Biostatistics (2009), **10**, 1, *pp.* 60–69
doi:10.1093/biostatistics/kxn015
Advance Access publication on June 6, 2008

Genomic outlier profile analysis: mixture models, null hypotheses, and nonparametric estimation

DEBASHIS GHOSH*

Department of Statistics and Department of Public Health Sciences, Pennsylvania State University, University Park, PA 16802, USA
ghoshd@psu.edu

ARUL M. CHINNAIYAN

Department of Pathology and Department of Urology, Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI 48109, USA

SUMMARY

In most analyses of large-scale genomic data sets, differential expression analysis is typically assessed by testing for differences in the mean of the distributions between 2 groups. A recent finding by Tomlins *and others* (2005) is of a different type of pattern of differential expression in which a fraction of samples in one group have overexpression relative to samples in the other group. In this work, we describe a general mixture model framework for the assessment of this type of expression, called outlier profile analysis. We start by considering the single-gene situation and establishing results on identifiability. We propose 2 nonparametric estimation procedures that have natural links to familiar multiple testing procedures. We then develop multivariate extensions of this methodology to handle genome-wide measurements. The proposed methodologies are compared using simulation studies as well as data from a prostate cancer gene expression study.

Keywords: Bonferroni correction; DNA microarray; False discovery rate; Goodness of fit; Multiple comparisons; Uniform distribution.

1. INTRODUCTION

With the advent of high-throughput gene assay technologies, scientists are now able to measure genome-wide mRNA expression levels in a variety of settings using DNA microarrays (Schena, 2000). One of the major tasks in studies involving these technologies is to find genes that are differentially expressed between 2 experimental conditions. The simplest example is to find genes that are up- or downregulated in cancerous tissue relative to healthy tissue. Typically in these experiments, the number of genes, represented as spots on the biochip, is much larger than the number of independent samples in the study. Consequently, assessing differential expression in this setting leads to performing several thousand hypothesis tests, which leads to the problem of multiple comparisons.

*To whom correspondence should be addressed.

There has been an enormous literature on statistical assessment of differential expression in genomic studies (e.g. Efron *and others*, 2001; Dudoit *and others*, 2002), along with multiple comparisons procedures for controlling a proper error rate, such as the familywise type-I error (FWER) (Shaffer, 1995) or the false discovery rate (FDR) (Benjamini and Hochberg, 1995). However, in most of these studies, differential expression is tested using a test for difference in mean expression or testing that the entire distribution functions for gene expression in the 2 conditions are the same. For the former scenario, the most commonly used procedure is the 2-sample *t*-test, while for the latter, the Wilcoxon rank sum test is used.

A more interesting differential expression pattern was observed by Tomlins *and others* (2005). They noticed that for certain genes, only a fraction of samples in one group were overexpressed relative to those in the other group; the remaining samples showed no evidence of differential expression. Tomlins *and others* (2005) developed a ranking method known as cancer outlier profile analysis (COPA) for calculating outlier scores using gene expression data. Their score was purely descriptive; they did not attempt to assign any measure of significance to the gene scores. More recently, Tibshirani and Hastie (2007) and Wu (2007) have shown that significance can be assigned using modifications of 2-sample *t*-tests. In addition, a nonparametric methodology proposed by Lyons-Weiler *and others* (2004) could be applied to this problem as well. We discuss all these proposals in Section 3.2.

We should mention that while the term “outlier’’ has a pejorative meaning in statistics, it is a very meaningful concept in a biological sense. As noted by Lyons-Weiler *and others* (2004) and subsequently by Tomlins *and others* (2005), the biology of oncogenesis permits that unique sets of genes may be involved in tumor development across patients. While statistical outliers refer to measurements that exceed the expected variation in a set of data, the oncogenetic outliers we seek to find will be putatively related to cancer processes.

