Biometrika (2010), **97**, 1, *pp*. 246–253
doi: 10.1093/biomet/asp075
Advance Access publication 25 January 2010

The distribution-based *p*-value for the outlier sum in differential gene expression analysis

BY LIN-AN CHEN

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan
lachen@stat.nctu.edu.tw

DUNG-TSA CHEN

Biostatistics Department, Moffitt Cancer & Research Institute, Department of Oncologic Sciences, University of South Florida, Tampa, Florida 33612, U.S.A.
dung-tsa.chen@moffitt.org

AND WENYAW CHAN

Division of Biostatistics, The University of Texas, Houston, Texas 77225, U.S.A.
wenyaw.chan@uth.tmc.edu

SUMMARY

Outlier sums were proposed by Tibshirani & Hastie (2007) and Wu (2007) for detecting outlier genes where only a small subset of disease samples shows unusually high gene expression, but they did not develop their distributional properties and formal statistical inference. In this study, a new outlier sum for detection of outlier genes is proposed, its asymptotic distribution theory is developed, and the *p*-value based on this outlier sum is formulated. Its analytic form is derived on the basis of the large-sample theory. We compare the proposed method with existing outlier sum methods by power comparisons. Our method is applied to DNA microarray data from samples of primary breast tumors examined by Huang et al. (2003). The results show that the proposed method is more efficient in detecting outlier genes.

Some key words: Asymptotic distribution; Cancer outlier profile analysis; Gene expression; Outlier robust *t*-statistic; Outlier sum; *p*-value; *t*-test.

1. INTRODUCTION

DNA microarray technology, which simultaneously probes thousands of gene expression profiles, has been successfully used in medical research for disease classification (Agrawal et al., 2002; Ohki et al., 2005). For example, in breast cancer research, Sorlie et al. (2003) used gene expression to classify malignant breast tumors into five distinct molecular subtypes. In lymphoma research, Alizadeh et al. (2000) reported that patients with one type of molecular pattern, germinal center B-like diffuse large B-cell lymphoma, had a significantly better chance of overall survival than those with another molecular pattern, activated B-like diffuse large B-cell lymphoma. Recently, investigators have extended the disease classification capability of microarray analysis to identifying outlier genes that are overexpressed only in a small number of disease samples. Tomlins et al. (2005) introduced an approach called cancer outlier profile analysis to identify outlier genes. This approach standardizes gene expression by centring at the median and scaling by the median absolute deviation. A kth percentile of the standardized expression value is then used as a cut-off point to determine an outlier gene. Later, two outlier sum approaches, the outlier sum statistic (Tibshirani & Hastie, 2007) and the outlier robust t-statistic (Wu, 2007), were developed to improve detection of outlier genes. Both approaches use the summation of extreme high expression

values, instead of the kth percentile of the standardized gene expression, to identify outlier genes. Empirical studies have shown that both outlier sum approaches are more powerful in detecting outlier genes than standard approaches, such as the t-test. However, the development of distribution theory for these statistics is still primitive. In this paper, we derive an explicit form for the p-value of an outlier sum test statistic. The approach is empirically proven to be generally more powerful than the other methods.

2. The new outlier sum and its p-value

In a study that consists of n_1 subjects in the normal control group and n_2 subjects in the disease group, suppose that there are m genes to be investigated. Their gene expression can be represented as $X_{ij}(i = 1, \ldots, n_1, j = 1, \ldots, m)$ for the normal control group and $Y_{ij}(i = 1, \ldots, n_2,\ j = 1, \ldots, m)$ for the disease group. For a fixed gene j, let μ_j represent the parameter for central tendency that is used for measuring distance for observations in the disease group, let η_j represent the cut-off point that is used for identifying an observation in the disease group as an outlier, and let σ_j represent the scale parameter that is used for standardizing the sum of outliers. Let $\hat{\mu}_j$, $\hat{\eta}_j$ and $\hat{\sigma}_j$ denote their corresponding estimators.

