國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

離群值比例之基因分析

Outlier Proportion Based Gene Expression Analysis

研 究 生：刁瀅潔

指導教授：陳鄰安　博士

中 華 民 國 九 十 九 年 六 月

離群值比例之基因分析

Outlier Proportion Based Gene Expression Analysis

研 究 生：刁瀅潔　　Student：Ying-Chieh Tiao

指導教授：陳鄰安　博士　　Advisor：Dr. Lin-An Chen

國 立 交 通 大 學

統 計 學 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

In Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2010

Hsinchu, Taiwan, Republic of China

中 華 民 國 九 十 九 年 六 月

離群值比例之基因分析

學生：刁瀅潔　　　　　　　　　　　　　　　　　　　指導教授：陳鄰安 博士

國立交通大學統計學研究所碩士班

摘　　　　要

藉由偵測病體樣本中的離群值而找出具有影響力的基因已是一種非常新且重要的基因分析方法。透過離群和或是離群平均可以偵測出離群資料中的集中趨勢是否有所改變，但是卻無法偵測出偏度等其它特徵量數。因此，我們希望可以提供一個容易實行且有較高檢定力的統計檢定，以作為基因分析的另一項替代選擇方法。我們將提出離群值比例的觀點，以離群值比例的近似分配為基礎，發展出一項統計檢定。此外，我們也將更進一步地比較離群值比例和離群平均兩者的檢定力表現。而為了避免估計尾端機率點的密度函數之困難，進而造成檢定力較低的缺點，因此我們將採用經驗分位數當作切點。

Outlier Proportion Based Gene Expression Analysis

Student: Ying-Chieh Tiao Advisor: Dr. Lin-An Chen

Institute of Statistics

National Chiao Tung University

Abstract

Discovering the influential genes through the detection of outliers in samples of disease group subjects is a very new and important approach for gene expression analysis. The outlier sum or outlier mean technique can detect the shift in central tendency for the outlier data but not other characteristics such as spreadness or others for the outlier data. It is desired to provide a test that is easy to implement and efficient in power performance as an alternative tool for gene expression analysis. We propose the concept of outlier proportion for developing a test based on asymptotic distribution of this statistics. We further compare it with the outlier mean for their power performances. To avoid the inefficiency in estimating densities at tail quantiles involved in estimation of outlier proportion variance, we further consider applying the empirical quantile as the cutoff point for an alternative outlier proportion based test which shows satisfactory role in gene expression analysis from the point of power performance.

致　　　　謝

從小到大，十八年的學生生涯即將奏起最終樂章，而這也意味著我將正式地離開學校生活，投入職場展翅飛翔。

首先我要由衷地感謝我的論文指導老師－陳鄰安教授。老師總是很用心地的指導著我，耐心地為我解惑。古人云：「師者，所以傳道、授業、解惑也。」這些都一再地在老師身上得到印證。不僅如此，老師也會不時地關心著我的生活情況，讓我備感溫暖，銘感五內。更因為老師的教導及用心，讓我無論是在課業或是論文方面，都得到了最多且珍貴的收穫，讓我覺得自己真的是很幸福，可以遇到這麼棒的一位論文指導老師，因此我非常真摯地說一句：「老師，謝謝您！」。此外，也要謝謝三位論文口試委員對這篇論文的指教和建議，使得整篇論文可以更加豐富與完整。

再者我要謝謝交大統計所 97 級的所有同學，感謝你們這兩年陪我一起成長，還有要謝謝郭姊這兩年辛苦地替我們打理研究所生活的一切事宜，真的辛苦您了。此外，我也要謝謝一直陪伴在我身邊的朋友、學長姐和學弟妹們，尤其是我最愛的大學七姊妹們，因為你們的鼓勵與陪伴，給了我最大的信心，我相信我們之間的這份友情，一輩子都不會改變。