The goal of this article is to describe a relatively general statistical model for the outlier approach of Tomlins *and others* (2005). By formulating the probabilistic model, we can clarify various issues in outlier profile analysis that have not been previously addressed and better situate the proposals of prior authors. In particular, their proposals are parametric in nature; we come up with alternative nonparametric procedures for outlier analysis with genomic data. As a by-product of our methods, we link multiple testing procedures with outlier detection. The paper is structured as follows: In Section 2, we describe the data setup and formulate the statistical model for outlier profile analysis in the case of a single gene. Doing this allows us to establish results about identifiability as well as develop a sample-specific hypothesis of interest. We also develop the proposed nonparametric estimation procedure and link it with multiple testing methodology. In Section 3, we describe the general procedure with genome-wide expression data sets and relate the prior proposals in the literature. In Section 4, we describe application of the proposed methodology to simulated data. Finally, we conclude with some discussion in Section 5.

2. Outlier profile analysis: single-gene case

2.1 *Data, inference, and proposed methodology*

The data consist of (Y_{gi}, Z_i) , where Y_{gi} is the gene expression measurement on the gth gene for the ith subject and Z_i is a binary indicator taking values 0 and 1, $g = 1, \ldots, m$, $i = 1, \ldots, n$. We will refer to the group with $Z = 0$ as nondiseased samples and $Z = 1$ as diseased samples. We will use the notation $\mathbf{Y}_{g\cdot}$ to denote the gene expression profile of the gth gene across all subjects and $\mathbf{Y}_{\cdot i}$ to represent the m-dimensional expression profile for the ith individual. We will assume that there are n_0 samples with $Z = 0$ and n_1 samples with $Z = 1$ so $n = n_0 + n_1$. Without loss of generality, we will assume that the first n_0 samples come from the undiseased samples.

We first consider a simple situation of $G = 1$ gene. Then, a simple model for modeling $Y_i \equiv Y_{gi}$ conditional on Z_i is the following:

$$
\begin{aligned}
Y_i|Z_i = 0 &\overset{\text{iid}}{\sim} F_0(y),\\
Y_i|Z_i = 1 &\overset{\text{ind}}{\sim} \pi_0 F_0(y) + (1-\pi_0)F_{1i}(y),
\end{aligned} \tag{2.1}
$$

where $F_0(y)$ is an unspecified distribution function, π_0 is a fraction of samples that do not differentially express the gene between the 2 groups, and F_{1i} is a family of distribution functions. The following lemma provides conditions under which such a scenario can be tested.

LEMMA 2.1

(a) If $\pi_0 \neq 1$ and $F_{1i} \neq F_0$, then model (2.1) can be tested given the observed data.

(b) If $Z_1, \ldots, Z_n$ are not observed, then model (2.1) is not identifiable based on $Y_1, \ldots, Y_n$.

Proof. The proof of (a) follows from arguments in Section 3.1 of Genovese and Wasserman (2004). For (b), we will not be able to distinguish between F_0 and F_{1i} without information on Z. □

There are several points we wish to make at this stage. First, we have statistical independence because the model is for 1 gene across the n samples and the samples are independent. Second, it is obvious that if $\pi_0 = 0$ and F_{1i} does not depend on i, then we are reduced to a usual 2-sample problem. For that scenario, a common hypothesis to test is $H_0^* : F_0 = F_1$. Third, and perhaps, most importantly, model (2.1) is also a model for outliers in that those observations with $Z_i = 1$ that come from F_{1i} represent the outliers. For the given gene, one can thus potentially test the hypothesis H_{0i}: the ith sample ($i = 1, \ldots, n$) is not an outlier versus H_{1i} : the ith sample is an outlier. We can actually test for a more specific hypothesis than has been discussed previously in the literature on outlier profile analysis, namely that for a given gene and given sample, the sample represents an outlier. Furthermore, the only assumption we need on F_{1i} is that it does not equal F_0. Note that the hypothesis being described here is more focused than that tested by Tibshirani and Hastie (2007) and Wu (2007). We will return to discussion of the hypothesis they test in Section 3.

