\}$, and get

$$\sup_{\|\mathbf{t}\|\le C(\log n)^{1/2}} \big\|\mathbf{r}_n(\mathbf{t}) - E(\mathbf{r}_n(\mathbf{t}))\big\| = O_p\big(n^{1/4}(\log n)^{3/4}\big). \quad \text{(A.6)}$$

From (A.1), we know that

$$\begin{aligned}&\sup_{\|\mathbf{t}\|\le C(\log n)^{1/2}} \big\|E(\mathbf{r}_n(\mathbf{t}))\big\| \\ &= \sup_{\|\mathbf{t}\|\le C(\log n)^{1/2}} \Big\|\sum_{ijk}\mathbf{z}_{ijk}^*\big[F_1(F_1^{-1}(\tau)) \\ &\qquad - F_1\big(F_1^{-1}(\tau) + n^{-1/2}(\mathbf{x}_{ijk}^T\mathbf{t})\big)\big]\Big\| \\ &= \sup_{\|\mathbf{t}\|\le C(\log n)^{1/2}} \Big\|\sum_{ijk}\mathbf{z}_{ijk}^*\big[f_1(F_1^{-1}(\tau))n^{-1/2}(\mathbf{x}_{ijk}^T\mathbf{t}) \\ &\qquad + f_1'(F_1^{-1}(\widetilde{\tau}))n^{-1}(\mathbf{x}_{ijk}^T\mathbf{t})^2\big]\Big\| \\ &= O(\log n), \quad \text{(A.7)}\end{aligned}$$

where the fact that $\mathbf{Z}_n^{*T}\mathbf{X} = 0$, and Assumptions A.1, A.2, and A.4 are used in the last step.

Combining (A.6) and (A.7), we have

$$\sup_{\|\mathbf{t}\|\le C(\log n)^{1/2}} n^{-1/2}\Big|\sum_{ijk}\mathbf{z}_{ijk}^*\big[\psi_\tau\big(u_{ijk}(\tau) - n^{-1/2}(\mathbf{x}_{ijk}^T\mathbf{t})\big) - \psi_\tau(u_{ijk}(\tau))\big]\Big| = o_p(1), \quad \text{(A.8)}$$

which together with Lemma A.1 completes the proof.

Lemma A.3. Under H_0 and Assumptions A.1–A.4, we have $\hat{\delta} \xrightarrow{P} \delta$ as $n \to \infty$.

Recall that $\delta = P(u_{111}(\tau) < 0, u_{112}(\tau) < 0)$. Lemma A.3 can be verified using lemma 4.1 of He and Shao (1996), which provides an uniform approximation of the sum in $\hat{\delta}$. We skip the details that are purely technical.

Proof of Theorems 1 and 2

Let

$$\mathbf{R}_{ij} = \sum_{k=1}^{K_{ij}} \mathbf{z}^*_{ijk} \psi_\tau (y_{ijk} - \mathbf{x}^T_{ijk} \boldsymbol{\alpha}(\tau)) = \sum_{k=1}^{K_{ij}} \mathbf{z}^*_{ijk} \big(\tau - I_{\{u_{ijk}(\tau)<0\}}\big),$$

where $u_{ijk}(\tau) = u_{ijk} - F_1^{-1}(\tau)$. Direct calculations show that

$$\operatorname{cov}(\mathbf{R}_{ij}) = \sum_{k=1}^{K_{ij}} \mathbf{z}^*_{ijk} \mathbf{z}^{*T}_{ijk} \tau(1-\tau) + \sum_{k_1 \neq k_2}^{K_{ij}} \mathbf{z}^*_{ijk_1} \mathbf{z}^{*T}_{ijk_2} [-\tau^2 + \delta].$$

Note that $\mathbf{S}^*_n = n^{-1/2} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \mathbf{R}_{ij}$ and that $\mathbf{R}_{ij}$ are independent entries. It follows from the Lindberg–Feller central limit theorem that

$$(\mathbf{Q}^*_n)^{-1/2} \mathbf{S}^*_n \xrightarrow{D} \mathrm{N}(\underline{\mathbf{0}}_q, \mathbf{I}_q) \tag{A.9}$$

as $n \to \infty$, where $\mathbf{Q}^*_n = n^{-1} \sum_{ijk} \mathbf{z}^*_{ijk} \mathbf{z}^{*T}_{ijk} \tau(1-\tau) + n^{-1} \times \sum_{ij} \sum_{k_1 \neq k_2}^{K_{ij}} \mathbf{z}^*_{ijk_1} \mathbf{z}^{*T}_{ijk_2} (-\tau^2 + \delta)$. The proof for Theorem 1 is therefore complete by combining (A.9), and Lemmas A.2 and A.3.