最後，我要感謝我的家人，因為你們一直是我的最佳後援部隊，給了我無限的支持與關心，讓我可以無後顧之憂的去追求我的夢想，真的非常謝謝你們。

在此，將本篇論文獻給我的家人、朋友、師長以及所有曾經幫助過我，陪伴過我的人。我將致上我最真摯的謝意，和你們分享這份成果與喜悅。

刁瀅潔　　謹誌于
國立交通大學統計學研究所
中華民國九十九年六月

Contents

摘要 i

Abstract.. ii

致謝 iii

1. Introduction 1

2. Outlier Proportion 3

3. A Test Based on Asymptotic Distribution of

Sample Outlier Proportion 4

4. An Outlier Proportion Test With Empirical

Quantile as Cutoff point 8

5. Simulation Study 12

6. Appendix 15

Reference 17

Outlier Proportion Based Gene Expression Analysis

SUMMARY

Discovering the influential genes through the detection of outliers in samples of disease group subjects is a very new and important approach for gene expression analysis. The outlier sum or outlier mean technique can detect the shift in central tendency for the outlier data but not other characteristics such as spread or others for the outlier data. It is desired to provide a test that is easy to implement and efficient in power performance as an alternative tool for gene expression analysis. We propose the concept of outlier proportion for developing a test based on asymptotic distribution of this statistic. We further compare it with the outlier mean for their power performances. To avoid the inefficiency in estimating densities at tail quantiles involved in estimation of outlier proportion variance, we further consider applying the empirical quantile as the cutoff point for an alternative outlier proportion based test which shows satisfactory role in gene expression analysis from the point of power performance.

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); Alizadeh et al. (2000); Ohki et al. (2005)); Sorlie et al. (2003)). Among the existed techniques in differential genes detection, common statistical methods for two-group comparisons such as t-test, are not appropriate due to a large number of genes expressions and a limited number of subjects available. Several statistical approaches have been proposed to identify those genes where only a subset of the sample genes has high expression. Among them, Tomlins et al. (2005) observed that there is small number of outliers in samples of differential genes and then introduced a method called cancer outlier profile analysis that identifies outlier profiles by a statistic based on the median and the median absolute deviation of a gene expression profile. With this observation, a sequence of approaches then concentrated on detecting differential genes based on out-

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX lier samples while Tibshirani and Hastie (2007) and Wu (2007) suggested to use an outlier sum, the sum of all the gene expression values in the disease group that are greater than a specified cutoff point. The common disadvantage of these techniques is that the distribution theory of the proposed methods has not been discovered so that the distribution based p value can not been applied. Recently Chen, Chen and Chan (2010) considered the outlier mean (average of outlier sum) and developed its large sample theory that allows us to formulate the distribution based p value. In specific, they considered the parametric study by specifying the normal distribution and performed simulation studies and data analysis for gene expression analysis.

According to Tomlins et al. (2005), it is desired to verify if the variables for disease group subjects and normal group subjects on the region excessed a cutoff point are identical. The outlier mean approach of Chen, Chen and Chan (2010) can detect if the excessive means are different. We know that summarizing the outlier data by its sum or mean (average) may be efficient when the central tendencies of two distributions on excessive region are significantly different. However, it is known that it is not enough to detect just the shift in mean while there may have a shift other than the central tendency. So, it requires to measure other characteristics showing in the outlier data as an alternative for detection of influential genes. Here, in this paper, we consider the proportion of outlier data, called the outlier proportion, to detect the influential genes. Interestingly this study shows that outlier proportion technique provides a technique very simple in computation but it is also much more efficient than the outlier mean test in detection of influential genes.

In Section 2, we introduce the concept of population outlier proportion and study the adequacy for using it in detection of distributional shift. In Section 3, we study large sample property of the outlier variance and we compare the power performances between the tests based on outlier mean and outlier proportion. In Section 4, we propose an alternative outlier proportion based test that avoids the estimation of densities on extreme quantiles for construction of test statistic.

2. Outlier Proportion

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 $X_{ij}, i = 1, 2, ..., n_1, j = 1, ..., m$ for normal control group and $Y_{ij}, i = 1, 2, ..., n_2, j = 1, 2, ..., m$ for the disease group.

For theoretical development, let us fix a gene and we drop the index j. Let X and Y be expression variables with expression $X_i, i = 1, ..., n_1$ for group of normal subject and $Y_i, i = 1, ..., n_2$ for group of disease subject, respectively, with distribution functions F_X and F_Y.

An important observation by Tomlins et al. (2005) from a study of prostate cancer, outlier genes are over-expressed only in a small number of disease samples. With defining a cutoff point $\hat{\eta}$ determined from the data of the variable X, Tibshirani and Hastie (2007) and Wu (2007) considered the sum of variables $Y_i's$ that are over higher cutoff point $\hat{\eta}$ given by $\sum_{i=1}^{n_2} Y_i I(Y_i \geq \hat{\eta})$ as a test statistic for detection if the disease group distribution is different from the normal group distribution. Latter, Chen, Chen and Chan (2010) developed the asymptotic distribution for its average, called the outlier mean, $L_Y = (\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}))^{-1} \sum_{i=1}^{n_2} Y_i I(Y_i \geq \hat{\eta})$ for constructing a distribution based p value. Let η be the population counterpart of the sample cutoff point $\hat{\eta}$. The idea behind the outlier mean approach considers a test based on L_Y to verify if its corresponding population outlier mean $\mu_{\ell_Y} = E(Y|Y \geq \eta)$ varied from the same population outlier mean when $F_Y = F_X$ as $\mu_{\ell_X} = E(X|X \geq \eta)$.