A standardized version of the outlier sum statistic for gene j defined by Tibshirani & Hastie (2007) and Wu (2007) may be represented as $\sum_{i=1}^{n_2} \hat{\sigma}_j^{-1}(Y_{ij} - \hat{\mu}_j)I(Y_{ij} > \hat{\eta}_j)$. The former uses combined observations of X_{ij}s and Y_{ij}s for estimation of μ_j, η_j and σ_j and the latter uses only X_{ij}s for estimation of the same set of parameters. In this paper, we present a nonstandardized outlier sum formulated in (1) and use its mean statistic to derive the p-value based on the large-sample theory of this statistic. In the following, we drop the gene index j from all mathematical expressions unless otherwise noted. The outlier sum and its mean statistic for a gene are, respectively, formulated as

$$L = \sum_{i=1}^{n_2} Y_i I(Y_i > \hat{\eta}), \quad \bar{L} = \frac{1}{\sum_{i=1}^{n_2} I(Y_i > \hat{\eta})} \sum_{i=1}^{n_2} Y_i I(Y_i > \hat{\eta}). \tag{1}$$

Let F_X and F_Y denote the distribution functions, respectively, for variables X and Y. For testing $H_0 : F_Y = F_X$, we will further assume that there are the unknown location and scale parameters for this outlier mean, denoted by μ_ℓ and σ_ℓ, respectively, and they satisfy

$$\mathrm{pr}_{H_0}\left\{\frac{n_2^{1/2}(\bar{L} - \mu_\ell)}{\sigma_\ell} \leqslant z\right\} \to \Phi(z), \quad n_2 \to \infty, \tag{2}$$

for any real number z, where Φ represents the cumulative distribution function of a standard normal variate. We also set two constraints on the cut-off point $\hat{\eta}$ and the sample sizes n_1 and n_2, stated in Assumptions 1 and 2 in the Appendix.

From the relationship between the outlier mean and the outlier sum in (1) and from Assumption 1 for the proportion of the outlier samples, we can infer from Slutsky's theorem and equation (2) that $(n_2^{1/2}\beta\sigma_\ell)^{-1}(L - n_2\beta\mu_\ell)$ has an approximate $N(0, 1)$ distribution. This provides a natural candidate for the test statistic based on the outlier sum statistic:

$$Z_{\text{test}} = \frac{L - n_2\hat{\beta}\hat{\mu}_\ell}{n_2^{1/2}\hat{\beta}\hat{\sigma}_\ell}, \tag{3}$$

where $\hat{\beta}$, $\hat{\mu}_\ell$ and $\hat{\sigma}_\ell$ are estimators for β, μ_ℓ and σ_ℓ, respectively.

The p-value for the outlier sum can be expressed as

$$p = 1 - \Phi(z_{\text{test}}), \tag{4}$$

where z_{test} is the sample realization of Z_{test} and estimates $\hat{\beta}$, $\hat{\mu}_\ell$ and $\hat{\sigma}_\ell$ are computed based on the observations x_is from the normal group.

3. ASYMPTOTIC PROPERTIES OF THE OUTLIER MEAN

In a study of the outlier mean or sum, the cut-off point is usually chosen to be a certain multiple of the interquartile range away from the median. For example, in the outlier robust t-statistic approach, one interquartile range away from the median is adopted as the cut-off point. In the proposed method, we allow a flexible selection of the cut-off point and express it as $\eta = F_X^{-1}(0{\cdot}5) + 1{\cdot}5k\{F_X^{-1}(0{\cdot}75) - F_X^{-1}(0{\cdot}25)\}$ for mathematical convenience, where $k > 0$ is a constant. To develop the p-value based on large sample theory, we suggest use of $\hat{\eta} = \hat{F}_X^{-1}(0{\cdot}5) + 1{\cdot}5k \times \mathrm{IQR}_X$ as the cut-off point for estimating η, where $\mathrm{IQR}_X = \hat{F}_X^{-1}(0{\cdot}75) - \hat{F}_X^{-1}(0{\cdot}25)$ is the sample interquartile range for the distribution of variables $\{X_i\}$. Consider that the underlying distribution F_X is normal, for $k = 1$, the population cut-off point becomes $\eta = \mu_X + 1{\cdot}5\{F_X^{-1}(0{\cdot}75) - F_X^{-1}(0{\cdot}25)\} = F_X^{-1}(0{\cdot}75) + \{F_X^{-1}(0{\cdot}75) - F_X^{-1}(0{\cdot}25)\}$. In this case, the cut-off point $\hat{\eta}$ for the outlier sum is equal to the outlier robust t-statistic approach. Hence, $L = \sum_{i=1}^{n_2} Y_i I[Y_i > \{\hat{F}_X^{-1}(0{\cdot}5) + 1{\cdot}5k\mathrm{IQR}_X\}]$ for $k > 0$ may serve as a generalization of the classical outlier sum. However, one interesting study is to see if $k = 1$ provides the desired results for gene expression analysis.