We now develop the proposed procedure for our situation. At the first stage, we estimate F_0 using the gene expression measurements with $Z_i = 0$. This yields an empirical distribution function $\hat{F}_0(y) = (n_0)^{-1}\sum_{i=1}^{n} I(Y_i \leqslant y, Z_i = 0)$. Next, we transform the gene expression measurements with $Z_i = 1$ using $\hat{F}_0$, which generates new variables $\hat{U}_i = 1 - \hat{F}_0(Y_i)$, $i = n_0 + 1, \ldots, n$. If F_0 were known, then for $i = 1, \ldots, n_1$,

$$U_i \overset{\text{iid}}{\sim} \pi_0 F_U(u) + (1-\pi_0)F_{Wi}(u),$$

where $F_U(u) = u$ is the cumulative distribution function (cdf) of a uniform(0,1) distribution and $F_{Wi}(u) = F_0 \circ \{F_{1i}^{-1}(u)\}$. We propose 2 algorithms for selecting outliers. Here is the first, referred to as the Bonferroni algorithm:

1. Set an error level α.
2. Reject H_{0i} for H_{1i} for the ith sample (i.e. declare the ith sample to be an outlier) if and only if

$$\hat{U}_i \leqslant \alpha/n_1.$$

We call this the Bonferroni algorithm because the rule in Step 2 of the algorithm is very similar to the Bonferroni correction for p-values in multiple testing. Here, the number of tests being performed is equal to the number of diseased samples in the data set. This is why we adjust the significance level by n_1 in Step 2.

The second algorithm we propose is to use the Benjamini–Hochberg (BH) (1995) algorithm for outlier detection. It proceeds by first sorting the $\hat{U}_i$s in increasing order, $\hat{U}_{(1)} \leqslant \hat{U}_{(2)} \leqslant \cdots \leqslant \hat{U}_{(n_1)}$, and then selecting outliers using the following 2-step algorithm:

1. Set an error rate α.
2. Take as outliers $\hat{U}_{(1)}, \dots, \hat{U}_{(\hat{k})}$, where $\hat{k} = \max\{1 \leqslant i \leqslant n_1 : \hat{U}_i \leqslant i\alpha/n_1\}$. If no such $\hat{k}$ exists, conclude that there are no outliers.

We have been assuming that $F_{1i}(y) \leqslant F_0(y)\ \forall\, y$ in (2.1). More generally, we could allow $F_{1i}(y) \neq F_0(y)$. However, then we would have to look for outliers that have both small and large values of U_i. Observe that $F_{1i}(y) \leqslant F_0(y)$ in (2.1) corresponds to gene expression being stochastically larger in diseased samples relative to nondiseased samples and $F_{1i}(y) > F_0(y)$ the opposite is true. In practice, we recommend running the procedure twice, one assuming that $F_{1i}(y) \leqslant F_0(y)$ to find outlying samples with overexpressed genes, the second time assuming the opposite.

2.2 *Outlier detection and multiple testing*

The outlier detection algorithms we have proposed have a very natural connection with multiple testing procedures. Since we can use (2.1) as a model for outliers, we can decide whether or not each sample is an outlier using a hypothesis test; this yields a total of n_1 tests of hypotheses. We can then cross-classify samples into the table based on their "true'' outlier status versus what we declare: this is shown in Table 1. Note that in the table, the only observed quantities are (n_1, R, Q). Everything else about the table is unobserved.

There is a direct correspondence between Table 1 and testing multiple hypotheses. Based on the table, we can construct appropriate error measures to control. By an error, we mean that we declare a sample to be an outlier when it is not an outlier in truth. Two popular error measures to control are the FWER (Shaffer, 1995) and the FDR (Benjamini and Hochberg, 1995). In words, the FWER is the probability of making at least 1 false declaration of a sample being an outlier, while the FDR is the average number of false outliers among the samples declared to be outliers. Using the notation of Table 1, FWER equals $\Pr(X \geqslant 1)$, while the FDR is $E[X/R|R > 0]\Pr(R > 0)$. Assume that F_0 is known. It is easy to then show the following results:

(a) The Bonferroni algorithm controls FWER and FDR at level α.
(b) The BH algorithm controls FDR at level α.