We consider here to establish a test based on the sample outlier proportion, a tail probability estimator, as

$$\hat{\beta}_Y = n_2^{-1} \sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}). \tag{2.1}$$

Hence, the idea behind this sample percentage is to verify if its corresponding population outlier proportion

$$\beta_Y = P\{Y \geq \eta\} \tag{2.2}$$

varied from the same population outlier proportion when $F_Y = F_X$ as $\beta_X = P\{X \geq \eta\}$.

To verify if this consideration is appropriate, we suggest the population cutoff point of the form $\eta = 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)$ and make a numerical comparison of two outlier proportions. We consider the following setting

$$\text{Normal}: X \sim N(0,1) \text{ and } Y \sim N(\theta,1),$$
$$\text{Mixed normal}: X \sim N(0,1), Y \sim 0.9N(0,1) + 0.1N(\theta,\sigma^2).$$

Population outlier proportions for variables X and Y under the above settings are displyed in Table 1 with the specified $\alpha's$ and $\theta's$.

Table 1. Population outlier proportions ($\sigma = 1$)

α	β_X	$\theta = 1$ β_Y	$\theta = 3$ β_Y	$\theta = 5$ β_Y
	$F_X = N(0,1)$	$F_Y = N(\theta,1)$		
0.01	$1.48E-12$	$1.12E-9$	$3.19E-7$	0.0239
0.05	$4.01E-7$	$4.16E-5$	0.0016	0.5260
0.1	$6.03E-5$	0.0022	0.0325	0.8760
0.2	0.0057	0.0636	0.2998	0.9933
0.25	0.0215	0.1530	0.4906	0.9985
0.35	0.1238	0.4380	0.8006	0.9999
0.45	0.3530	0.7333	0.9477	0.9999
		Mixed Normal		
0.01		$1.13E-10$	$3.19E-8$	0.0023
0.05		$4.52E-6$	$1.67E-4$	0.0526
0.1		$2.76E-4$	0.0033	0.0876
0.2		0.0115	0.0351	0.1045
0.25		0.0346	0.0684	0.1192
0.35		0.1552	0.1915	0.2114
0.45		0.3911	0.4125	0.4177

Conceptually the bigger the difference $\beta_Y - \beta_X$, the easier to establish a test in detection of distributional shift. From Table 1, we expect that larger $\alpha's$ make the detection by outlier proportion more powerful. We will evaluate this point in the subsequent sections.

3. A Test Based on Asymptotic Distribution of Sample Outlier Proportion

The sample outlier proportion is defined by

$$\hat{\beta}_Y = \frac{1}{n_2}\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta})$$

where cutoff point estimator is $\hat{\eta} = 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha)$ and where $\hat{F}_X^{-1}(\delta)$ is the δth empirical quantile based on sample $X_i, i = 1, ..., n_1$.

To construct a distribution based test statistic by this outlier proportion, we state an asymptotic distribution for this statistic in the following theorem where its proof is given in Appendix.

Theorem 3.1. Suppose that assumptions (A_2) and (A_3) in the Appendix are true. Then $n_2^{1/2}(\hat{\beta}_Y - \beta_Y)$ converges in distribution to $N(0, \sigma_\beta^2)$ where

$$\sigma_\beta^2 = \alpha(\alpha b_1 - (1-\alpha) b_2)^2 + (1-2\alpha)\alpha^2(b_1+b_2)^2 + \alpha(-(1-\alpha)b_1 + \alpha b_2)^2 + \beta_Y(1-\beta_Y).$$

Here we let

$$\begin{aligned} b_1 =& 2\gamma f_Y(\eta) f_X^{-1}(F_X^{-1}(1-\alpha)), \\ b_2 =& \gamma f_Y(\eta) f_X^{-1}(F_X^{-1}(\alpha)). \end{aligned}$$

This theorem indicates, under $H_0 : F_x = F_y$, the following

$$P_{H_0}\{\sqrt{n_2}(\frac{\hat{\beta}_Y - \beta_X}{\sigma_\beta}) \leq z\} \to \int_{-\infty}^{z} \phi(z)dz$$

for $z \in R$ where ϕ represents the probability density function of $N(0,1)$. Suppose that we have estimates $\hat{\sigma}_\beta$ and $\hat{\beta}_X$, a test based on the sample outlier proportion is

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}(\frac{\hat{\beta}_Y - \hat{\beta}_X}{\hat{\sigma}_\beta}) \geq z_{\alpha^*}. \tag{3.1}$$

The test tries to see if outlier proportion for disease group subjects is different from it for normal group subjects. As a nonparametric approach, this test statistic involves the estimation of some density points f_X and f_Y.