The proof of Theorem 2 is a direct extension of that of Theorem 1, where we approximate each $\mathbf{S}^{(a)}_n$ in $\mathbf{W}_n$ by $\mathbf{S}^{(a)*}_n = n^{-1/2} \sum_{ijk} \mathbf{z}^*_{ijk} \times \psi_{\tau_a}(u_{ijk}(\tau_a))$ for $1 \leq a \leq l$. For details, refer to the doctoral thesis of Wang (2006).

[Received April 2006. Revised July 2006.]

REFERENCES

Affymetrix (2001), *Statistical Algorithm Reference Guide* (Version 5), Santa Clara, CA: Affymetrix.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Chu, T., Weir, B., and Wolfinger, R. (2002), "A Systematic Statistical Linear Modeling Approach to Oligonucleotide Array Experiments," *Mathematical Biosciences*, 176, 35–51.

Fan, J., Chen, Y., Chan, H. M., Tam, P. K. H., and Ren, Y. (2005), "Removing Intensity Effects and Identifying Significant Genes for Affymetrix Arrays in MIF-Suppressed Neuroblastoma Cells," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 17751–17756.

Fan, J., Huang, T., and Peng, H. (2005), "Semilinear High-Dimensional Model for Normalization of Microarray Data: A Theoretical Analysis and Partial Consistency" (with discussion), *Journal of the American Statistical Association*, 100, 781–813.

Gutenbrunner, C., Jurêcková, J., Koenker, R., and Portnoy, S. (1993), "Tests of Linear Hypotheses Based on Regression Rank Scores," *Journal of Nonparametric Statistics*, 2, 307–333.

He, X., and Shao, Q. M. (1996), "A General Bahadur Representation of *M*-Estimators and Its Application to Linear Regression With Nonstochastic Designs," *The Annals of Statistics*, 24, 2608–2630.

He, X., Zhu, Z. Y., and Fung, W. K. (2002), "Estimation in a Semiparametric Model for Longitudinal Data With Unspecified Dependence Structure," *Biometrika*, 89, 579–590.

Hu, J., and He, X. (2006), "Enhanced Quantile Normalization of Microarray Data to Reduce Loss of Information in the Gene Expression Profile," *Biometrics*, **??**, **??–??**.

Irizarry, R. A., Hobbs, F. C. B., Beaxer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. P. (2003), "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data," *Biostatistics*, 4, 249–264.

Khuri, A. I., Mathew, T., and Sinha, B. K. (1998), *Statistical Tests for Mixed Linear Models*, New York: Wiley.

Kocherginsky, M., He, X., and Mu, Y. (2005), "Practical Confidence Intervals for Regression Quantiles," *Journal of Computational and Graphical Statistics*, 14, 41–55.

Koenker, R. (2004), "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis*, 91, 74–89.

——— (2005), *Quantile Regression*, Cambridge, U.K.: Cambridge University Press.

Koenker, R., and Bassett, G. J. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Portnoy, S. (1987), "*L*-Estimation for Linear Models," *Journal of the American Statistical Association*, 82, 851–857.

Li, C., and Wong, W. H. (2001), "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection," *Proceedings of the National Academy of Sciences of the United States of America*, 98, 31–36.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute.

Lönnstedt, I., and Speed, T. P. (2002), "Replicated Microarray Data," *Statistica Sinica*, 12, 31–46.

Mecham, B. H., Wetmore, D. Z., Szallasi, Z., Sadovsky, Y., Kohane, I., and Mariani, T. J. (2004), "Increased Measurement Accuracy for Sequence-Verified Microarray Probes," *Physiological Genomics*, 18, 308–315.

Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., and Brody, J. S. (2004), "Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome," *Proceedings of the National Academy of Sciences of the United States of America*, 101, 10143–10148.

Wang, H. (2006), "Inference on Quantile Regression for Mixed Models With Applications to GeneChip Data," unpublished Ph.D. dissertation, University of Illinois at Urbana–Champaign, IL.

Wu, C., Morris, J. S., Baggerly, K. A., Coombes, K. R., Girard, L., Minna, J. D., and Zhang, L. (2005), "A Probe-to-Transcripts Mapping Method for Microarray Data Preprocessing to Reflect Transcriptome Complexity," preprint, **???**.