Let us now study the asymptotic distribution of the outlier mean $\bar{L}$ in this specification of $\hat{\eta}$. A Bahadur representation for the outlier mean can be stated in the following theorem.

THEOREM 1. *If Assumptions 1, 2, 3 and 4 in the Appendix are true, a Bahadur representation of the outlier mean is*

$$
\begin{aligned}
n_2^{1/2}(\bar{L} - \mu_\ell) = {} & \frac{1}{4}(2b_1 + b_2 - 3b_3) \times n_1^{-1/2} \sum_{i=1}^{n_1} I\left\{X_i \leqslant F_X^{-1}(0{\cdot}25)\right\} \\
& + \frac{1}{4}(2b_1 + b_2 + b_3) \times n_1^{-1/2} \sum_{i=1}^{n_1} I\left\{F_X^{-1}(0{\cdot}25) \leqslant X_i \leqslant F_X^{-1}(0{\cdot}5)\right\} \\
& + \frac{1}{4}(-2b_1 + b_2 + b_3) \times n_1^{-1/2} \sum_{i=1}^{n_1} I\left\{F_X^{-1}(0{\cdot}5) \leqslant X_i \leqslant F_X^{-1}(0{\cdot}75)\right\} \\
& + \frac{1}{4}(-2b_1 - 3b_2 + b_3) \times n_1^{-1/2} \sum_{i=1}^{n_1} I\left\{X_i \geqslant F_X^{-1}(0{\cdot}75)\right\} \\
& + \frac{1}{\beta} n_2^{-1/2} \sum_{i=1}^{n_2} \{(Y_i - \mu_Y) I(Y_i > \eta) - \mu_\eta\} + o_p(1),
\end{aligned}
$$

where $\mu_\eta = \beta^{-1} \int_\eta^\infty (y - \mu_Y) f_Y(y) dy$, $\mu_\ell = \beta^{-1} \int_\eta^\infty y f_Y(y) dy$, $b_1 = (\eta - \mu_Y)\beta^{-1} f_Y(\eta) \gamma^{1/2} f_X^{-1}\{F_X^{-1}(0{\cdot}5)\}$, $b_2 = (1{\cdot}5k)(\eta - \mu_Y)\beta^{-1} f_Y(\eta)\gamma^{1/2} f_X^{-1}\{F_X^{-1}(0{\cdot}75)\}$ and $b_3 = (1{\cdot}5k)(\eta - \mu_Y)\beta^{-1} f_Y(\eta)\gamma^{1/2} f_X^{-1}\{F_X^{-1}(0{\cdot}25)\}$. In this case, $\beta = \mathrm{pr}(Y \geqslant \eta)$.

The asymptotic distribution of the outlier mean can be obtained from the central limit theorem.

THEOREM 2. *If Assumptions 2, 3 and 4 in the Appendix are true, then $n_2^{1/2}(\bar{L} - \mu_\ell)$ converges in distribution to $N(0, \sigma_\ell^2)$, where*

$$
\begin{aligned}
\sigma_\ell^2 &= \sigma_\ell^2(b_1, b_2, b_3, v) \\
&= \frac{3}{256}\{(2b_1 + b_2 - 3b_3)^2 + (2b_1 + b_2 + b_3)^2 + (-2b_1 + b_2 + b_3)^2 + (-2b_1 - 3b_2 + b_3)^2\} + v, \\
v &= \frac{1}{\beta^2}\mathrm{var}\{(Y - \eta) I(Y > \eta)\}.
\end{aligned}
$$

After replacing the population parameters by their corresponding estimators in the formulas of Theorem 2 and in (3), the p-value in (4) can be computed.

There are two scenarios that require different approaches to find estimators of β, μ_ℓ and σ_ℓ. When distributions F_X and F_Y are known but involve some unknown parameters, maximum likelihood estimation

can be used. When the distributions F_X and F_Y are unknown, a nonparametric technique can be used for estimating β, μ_ℓ and σ_ℓ by first estimating the mean μ_Y, percentile $F_X^{-1}(\alpha)$, truncated mean μ_ℓ, truncated variance σ_ℓ^2 and densities f_X and f_Y. Theorem 2, along with equation (4), provides a new method for computing the p-value that can be useful in different scenarios. Although the nonparametric approach can broaden the application areas, this paper focuses on the parametric model.