Having this sample outlier proportion based nonparametric test, it is desired to verify the power performance of this test when there exists distributional shift for the disease group distribution. An approximate power with significant level α^* may be derived as bellows

$$
\begin{aligned}
p_p &= P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\beta}_Y - \hat{\beta}_X}{\hat{\sigma}_\beta}) \geq z_{\alpha^*}\} \\
&= P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\beta}_Y - \beta_Y}{\sigma_\beta}) \geq \sqrt{n_2}(\frac{\frac{z_{\alpha^*}\hat{\sigma}_\beta}{\sqrt{n_2}} + \hat{\beta}_X - \beta_Y}{\sigma_\beta})\} \\
&\approx P\{Z \geq z_{\alpha^*} + \frac{\sqrt{n_2}(\beta_X - \beta_Y)}{\sigma_\beta}\}. \qquad (3.2)
\end{aligned}
$$

Considering the following distributional settings,

Normal: $X \sim N(0,1), Y \sim N(\theta,1)$

Laplace distribution: $X \sim Laplace(0,1), Y \sim Laplace(\theta,1)$

t distribution :$X \sim t(5), Y \sim t(5)+\theta$,

we display the powers p_m, for outlier mean based test, and p_p, for outlier proportion based test, in Table 2.

Table 2 Approximate powers of outlier mean and outlier proportion

α	$\theta = 1$	$\theta = 2$	$\theta = 3$
Normal			
$\alpha = 0.45, p_m$	0.523	0.999	1
p_p	0.908	1	1
$\alpha = 0.35, p_m$	0.144	0.844	1
p_p	0.537	1	1
$\alpha = 0.25, p_m$	0.063	0.294	0.863
p_p	0.262	0.992	0.999
$\alpha = 0.15, p_m$	0.052	0.111	0.151
p_p	0.122	0.537	0.579
Laplace			
$\alpha = 0.45, p_m$	0.289	0.993	1
p_p	0.979	1	1
$\alpha = 0.35, p_m$	0.050	0.414	0.999
p_p	0.390	1	1
$\alpha = 0.25, p_m$	0.050	0.255	0.490
p_p	0.219	0.798	0.999
$\alpha = 0.15, p_m$	0.050	0.05	0.050
p_p	0.123	0.441	0.260
t-distrib			
$\alpha = 0.45, p_m$	0.412	0.994	1
p_p	0.898	1	1
$\alpha = 0.35, p_m$	0.077	0.418	0.999
p_p	0.543	1	1
$\alpha = 0.25, p_m$	0.043	0.052	0.518
p_p	0.332	0.995	0.998
$\alpha = 0.15, p_m$	0.046	0.027	0.016
p_p	0.203	0.828	0.687

How surprisingly the outlier proportion performs much better than the outlier mean in these three location distributional shifts.

According to Tomlins et al. (2005), it is desired to verify the power performance of the outlier proportion when there is only a small percentage of outliers in the data of Y. For this, we consider the following distributional setting:

$$X \sim Laplace(0,1), Y \sim 0.9Laplace(0,1) + 0.1Laplace(\theta, \sigma)$$

Table 3 Approximate powers of outlier mean and outlier proportion for Laplace mixture

α	p_m ($\theta = 3$)	p_p	p_m ($\theta = 5$)	p_p	p_m ($\theta = 10$)	p_p
$\sigma = 3$						
$\alpha = 0.45$	0.184	0.707	0.253	0.734	0.368	0.756
$\alpha = 0.35$	0.189	0.846	0.276	0.875	0.420	0.896
$\alpha = 0.25$	0.180	0.941	0.299	0.965	0.550	0.978
$\alpha = 0.15$	0.150	0.975	0.255	0.990	0.742	0.996
$\alpha = 0.05$	0.105	0.987	0.130	0.992	0.485	0.999

This computation shows that the outlier proportion is still a satisfactory one in this case of mixed distribution. This further support the use of outlier proportion in gene expression analysis.