These are exact results for finite samples; one can invoke the theoretical results of Genovese and Wasserman (2004) in order to study the asymptotic properties of the parameter estimators in the model. It becomes more difficult to prove results about error control with the proposed procedure because it involves $\hat{F}_0$ rather than F_0. Since we normalize the gene expression measurements for the $Z = 1$ group by $\hat{F}_0$, the transformed observations are not statistically independent. Using the notation of Genovese and Wasserman, define T as a mapping from $[0, 1]^{n_1}$ into $[0, 1]$; we can then define the Bonferroni and BH

Table 1. *Outcomes of n_1 tests of hypotheses regarding outlying samples*

	Decide outlier	Decide nonoutlier	Total
True nonoutlier	X	V	n_{10}
True outlier	B	A	n_{11}
	R	Q	n_1

Note: The rows represent each sample being a true outlier or a true nonoutlier. In the columns, decide outlier means that we reject H_{0i} and decide nonoutlier means that we fail to reject H_{0i}.

algorithms as

$$T_{\mathrm{B}}(\hat{\mathbf{U}}) = \alpha/n_1$$

and

$$T_{\mathrm{BH}}(\hat{\mathbf{U}}) = \sup\{t : \hat{M}_{n_1}(t) = t/\alpha\},$$

where $\hat{\mathbf{U}} = (\hat{\mathbf{U}}_1, \ldots, \hat{U}_{n_1})$ and $\hat{M}_{n_1}(t) = (n_1)^{-1} \sum_{i=1}^{n_1} I(\hat{U}_i \leqslant t)$. However, the results of Genovese and Wasserman (2004) do not directly apply to this problem because of the dependence in the transformed observations $\hat{\mathbf{U}}$. Assume that the densities corresponding to F_0 and F_1, f_0 and f_1, are continuous and that f_1 is strictly positive on $\{y : 0 < F_1(y) < 1\}$. Then, this is sufficient to guarantee the convergence of $\hat{F}_0$ to F_0; by the continuous mapping theorem (Van der Vaart and Wellner, 1996), this implies that the Bonferroni procedure will asymptotically control the FWER. For the BH procedure, we make the additional 2 assumptions. First, we assume that π is identifiable; conditions guaranteeing this are given in Genovese and Wasserman (2004, Section 3.1). Second, we assume that the range of $F_0 \circ F_1^{-1}$ is [0, 1]. This guarantees the uniform convergence of $\hat{M}_{n_1}(t)$ to its population limit. By another application of the argmax continuous mapping theorem, we have that T_{BH} controls the FDR.

3. Outlier profile analysis: genome-wide case

3.1 *Multivariate extensions: model and methodology*

Now, we wish to consider the outlier profile analysis problem for genome-scale data such as data generated by a gene expression microarray experiment. Then, model (2.1) becomes the following:

$$Y_{gi}|Z_i = 0 \overset{\text{ind}}{\sim} F_{0g}(y),$$

$$Y_{gi}|Z_i = 1 \overset{\text{ind}}{\sim} \pi_{0g}F_{0g}(y) + (1 - \pi_{0g})F_{1gi}(y), \tag{3.1}$$

$$\pi_{01}, \ldots, \pi_{0G} \sim \mathcal{P}, \tag{3.2}$$

where $\mathcal{P}$ is an arbitrary distribution function. Note that we are leaving the structure of F_{0g} and F_{1gi} unspecified. We will return to this point later in Section 3.3. Now, suppose that in (3.2), the π_{0g} ($g = 1, \ldots, G$) are a mixture themselves of a point mass at 1 and alternative distribution function so that

$$\pi_{01}, \ldots, \pi_{0G} \overset{\text{iid}}{\sim} p\delta_1 + (1 - p)F_P(\pi). \tag{3.3}$$

Now, if π_{0g} comes from the first mixture component, then there is no differential expression for the gth gene. To make the model sensible, we need that smaller values of π_{0g} correspond to an increased likelihood of coming from the distribution function F_P.

What most previous authors have tested within this model (Lyons-Weiler *and others*, 2004; Tibshirani and Hastie, 2007; Wu, 2007) is $H_{0g} : \pi_{0g} = 1, g = 1, \ldots, G$. In contrast to the hypothesis described in Section 2, which is a sample-specific hypothesis involving outliers, the null hypothesis H_{0g} here is a gene-specific one. When we think about assessing significance now, any multiple testing adjustment needs to account for the multiplicity of genes in the study and not the number of samples.