In the rest of this section, we develop an explicit form for the p-value under an added assumption that variables X and Y have normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. Under these assumptions, we have $F_X^{-1}(\alpha) = \mu_X + z_\alpha \sigma_X$, which implies $F_X^{-1}(0{\cdot}5) + 1{\cdot}5k\{F_X^{-1}(0{\cdot}75) - F_X^{-1}(0{\cdot}25)\} = \mu_X + 3kz_{0{\cdot}75}\sigma_X$, where z_α is the 100αth percentile of the standard normal distribution. This gives the outlier sum

$$L = \sum_{i=1}^{n_2} Y_i I(Y_i > \hat{\mu}_X + 3kz_{0{\cdot}75}\hat{\sigma}_X) \tag{5}$$

after replacing μ_X and σ_X by their estimators. Using equation (5) and the aforementioned properties and substituting the population parameters by their estimates, i.e. $\hat{\gamma} = n_2/n_1$, $\hat{\mu}_X = \bar{x} = n_1^{-1}\sum_{i=1}^{n_1} x_i$, $\hat{\sigma}_X^2 = s_x^2 = (n_1 - 1)^{-1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2$, we obtain

$$\ell = \sum_{i=1}^{n_2} y_i I(y_i > \bar{x} + 3kz_{0{\cdot}75}s_x), \qquad \beta = 1 - \Phi(3kz_{0{\cdot}75}),$$

$$\hat{\mu}_\ell = \bar{x} + \frac{s_x}{\beta}\int_{(3kz_{0{\cdot}75})}^{\infty} z\phi(z)dz, \qquad \hat{b}_1 = \frac{(2\pi)^{1/2}}{\beta} 3kz_{0{\cdot}75}s_x\phi(3kz_{0{\cdot}75})\hat{\gamma}^{1/2},$$

$$\hat{b}_2 = \hat{b}_3 = \frac{1{\cdot}5k\hat{b}_1\phi^{-1}(z_{0{\cdot}75})}{(2\pi)^{1/2}}, \qquad \hat{v} = \frac{s_x^2}{\beta^2}\left[\int_{(3kz_{0{\cdot}75})}^{\infty} z^2\phi(z)dz - \left\{\int_{(3kz_{0{\cdot}75})}^{\infty} z\phi(z)dz\right\}^2\right],$$

where β is a known constant in this parametric setting.

With the above formulas, Theorem 2 permits the computation of $\hat{\sigma}_\ell^2 = \sigma_\ell^2(\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{v})$. By plotting estimates of μ_x and σ_x , the p-value of (4) can be obtained. This technique can be extended to other known population distributions of X and Y.

4. Simulation study

4·1. *Type I error*

This simulation study examines whether the location of the cut-off point affects Type I error. From (5), we will make comparisons based on the value of k. We assume that gene expression in both control and disease groups follows a standard normal distribution, and each has a sample size $n = 20$. The p-value is computed based on (4). Type I error is calculated as the proportion of the p-values less than 0·05 among simulation runs. The results show that as k increases, or equivalently, the cut-off point shifts more away from the median, Type I error gets smaller and the chance of rejecting the null hypothesis decreases. Type I errors are 0·058, 0·0075, 0·00043, 0·00002 and 0·000002 for $k = 1, 1{\cdot}5, 2, 2{\cdot}5$ and 3, respectively, at $\alpha = 0{\cdot}05$. Clearly, when $k = 1$, Type I error is closest to the targeted significance level. When the adjusted p-value cut-off is 0·038, Type I error reaches 0·05 for $k = 1$. Thus, we opt for $k = 1$ for the distribution-based p-value approach to compare with other outlier statistics and choose 0·038 as the significance level to control for Type I error at 0·05 in the power study.

4·2. *Power analysis*

In the power study, we compare the distribution-based p-value approach with four other approaches: two-sample t-test, the cancer outlier profile analysis approach, the outlier sum approach, and the outlier robust t-statistic approach. Three scenarios are examined: (a) the genes in the control and disease groups follow the $N(0, 1)$ distribution, except for the first gene in the n_0 outlier samples from the disease group that follows the $N(d, 1)$ distribution, where d is the effect size; (b) the genes in the control and disease groups follow the t-distribution with four degrees of freedom, except for the first gene in the n_0 outlier samples from the disease group, which follows the noncentral t-distribution with four degrees of freedom