4. An Outlier Proportion Test With Empirical Quantile as Cutoff point

We have observed that the outlier proportion may have satisfactory power performance when we have consistent estimators $\hat{\beta}_X$ and $\hat{\sigma}_\beta$ to construct test in (3.1). However, $\hat{\sigma}_\beta$ involves estimations of density points f_Y and f_X while estimation of density function at tail quantile points is extremely difficult in practice. Without an alternative proposal avoiding this density estimation, the outlier proportion based test won't be practically powerful in detection of influential genes unless n_1 and n_2, the numbers of disease group subjects and number of normal group subjects, are very large.

In this section, we choose cutoff point $\hat{\eta} = \hat{F}_X^{-1}(\gamma)$ for some $\gamma > 0$. For not being confused, we denote the outlier proportion as

$$\hat{\beta}_Y^* = \frac{1}{n_2} \sum_{i=1}^{n_2} I(Y_i \geq \hat{F}_X^{-1}(\gamma))$$

for estimating $\beta_Y^* = P(Y \geq F_X^{-1}(\gamma))$. We first study the differences of two population outlier proportions under the following distribution setting:

$$X \sim Laplace(0,1), Y \sim 0.9Laplace(0,1) + 0.1Laplace(\theta, \sigma).$$

Table 4. Population outlier proportions

σ	β_X	$\theta = 3$ β_Y	$\theta = 5$ β_Y	$\theta = 10$ β_Y
$\gamma = 0.9$				
$\sigma = 3$	0.203	0.751	0.872	0.975
$\sigma = 5$	0.203	0.671	0.779	0.918
$\sigma = 10$	0.203	0.594	0.668	0.798
$\gamma = 0.95$				
$\sigma = 3$	0.193	0.747	0.870	0.975
$\sigma = 5$	0.193	0.668	0.777	0.918
$\sigma = 10$	0.193	0.592	0.666	0.797

It is seen that the differences between two population proportions are quite significant when the quantile percentage γ is 0.9 or 0.95. This shows that using quantile as cutoff point in detection of outliers is quite satisfactory.

A large sample theory for this quantile based outlier proportion is stated below.

Theorem 4.1. Suppose that assumptions (A_2) and (A_3) in the Appendix are true. Then, $n_2^{1/2}(\hat{\beta}_Y^* - \beta_Y^*)$ converges in distribution to $N(0, \sigma_{\beta,Y}^2)$ where

$$\sigma_{\beta,Y}^2 = \gamma(1-\gamma)\gamma_{xy} f_Y^2(F_X^{-1}(\gamma)) f_X^{-2}(F_X^{-1}(\gamma)) + \beta_Y^*(1-\beta_Y^*).$$

To construct a test statistic based on the above theorem, we still face the problem of requiring estimation of $\sigma_{\beta,Y}^2$ that involved prediction of density points $f_Y(F_X^{-1}(\gamma))$ and $f_X(F_X^{-1}(\gamma))$ which is difficult unless there is huge sample. However, under H_0 we may replace f_Y by f_X and then $\sigma_{\beta,Y}^2$ is induced as

$$\sigma_{\beta,X}^2 = \gamma(1-\gamma)\gamma_{xy} + \beta_Y^*(1-\beta_Y^*).$$

In this setting, we need only to find estimates $\hat{\beta}_Y^*$ and $\hat{\beta}_X^*$ to build the outlier proportion based test as

$$\text{rejecting } H_0 \text{ if } \sqrt{n_2}\left(\frac{\hat{\beta}_Y^* - \hat{\beta}_X^*}{\hat{\sigma}_{\beta,X}}\right) \geq z_{\alpha^*}. \tag{4.1}$$

An approximate power for outlier proportion based on this quantile cutoff point at significance level α^* may be derived as bellows

$$\begin{aligned}
&P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\beta}^*_Y - \hat{\beta}^*_X}{\hat{\sigma}_{\beta,X}}) \geq z_{\alpha^*}\} \\
&\quad = P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\beta}^*_Y - \beta^*_Y}{\sigma_{\beta,Y}}) \geq \sqrt{n_2}(\frac{\frac{z_{\alpha^*}\hat{\sigma}_{\beta,X}}{\sqrt{n_2}} + \hat{\beta}^*_X - \beta^*_Y}{\sigma_{\beta,Y}})\} \\
&\quad \approx P\{Z \geq z_{\alpha^*}\frac{\sigma_{\beta,X}}{\sigma_{\beta,Y}} + \frac{\sqrt{n_2}(\beta^*_X - \beta^*_Y)}{\sigma_{\beta,Y}}\}. \qquad (4.2)
\end{aligned}$$

where $\beta^*_X = P(X \geq F_X^{-1}(\gamma))$.