Our approach is to have the following class of scores:

$$S_g = \sum_{i=1}^{n} W_{gi} I\{D_i = 1, \hat{U}_{gi} \leqslant c_{i,\alpha}^{g}\}, \tag{3.4}$$

where $\hat{U}_{gi}$ is the gene-specific analog of $\hat{U}_i$ from Section 2.2, W_{gi} is a weight function, and $c_{i,\alpha}^{g}$ is a critical value depending on the particular procedure being used (Bonferroni 1, Bonferroni 2, or BH). Two natural choices for W are $W_{gi} = 1$ and $W_{gi} = Y_{gi}$. Using the first weight function will make the statistic S_g fairly discrete, while using the second weight function will make the statistic S_g be more continuous. In this paper, we use the second one.

To derive the null distribution of (3.4), we permute the class labels (Z) between the cases and the controls. During this permutation, we recalculate $\hat{F}_0$. Based on this, we can then perform the usual multiple testing adjustments controlling either the FWER or the FDR. Note that for this situation, we must adjust for the number of genes since the number of hypotheses being tested is equal to the number of genes on the microarray. Based on the permutations, we can then adjust the p-values for multiple testing. A variety of procedures for doing so based on FWER can be found in Dudoit *and others* (2002). For presenting scientists with a list of genes calibrated in evidence for outlierness, we use the q-value approach of Storey and Tibshirani (2003). In words, the q-value is approximately the smallest FDR at which we would reject the null hypothesis that there is no outlying expression for the gth gene in diseased relative to nondiseased samples. We then rank genes based on the q-value.

3.2 *Comparison with previous methods*

It is instructive to consider the difference between the proposed methodology versus those previous authors have constructed. We first start with the approach of Tomlins *and others* (2005). They standardize the data across all samples and create new measurements

$$Y^*_{gi} = \frac{\{Y_{gi} - \text{median}(\mathbf{Y}_{g\cdot})\}}{\text{MAD}(\mathbf{Y}_{g\cdot})},$$

where median($\mathbf{Y}$) is the median of the vector $\mathbf{Y}$ and MAD($\mathbf{Y}$) denotes the median absolute deviation of $\mathbf{Y}$. The COPA score of Tomlins *and others* (2005) for the gth gene is the following:

$$\text{COPA}_g = q_r(\mathbf{Y}^*_{g\cdot}), \tag{3.5}$$

where $q_r(\mathbf{Y})$ denotes the rth percentile of the vector $\mathbf{Y}$. In words, Tomlins *and others* (2005) use the rth percentile of $Y^*_{g\cdot}$ for all samples; in practice, they consider r to be 75, 90, and 99. What Tibshirani and Hastie (2007) propose is a modified t-test for finding this pattern; they use as their statistic

$$\text{OS}_g = \sum_{i=1}^{n} Y^*_{gi} I\{Z_i = 1, Y^*_{gi} > q_{75}(\mathbf{Y}^*_{g\cdot}) + \text{IQR}(\mathbf{Y}^*_{g\cdot})\}, \tag{3.6}$$

where IQR($\mathbf{Y}$) denotes the interquartile range of a vector $\mathbf{Y}$. Tibshirani and Hastie (2007) argue that the outlier sum (OS) score (3.6) is potentially more efficient than the COPA score because it sums over all outlying disease samples.

Wu (2007) develops an approach called the outlier robust t-statistic (ORT). He seeks to separate the diseased and undiseased populations as much as possible because he argues that it is possible for the distributions of gene expression measurements for the 2 groups to be different. His statistic is

$$\text{ORT}_g = \frac{\sum_{i=1}^{n} Y_{gi} I\left\{Z_i = 1, Y^*_{gi} > q_{75}\left(Y_{g1}, \ldots, Y_{gn_0}\right) + \text{IQR}\left(Y_{g1}, \ldots, Y_{gn_0}\right)\right\}}{\text{median}\left(\text{median}_{1\leqslant i\leqslant n_0}\left|Y_{gi} - \text{median}_{1\leqslant i\leqslant n_0} Y_{gi}\right|, \text{median}_{n_0+1\leqslant i\leqslant n}\left|Y_{gi} - \text{median}_{n_0+1\leqslant i\leqslant n} Y_{gi}\right|\right)}. \tag{3.7}$$