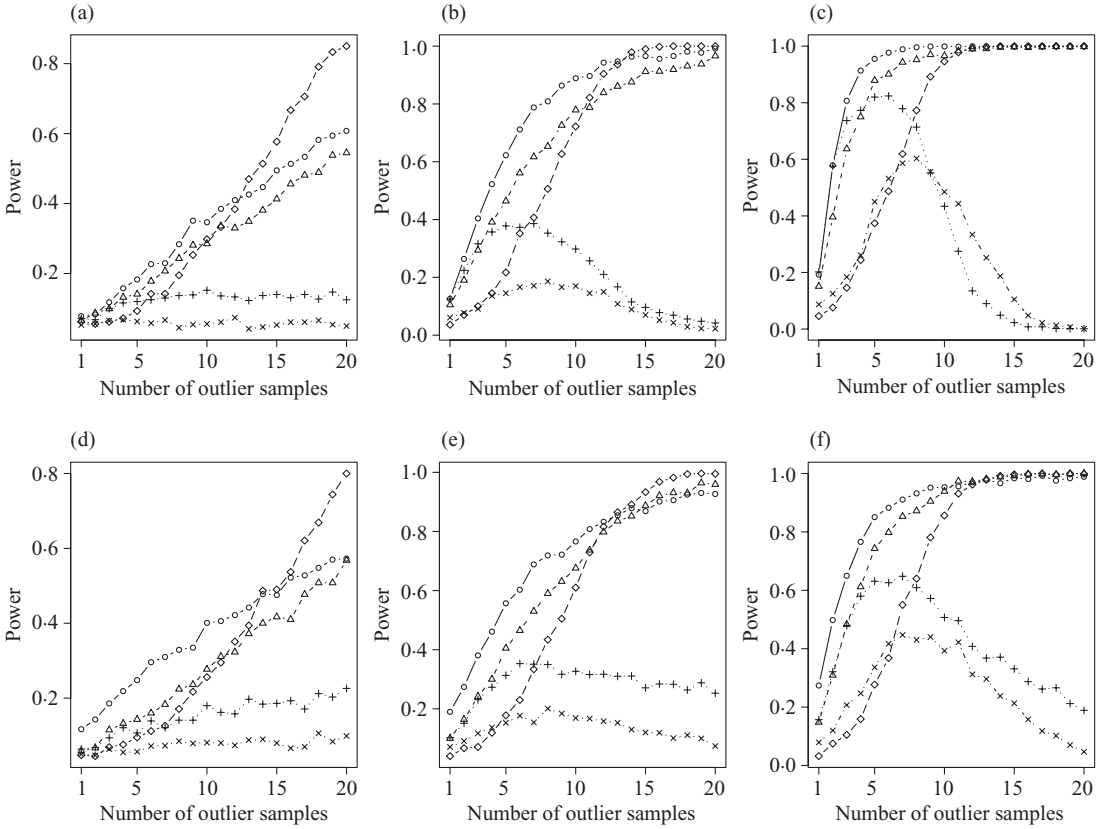

Fig. 1. Power comparison of various effect sizes with equal variance for (a)–(c) normal distribution with an effect size of 1, 2 and 3, respectively, and for (d)–(f) t-distribution with 4 degrees of freedom and noncentrality of 1, 2 and 3, respectively. The following approaches are illustrated: distribution-based p-value for the outlier sum (○); outlier robust t-statistic (△); outlier sum (+); cancer outlier profile analysis (×); and two-sample t-test (◇).

and noncentrality d; (c) the genes in the control and disease groups follow the $N(0, \sigma^2)$ distribution, except for each gene in the n_0 outlier samples from the disease group that follows the $N(2\sigma, \sigma^2)$ distribution. For scenarios (a) and (b), we evaluate how the effect size, noncentrality, and number of outlier samples affect the power. Specifically, we simulate 1000 genes for each of the 20 control and 20 disease samples and choose d from 1, 1·5, 2, 3, to 4. In addition, we let n_0 vary from 1 to 20. For scenario (c), each gene has a different variance, but the effect size, the ratio of mean and standard deviation, is each fixed at 2. To replicate our data sample presented in § 5, we simulate 12 625 genes for each of the 19 control and 18 disease samples and allow each gene to have a separate σ, estimated from the data sample and ranging from 0·1 to 2·4. In this scenario, we let n_0 vary from 1 to 18.

Power is calculated as the proportion of 1000 simulation runs that have a significant difference based on Type I error 0·05. For scenarios (a) and (b), power calculation is based on the first gene. For scenario (c), the average power of all genes is used. Significance is determined by the p-value when methods have a p-value formula available. Otherwise, significance is determined by whether the test statistic is greater than the 95th percentile of the outlier statistic, based on the parametric bootstrapping method.