It is interested to compare outlier mean and outlier proportion both using quantile cutoff point in terms of powers. First, we consider the following two location shift models:

Case 1:$X \sim N(0,1)$ and $Y \sim N(\theta,1)$

Case 2:$X \sim Laplace(0,1)$ and $Y \sim Laplace(\theta,1)$

We display the results of power in the following table.

Table 5 Approximate powers of outlier mean and outlier proportion

Power	$\theta = 1$	$\theta = 2$	$\theta = 4$
Case 1			
$(\gamma = 0.9)p_m$	0.180	0.858	1.0
p_p	0.687	0.987	1.0
$(\gamma = 0.95)p_m$	0.122	0.558	1.0
p_p	0.407	0.961	1.0
Case 2			
$(\gamma = 0.9)p_m$	0.050	0.192	1.0
p_p	0.389	0.771	1.0
$(\gamma = 0.95)p_m$	0.050	0.090	1.0
p_p	0.235	0.581	1.0

In this location shift models, it still shows that the outlier proportion is better than the outlier mean. This further indicates the appropriateness of applying the outlier proportion in gene expression analysis.

With observation from Tomlins et al. (2005), it is interested to further investigate a power comparison when there is only a small percentage of outliers in distribution of Y. We evaluate the approximate power for the following two mixed distributions:

$$\text{Case A} : X \sim Laplace(0,1), Y \sim 0.7Lapace(0,1) + 0.3N(\theta,1)$$
$$\text{Case B} : X \sim t(5), Y \sim 0.7t(5) + 0.3Laplace(\theta,1)$$

The results are listed in Table 6.

Table 6 Approximate powers of outlier mean and outlier proportion

Power	$\theta = 2$	$\theta = 3$	$\theta = 4$
Case A			
$(\gamma = 0.85)p_m$	0.107	0.553	0.986
p_p	0.634	0.809	0.839
$(\gamma = 0.9)p_m$	0.086	0.252	0.504
p_p	0.565	0.815	0.878
$(\gamma = 0.95)p_m$	0.125	0.156	0.237
p_p	0.424	0.690	0.881
Case B			
$(\gamma = 0.85)p_m$	0.335	0.926	0.999
p_p	0.637	0.774	0.818
$(\gamma = 0.9)p_m$	0.185	0.640	0.987
p_p	0.623	0.805	0.858
$(\gamma = 0.95)p_m$	0.177	0.205	0.458
p_p	0.499	0.779	0.880

The approximate powers showing in Table 6 indicates that the outlier proportion is still a right choice in these distributional settings. Let us further consider one more distributional setting as

$$\text{Mixed t} : X \sim t(10), Y \sim 0.9t(10) + 0.1(\chi^2(10) + \theta)$$

for comparison. The results are displayed in Table 7.

Table 7 Approximate powers of outlier mean and outlier proportion for some mixed distributions

Power	$\theta = 2$	$\theta = 4$	$\theta = 6$
$(\gamma = 0.9)p_m$	0.879	0.895	0.905
p_p	0.873	0.953	0.960
$(\gamma = 0.95)p_m$	0.873	0.892	0.903
p_p	0.900	0.957	0.970

Both methods are with high powers in this distributional setting, however, the outlier proportion based test is still a better one.

5. Simulations Study

Suppose that now we have estimates $\hat{\beta}^*_X$ and $\hat{\sigma}_{\beta,X}$ for β^*_X and $\sigma_{\beta,X}$ respectively. A test based on quantile based outlier probability is stated in (4.1). Let $\hat{\beta}^*_X = \frac{1}{n_1}\sum_{i=1}^{n_1} I(X_i \geq \hat{F}_X^{-1}(\gamma))$, $\hat{\gamma}_{xy} = \frac{n_2}{n_1}$ and $\hat{\sigma}_{\beta,X} = \gamma(1-\gamma)\hat{\gamma}_{xy} + \hat{\beta}_Y(1-\hat{\beta}_Y)$. A question is that is this practically a level α test?

Theoretically the critical point z_{α^*} is 1.645 when we expect the significance level is 0.05. We conduct $m = 100,000$ replications to simulate the following simulated probablity

$$p_p = \frac{1}{m}\sum_{j=1}^{m} I(n_2^{1/2}(\frac{\hat{\beta}^*_Y - \hat{\beta}^*_X}{\hat{\sigma}_{\beta,X}}) \geq \ell) \tag{5.1}$$

When we set $\ell = 1.645$ (5.1) represents the probability of type I error. with some distributions been used and various sample sizes that the results are displayed in the following table.