There is a procedure proposed by Lyons-Weiler *and others* (2004), called the permutation percentile separability test (PPST), that could also be applied to this problem. Their statistic is

$$\text{PPST}_g = \sum_{i=1}^{n} Y_{gi} I\left\{Z_i = 1, Y_{gi} > q_{95}\left(Y_{g1}, \ldots, Y_{gn_0}\right)\right\}. \tag{3.8}$$

For the last 3 formulae (3.6–3.8), the authors derive null distributions using permutation of the nondiseased and diseased samples. In comparing (3.5–3.8), we highlight several points. First, the original COPA measure (3.5) of Tomlins *and others* (2005) did not attempt to ascribe any measure of significance and defines the threshold based on all samples. The other approaches all attempt to use more statistically motivated criteria for assessing outliers, and they all sum over all n_1 samples in the diseased population. The OS approach (3.6) uses all samples for ranking, but the other 2 approaches (3.7) and (3.8) only construct a cutoff using the samples in the nondiseased category. The latter 3 approaches (3.6–3.8) all test the hypothesis $H_{0g} : \pi_g = 1$.

Our approach in (3.4) differs in one major respect from the scores (3.6–3.8). We seek to control an error rate measure, which none of the other proposals do. This has the effect of creating a data-dependent threshold that incorporates variability in a more flexible way than by use of the interquartile range such as in (3.6) and (3.7).

4. NUMERICAL EXAMPLES

4.1 *Simulation studies*

To assess the performance of the methodology, we first conducted some simulation studies. In particular, we generated gene expression measurements for 1000 genes and allowed for 50 genes to have a differential expression pattern different between 2 groups, each with $n = 20$ samples. We considered differential expression in $k = 5, 10$, and 15 samples. For each simulation scenario, 100 data sets were generated. We took the baseline distribution of gene expression to be exponential with mean 1 and the differential expression to be exponential with mean 2. We compared the performance of the proposed methodology to the methods discussed in Section 3.2 using receiver operating characteristic (ROC) curves, averaged over the simulations. In terms of performance, ROC curves close to the diagonal indicate poor performance, while those closer to the upper left-hand corner indicate better performance. For the Bonferroni and BH methods, we took $\alpha = 0.05$. The simulation results are indicated in Figure 1. We did not use the original COPA method of Tomlins *and others* (2005).

Based on the curves, we find that for small values of the false-positive rate, the proposed methodology using the BH procedure performs the best among all methods, while that using the OS method performs the worst. One point of note is that the PPST method (3.8), which has not been previously explored in the literature, tends to perform better than the OS and the outlier robust t-statistic and actually does better as k increases. The proposed methods are always competitive in these situations.

We next performed a simulation that mimicked a setup used by Wu (2007). We took the baseline distribution of genes to be standard normal with mean zero and variance one; in a fraction of samples, 50 genes had a normal distribution with mean 2 and variance 1. The simulation results are given in Figure 2. For small k, the BH procedure tends to perform the best, while for larger k, the PPST and proposed Bonferroni methods tend to perform much better.

4.2 *Prostate cancer data set*

We now apply the proposed methodology to data from a gene expression study in prostate cancer. There is a total of 101 samples in the study: 22 noncancerous samples and 79 cancerous samples; the samples were profiled using 2-color (red/green) microarrays. There were a total of 9984 genes on the original microarray; the following preprocessing steps were applied before using the methodology:

1. Genes with more than 50% missing values across all samples were removed from the study.
2. Missing values were imputed using a nearest neighbors algorithm (Troyanskaya *and others*, 2001), where the number of nearest neighbors is set to 10.

This left a total of 9272 genes for analysis.