From Fig. 1, the distribution-based p-value approach has the highest power when the number of outlier samples is smaller than 10. In contrast, when the number of outlier samples becomes large, such as greater than 15, the t-test yields the highest power. Interestingly, the distribution-based p-value approach and the outlier robust t-statistic approach also perform well. On the other hand, the cancer outlier profile analysis and the outlier sum have low power overall. From Fig. 2, with three outlier samples or less, power increases as effect size increases in the distribution-based p-value, the outlier robust t-statistic, and the outlier sum approaches, but not in the other two approaches for both normal and t-distributions. The distribution-based p-value approach performs best in most cases. As the number of outlier samples increases to 15 or more,

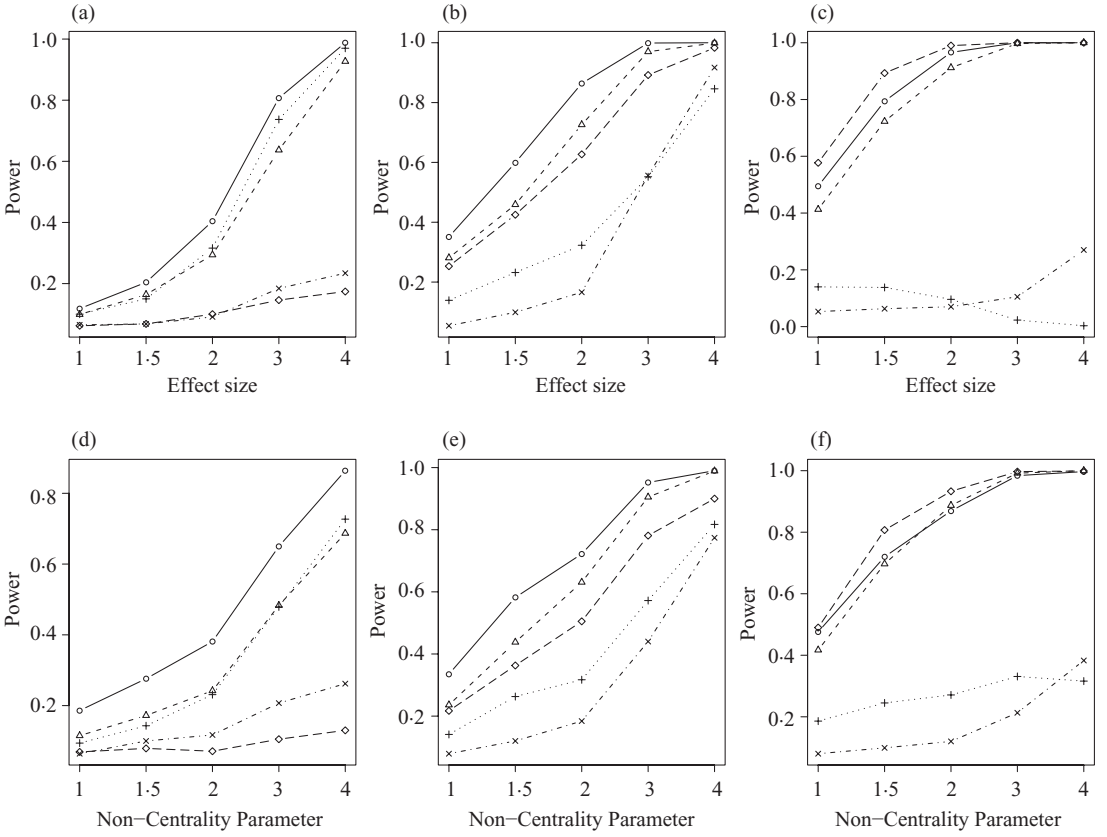

Fig. 2. Power comparison for various numbers of outlier tissues for (a)–(c) normal distribution with an effect size of 1, 2 and 3, respectively, and for (d)–(f) t-distribution with 4 degrees of freedom and noncentrality of 1, 2 and 3, respectively. The following approaches are illustrated: distribution-based p-value for the outlier sum (○); outlier robust t-statistic (△); outlier sum (+); cancer outlier profile analysis (×); and two-sample t-test (◇).

the t-test performs best, but the distribution-based p-value approach and the outlier robust t-statistic approach yield a comparable power. In contrast, the other two approaches have a power less than 0·4. The results demonstrate that the distribution-based p-value approach has better power to identify outliers. When the number of outlier samples is large, the strength of identifiability diminishes. For example, when all disease samples are outliers in the extreme case, the statistical issue becomes a standard two-group comparison and hence the identifiability of the outlier approaches disappears. For scenario (c) of differing variances, the results, not presented, are similar to those of Fig. 1 at the effect size equal to 2. Moreover, all five approaches show a very weak correlation between power and standard deviation. This phenomenon indicates that the power to detect outliers is independent of the variance as long as the effect size is fixed.