Table 8. Simulated probability of type I error when $z_{\alpha^*} = 1.645$

sample size	$N(0,1)$	$t(10)$	$Laplace(0,1)$
$n = 30$	0.1156	0.1178	0.1174
$n = 50$	0.1328	0.1327	0.1341
$n = 100$	0.1133	0.1125	0.1134
$n = 200$	0.1258	0.1238	0.1243
$n = 500$	0.1197	0.1211	0.1198
$n = 1000$	0.1285	0.1273	0.1264
$n = 10,000$	0.1203	0.1213	0.1205
$n = 100,000$	0.1199	0.1201	0.1198

Unfortunately (4.1) is not practically a level 0.05 test. We now, for each distribution, choose a constant ℓ such that (5.1) is approximately equal to 0.05 and then further to simulate the power of (5.1) under case I and case II distributions as follows

Case I: $X \sim N(0,1)$ and $Y \sim 0.9N(0,1) + 0.1(\chi^2(10) + \theta)$

Case II: $X \sim t(10)$ and $Y \sim 0.9t(10) + 0.1(\chi^2(10) + \theta)$.

The results are displayed in Table 9 and Table 10.

Table 9. Power performance comparison by simulation (Case I)

	H_0	$\theta = 2$	$\theta = 4$	$\theta = 6$
$\gamma = 0.5$				
$p_m(c = 2.16)$	0.0527	0.9109	0.9303	0.9419
$p_p(c = 2.38)$	0.0516	0.9526	0.9671	0.9782
$\gamma = 0.55$				
$p_m(c = 2.23)$	0.0501	0.9167	0.9332	0.9443
$p_p(c = 2.44)$	0.0504	0.9569	0.9685	0.9868
$\gamma = 0.6$				
$p_m(c = 2.28)$	0.0504	0.9192	0.9355	0.9443
$p_p(c = 2.51)$	0.0508	0.9739	0.9828	0.9983
$\gamma = 0.65$				
$p_m(c = 2.37)$	0.0523	0.9227	0.9394	0.9474
$p_p(c = 2.62)$	0.0513	0.9647	0.9761	0.9802
$\gamma = 0.7$				
$p_m(c = 2.48)$	0.0513	0.9227	0.9387	0.9469
$p_p(c = 2.7)$	0.0496	0.9716	0.9826	0.9956
$\gamma = 0.75$				
$p_m(c = 2.74)$	0.0511	0.9225	0.9388	0.9493
$p_p(c = 2.78)$	0.0505	0.9623	0.9764	0.9890
$\gamma = 0.8$				
$p_m(c = 2.96)$	0.0526	0.9243	0.9388	0.9486
$p_p(c = 2.83)$	0.0510	0.9674	0.9891	0.9912
$\gamma = 0.85$				
$p_m(c = 3.8)$	0.0508	0.9169	0.9332	0.942
$p_p(c = 2.95)$	0.0513	0.9598	0.9864	0.9946
$\gamma = 0.9$				
$p_m(c = 4.81)$	0.051	0.9034	0.926	0.9368
$p_p(c = 3.19)$	0.0497	0.9580	0.9681	0.9767
$\gamma = 0.95$				
$p_m(c = 20.8)$	0.0502	0.6608	0.7208	0.7659
$p_p(c = 3.58)$	0.0506	0.8774	0.9105	0.9423

Table 10. Power performance comparison by simulation (Case II)

	H_0	$\theta = 2$	$\theta = 4$	$\theta = 6$
$\gamma = 0.5$				
$p_m(c = 2.35)$	0.0497	0.8881	0.9119	0.9281
$p_p(c = 2.18)$	0.0508	0.9636	0.9870	0.9958
$\gamma = 0.55$				
$p_m(c = 2.42)$	0.0501	0.8932	0.9166	0.9304
$p_p(c = 2.29)$	0.0506	0.9626	0.9863	0.9961
$\gamma = 0.6$				
$p_m(c = 2.47)$	0.0508	0.8918	0.9159	0.9336
$p_p(c = 2.35)$	0.0498	0.9540	0.9847	0.9953
$\gamma = 0.65$				
$p_m(c = 2.65)$	0.0492	0.8925	0.9167	0.9316
$p_p(c = 2.42)$	0.0509	0.9391	0.9794	0.9937
$\gamma = 0.7$				
$p_m(c = 2.75)$	0.051	0.8956	0.917	0.9344
$p_p(c = 2.5)$	0.0495	0.9501	0.9836	0.9951
$\gamma = 0.75$				
$p_m(c = 3.05)$	0.05	0.8924	0.9168	0.9308
$p_p(c = 2.57)$	0.0510	0.9207	0.9693	0.9900
$\gamma = 0.8$				
$p_m(c = 3.36)$	0.0494	0.8847	0.9109	0.9288
$p_p(c = 2.73)$	0.0497	0.9413	0.9786	0.9925
$\gamma = 0.85$				
$p_m(c = 4.25)$	0.0503	0.868	0.9001	0.9185
$p_p(c = 2.98)$	0.0502	0.9164	0.9485	0.9799
$\gamma = 0.9$				
$p_m(c = 5.45)$	0.0505	0.8366	0.8775	0.9019
$p_p(c = 3.21)$	0.0509	0.8936	0.9167	0.9549
$\gamma = 0.95$				
$p_m(c = 23)$	0.0502	0.5262	0.588	0.6364
$p_p(c = 3.45)$	0.0503	0.7492	0.8406	0.9041