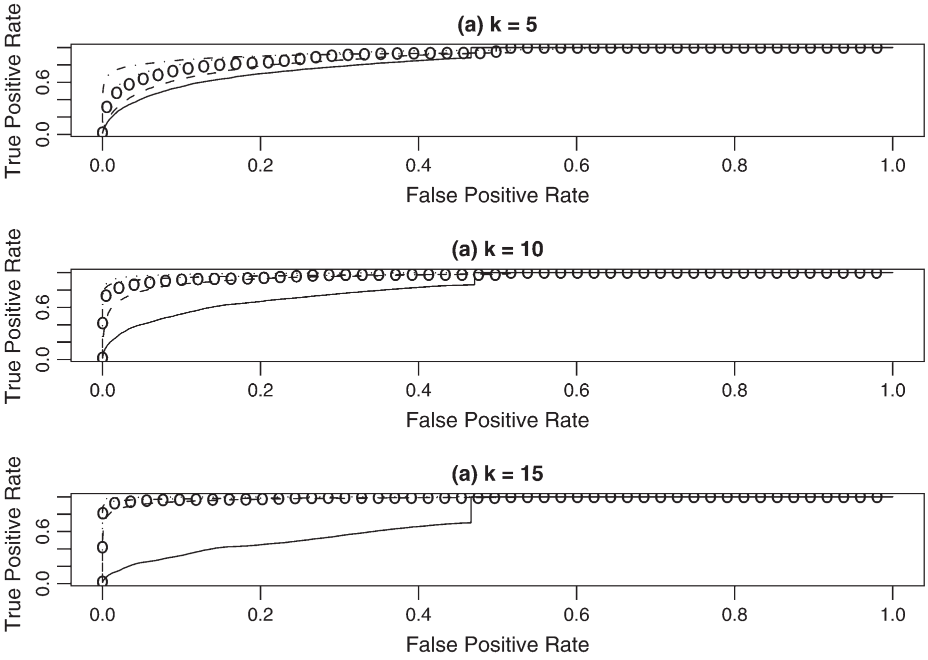

Fig. 1. Average ROC curves of various outlier detection procedures using first simulation scenario. Solid line indicates OS method of Tibshirani and Hastie (2007). Dotted line indicates percentile-specific method (PPST) of Lyons-Weiler *and others* (2004). Dashed line indicates outlier robust t-statistic of Wu (2007). Circles indicate proposed Bonferroni method. Dashed and dotted line indicates proposed BH method for outlier detection.

As discussed in Section 2, one of the advantages of the proposed methodology here is that we can determine which are specific samples that show evidence of outlying expression with respect to a particular gene. As an example, we take ERG (v-ets erythroblastosis virus E26 oncogene homolog), which was found to be part of a gene fusion product that appears to be quite common in prostate cancer (Tomlins *and others*, 2005). If we apply the methods from Section 2 to ERG, we find that there are 40 samples that show evidence of outlying expression in the cancerous samples relative to the noncancerous samples using the Bonferroni method with $\alpha = 0.05$. We get the same answer using the BH procedure with $\alpha = 0.05$; in fact, the set of samples called outliers is the same using either method. When we apply the method, switching the 2 groups of samples, there are no samples in the noncancerous group that show evidence of outlying expression relative to the cancerous samples.

Next, we applied the methods from Section 3 in order to do a more global search of genes that show evidence for outlying expression. Here, we only focus on genes that show evidence of overexpression in the cancerous samples relative to the noncancerous samples. A comparison of correlation between the ranks of the genes based on the outlier score methods is given in Table 2. Based on Table 2, we find that the proposed methods give highly concordant results. There is less concordance with the other 3 methods.

Next, we performed the q-value analysis of Storey and Tibshirani (2003). This was performed after calculating the permutation distribution using 10 000 samples. Interestingly, while the estimated π_0 was one using the PPST, ORT, and OS methods, it was 0.19 for the Bonferroni procedure and 0.29 for the BH procedure. This leads to many more genes being called significantly differentially expressed using the latter 2 procedures versus the existing methods at any q-value cutoff.