5. Data example

We apply all five approaches discussed in § 4 to the breast cancer microarray data reported by Huang et al. (2003). This dataset contains the expression levels of 12 625 genes from 37 breast tumor samples: 19 of them have no positive nodes discovered and are treated as the control group and the other 18 samples have identifiably positive nodes and are treated as the disease group. Since we test 12 625 genes simultaneously, errors in inference are more likely to occur without p-value adjustment. To account for multiple testing, the false discovery rate method (Benjamini & Hochberg, 1995) is used to correct for the p-value at the 0·05 level for the distribution-based p-value approach and the t-test. For the other three approaches, the 99th percentile of each outlier statistic is used as the cut-off point to identify significant outlier genes based on the parametric bootstrapping method.

Results show that the t-test does not identify any significant outlier genes. In contrast, the other four approaches identify 584, 535, 740 and 695 outlier genes, respectively, for the distribution-based p-value

approach, the outlier robust t-statistic approach, the outlier sum approach and the cancer outlier profile analysis approach. We further find that four disease samples are consistently ranked in the top four in number of outlier genes by the distribution-based p-value approach, the outlier robust t-statistic approach and the outlier sum approach. Each of these four disease samples has at least 20% of the outlier genes with a higher expression, suggesting abnormal up-regulation in gene expression. In addition, there are eight outlier genes consistently showing high expression in all four samples.

This application highlights the strength of the four outlier approaches, which identify many significant outlier genes and point out several disease samples with an abnormally large number of outlier genes. On the other hand, the t-test fails to detect any outlier gene. In addition, computation of the p-value is easier with the distribution-based p-value approach than the other three outlier approaches. These three approaches lack a distribution theory and hence need special procedures, such as the parametric bootstrapping method, to define a cut-off point for each outlier gene as the variance changes.

ACKNOWLEDGEMENT

The authors are grateful to the referees, an associate editor and the editor for comments that led to improvements in this manuscript. The research of the first author was partially supported by a grant from the National Science Council of Taiwan. The research of the second and third authors was partially supported by various grants from the National Institutes of Health, U.S.A.

APPENDIX

Four assumptions for the outlier sum test statistic are as follows.

Assumption 1. The proportion of outlier samples, $n_2^{-1}\sum_{i=1}^{n_2} I(Y_i > \hat{\eta})$, converges in probability to β with $0 < \beta < 1$.

Assumption 2. The limit $\gamma = \lim_{n_2,n_1\to\infty} n_2/n_1$ exists.

Assumption 3. The probability density function f_X is bounded away from zero in a neighbourhood of $F_X^{-1}(\alpha)$ for $\alpha \in (0, 1)$.

Assumption 4. The probability density function f_Y is bounded away from zero in a neighbourhood of the population cut-off point η.

Proof of Theorem 1. With Assumption 3, a representation of $\hat{F}_X^{-1}(\alpha)$ such as

$$n_1^{1/2}\{\hat{F}_X^{-1}(\alpha) - F_X^{-1}(\alpha)\} = f_X^{-1}\{F_X^{-1}(\alpha)\}n_1^{-1/2}\sum_{i=1}^{n_1}\left[\alpha - I\{X_i \leqslant F_X^{-1}(\alpha)\}\right] + o_p(1), \tag{A1}$$

implies that $\hat{\eta} = \hat{F}_X^{-1}(0{\cdot}5) + 1{\cdot}5k \times \mathrm{IQR}_X$ satisfies $T = n_1^{1/2}(\hat{\eta} - \eta) = O_p(1)$ (Ruppert & Carroll, 1980). From the expression of the outlier mean in (1), we have

$$\begin{aligned}
\bar{L} &= \left\{\sum_{i=1}^{n_2} I(Y_i > \hat{\eta})\right\}^{-1}\sum_{i=1}^{n_2} Y_i I(Y_i > \hat{\eta}) \\
&= \left\{\sum_{i=1}^{n_2} I(Y_i > \hat{\eta})\right\}^{-1}\sum_{i=1}^{n_2} Y_i I\left(Y_i > \eta + n_1^{-1/2}T\right) \\
&= \mu_Y + \left\{\sum_{i=1}^{n_2} I(Y_i > \hat{\eta})\right\}^{-1}\sum_{i=1}^{n_2}\left(Y_i - \mu_Y\right)I\left(Y_i > \eta + n_1^{-1/2}T\right).
\end{aligned}$$