The outlier mean and outlier proportion techniques are both powerful in these settings of distribution. More interestingly the outlier proportion is the more efficient method in this comparison.

6. Appendix

Three assumptions for the asymptotic representation of the sample outlier proportion test are as follows.

1. The limit $\gamma_{xy} = lim_{n_1, n_2 \to \infty} \frac{n_2}{n_1}$ exists.
2. Pobability density function f_X of distribution F_X is bounded away from

zero in neighborhoods of $F_X^{-1}(\alpha)$ for $\alpha \in (0,1)$ and the population cutoff point η.

3. Probability density function f_Y is bounded away from zero in a neighborhood of the population cutoff point η.

Proof of theorem 3.1.

From the expression of $\hat{\beta}_Y$ in (3.1), we have

$$n_2^{1/2}(\hat{\beta}_Y - \beta_Y) = -n_2^{-1/2}\sum_{i=1}^{n_2}[I(Y_i \le \eta + n_1^{-1/2}T_n) - I(Y_i \le \eta)] + n_2^{-1/2}\sum_{i=1}^{n_2}(I(Y_i \ge \eta) - \beta_Y). \tag{6.1}$$

where

$$T_n = n_1^{1/2}(\hat{\eta} - \eta) = n_1^{1/2}([2(\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha) - (2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha))].$$

With assumption (3), the key in this proof is that

$$\begin{aligned} & n_2^{-1/2}\sum_{i=1}^{n_2}[I(Y_i \le \eta + n_1^{-1/2}T_n) - I(Y_i \le \eta)] \\ & = -\gamma_{xy}^{1/2} f_Y(\eta) T_n + o_p(1) \end{aligned} \tag{6.2}$$

which may seen in Ruppert and Carroll (1980) and Chen and Chiang (1996). With the following representation of empirical quantile,

$$\begin{aligned} & \sqrt{n_1}(\hat{F}_X^{-1}(\alpha) - F_X^{-1}(\alpha)) \\ & = f_X^{-1}(F_X^{-1}(\alpha)) n_1^{-1/2}\sum_{i=1}^{n_1}[\alpha - I(X_i \le F_X^{-1}(\alpha))] + o_p(1), \end{aligned} \tag{6.3}$$

(see, for example, Ruppert and Carroll (1980)), a Bahadur representation of the outlier proportion is induced from (6.1)-(6.3) as

$$\begin{aligned} n_2^{1/2}(\hat{\beta}_Y - \beta_Y) = & n_1^{-1/2}\sum_{i=1}^{n_1}[(\alpha b_1 - (1-\alpha)b_2) I(X_i \le F_X^{-1}(\alpha)) + \alpha(b_1 + b_2) \\ & I(F_X^{-1}(\alpha) \le X_i \le F_X^{-1}(1-\alpha)) + (-(1-\alpha)b_1 + \alpha b_2) \\ & I(X_i \ge F_X^{-1}(1-\alpha))] + n_2^{-1/2}\sum_{i=1}^{n_2}[I(Y_i \ge \eta) - \beta_Y] + o_p(1). \end{aligned}$$

The asymptotic distribution in Theorem 3.1 is induced from the Central Limit Theorem. □

References

Chen, L.-A. and Chiang, Y. C. (1996). Symmetric type quantile and trimmed means for location and linear regression model. *Journal of Nonparametric Statistics.* 7, 171-185.

Chen, L.-A., Chen, Dung-Tsa and Chan, Wenyaw. (2010). The p Value for the Outlier Sum in Differential Gene Expression Analysis. *Biometrika*, 97, 246-253.

Ruppert, D. and Carroll, R.J. (1980). Trimmed least squares estimation in the linear model. *Journal of American Statistical Association* **75**, 828-838.

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

Tomlins, S. A., Rhodes, D. R., Perner, S., eta l. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644-648.

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