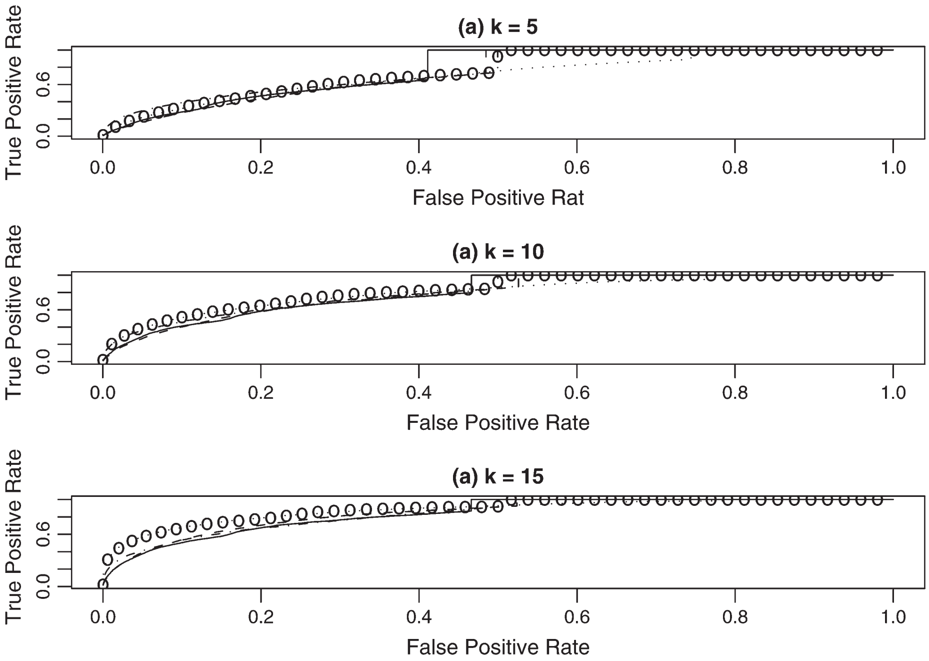

Fig. 2. Average ROC curves of various outlier detection procedures in second simulation scenario. See Figure 1 for methods corresponding to different symbols.

5. Discussion

In this article, we have placed a very formal statistical framework for outlier detection using genomic data. By formulating the problem using mixture models, we are able to clarify what hypotheses can be tested. Doing this also allows us to clarify the statistical contributions of previous work on this subject.

Another theme in this work is the relative utility of nonparametric methods. While much of the previous literature on outlier detection has used modified t-statistics, the empirical cdf-based approach proposed here tends to give very good performance in the simulation settings considered. While the t-statistic methods will be powerful in cases where the data are Gaussian, they will be less so in non-Gaussian settings. By contrast, the performance of the proposed nonparametric methods will be more robust to the choice of the data-generating mechanism.

One of the other facts noted by Tomlins *and others* (2005) was that there was a particular expression pattern to the ERG–ETV1 gene pair. In a fraction of samples, one of these genes would be overexpressed,

Table 2. *Correlation between outlier scores using methods in Section* 3

	PPST	BONF	ORT	OS	BH
PPST	1.00	0.78	0.71	0.52	0.76
BONF		1.00	0.58	0.50	0.99
ORT			1.00	0.61	0.57
OS				1.00	0.45
BH					1.00

Note: Correlation in Table 2 calculated using Spearman’s ρ. BONF, Bonferroni.

while the other would not show any expression. This also suggests another type of gene expression pattern to search for; it is bivariate in nature. While a threshold-based method for assessing significance was proposed by MacDonald and Ghosh (2006), it would be desirable to extend the approach here to that problem as well. This is currently under investigation.

Acknowledgments

Conflict of Interest: None declared.

Funding

The Huck Institute of Life Sciences and the National Institutes of Health (R01-GM72007).

References

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 Series B* **57**, 289–300.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* **12**, 111–140.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* **32**, 1035–1061.

Lyons-Weiler, J., Patel, S., Becich, M. J. and Godfrey, T. E. (2004). Tests for finding complex patterns of differential expression in cancers: towards individualized medicine. *BMC Bioinformatics* **125**, 110.

MacDonald, J. W. and Ghosh, D. (2006). COPA–cancer outlier profile analysis. *Bioinformatics* **22**, 2950–2951.

Schena, M. (2000). *Microarray Biochip Technology*. Sunnyvale, CA: Eaton.

Shaffer, J. (1995). Multiple hypothesis testing. *Annual Reviews of Psychology* **46**, 561–584.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

Tibshirani, R. and Hastie, T. (2007). Outlier sums for differential gene expression analysis. *Biostatistics* **8**, 2–8.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R. *and others* (2005). Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* New York: Springer.

Wu, B. (2007). Cancer outlier differential gene expression detection. *Biostatistics* **8**, 566–575.

[*Received December 13, 2007; revised April 3, 2008; accepted for publication April 29, 2008*]