The above expression can be rewritten as

$$n_2^{1/2}(\bar{L} - \mu_Y) = \left\{\sum_{i=1}^{n_2} I(Y_i > \hat{\eta})\right\}^{-1} n_2^{1/2}\sum_{i=1}^{n_2}(Y_i - \mu_Y)I\left(Y_i > \eta + n_1^{-1/2}T\right). \tag{A2}$$

With (A1), Assumptions 2 and 4, and techniques from Ruppert & Carroll (1980) and Chen & Chiang (1996), a modified second term on the right-hand side of (A2), can be expressed as

$$\begin{aligned} & n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_Y)\{I(Y_i>\eta+n_1^{-1/2}T)-I(Y_i>\eta)\}+n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_Y)I(Y_i>\eta) \\ & =-n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_Y)\{I(Y_i\leqslant\eta+n_1^{-1/2}T)-I(Y_i\leqslant\eta)\}+n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_Y)I(Y_i>\eta) \\ & =-(\eta-\mu_Y)f_Y(\eta)\gamma^{1/2}T+n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_Y)I(Y_i>\eta)+o_p(1). \end{aligned} \tag{A3}$$

By the same rationale and the weak law of large numbers, we can derive

$$n_2^{-1}\sum_{i=1}^{n_2}I[Y_i>\hat{F}_X^{-1}(0{\cdot}5)+1{\cdot}5k\{\hat{F}_X^{-1}(0{\cdot}75)-\hat{F}_X^{-1}(0{\cdot}75)\}]=n_2^{-1}\sum_{i=1}^{n_2}I(Y_i>\eta)+o_p(1) \tag{A4}$$

which converges to $\mathrm{pr}(Y\geqslant\eta)$.

Combining (A2), (A3) and (A4), a Bahadur representation of the outlier mean is

$$\begin{aligned} n_2^{1/2}(\bar{L}-[\mu_Y+\beta^{-1}E\{(Y-\eta)I(Y\geqslant\eta)\}]) &= -\beta^{-1}(\eta-\mu_Y)f_Y(\eta)\gamma^{1/2} \\ &\quad\times n_1^{1/2}(\hat{\eta}-\eta)+\beta^{-1}n_2^{-1/2}\sum_{i=1}^{n_2}\{(Y_i-\mu_Y)I(Y_i>\eta)-E\{(Y-\eta)I(Y\geqslant\eta)\}+o_p(1). \end{aligned}$$ □

References

Agrawal, D., Chen, T., Irby, R., Quackenbush, J., Chambers, A. F., Szabo, M., Cantor, A., Coppola, D. & Yeatman, T. J. (2002). Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *J. Nat. Cancer Inst.* **94**, 513–21.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–11.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Chen, L.-A. & Chiang, Y.-C. (1996). Symmetric quantile and trimmed means for location and linear regression model. *J. Nonparam. Statist.* **7**, 171–85.

Huang, E., Cheng, S. H., Dressman, H., Pittman, J., Tsou, M. H., Horng, C. F., Bild, A., Iversen, E. S., Liao, M., Chen, C. M., West, M., Nevins, J. R. & Huang, A. T. (2003). Gene expression predictors of breast cancer outcomes. *Lancet* **361**, 1590–6.

Ohki, R., Yamamoto, K., Ueno, S., Mano, H., Misawa, Y., Fuse, K., Ikeda, U. & Shimada, K. (2005). Gene expression profiling of human atrial myocardium with atrial fibrillation by DNA microarray analysis. *Int. J. Cardiol.* **102**, 233–8.

Ruppert, D. & Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Am. Statist. Assoc.* **75**, 828–38.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lonning, P. E., Brown, P. O., Borresen-Dale, A. L. & Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Nat. Acad. Sci.* **100**, 8418–23.

Tibshirani, R. & Hastie, T. (2007). Outlier sums 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., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A. & Chinnaiyan, A. M. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–8.

Wu, B. (2007). Cancer outlier differential gene expression detection. *Biostatistics* **8**, 566–75.

[*Received July* 2008. *Revised September* 2009]