國立交通大學

統計學研究所

碩士論文

基因離群變異分析

Outlier Variance for Gene Expression Analysis

研 究 生：鄭炑煒

指導教授：陳鄰安　博士

中華民國九十九年六月

基因離群變異分析

Outlier Variance for Gene Expression Analysis

研究生：鄭炑煒　　Student：Ciou-Wei Jheng

指導教授：陳鄰安　博士　　Advisor：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

基因離群變異分析

學生：鄭秌煒　　　　　　　　　　　　指導教授：陳鄰安 博士

國立交通大學統計學研究所

摘要

在基因分析的研究中，尋找致病基因觀察值中的離群樣本是很重要的。離群和或離群平均可以檢測離群值分配的集中趨勢，但無法偵測其他特性如離中趨勢等。我們提出用離群變異數來做一個基因分的工具。我們推倒了離群變異的大樣本分配並由此建立一個檢定方法。這個檢定方法與離群平均也一起比較他們的檢力表現。另外我們提出由分位值來建立一個離群變異，這個離群變異量可以省去估計機率密度的困難。

誌　謝

感謝兩年來所上老師們的教導，尤其是陳鄰安老師的指導，才能順利完成這篇論文，從老師身上除了學到解決問題的能力之外，感受到老師對研究和教學上的熱情，會以老師為榜樣，在未來的日子，懷抱熱情並以求知的精神走下去。同時也要謝謝所上行政人員，在剛來統研所這個新環境時的幫助，還有同學們在課業上互相學習、競爭，這對我助益良多。很高興能在統研所和老師、同學一起度過兩年。

最後，感謝父母給於生活上的安定，讓我可以專心於學業上，不必為其他事操煩，有父母的付出、老師的教導、同學們的關愛讓我能順順利利完成學業。

鄭秋煒 謹至于

國立交通大學統計學研究所

中華民國九十九年六月

Abstract (in Chinese) i

Acknowledgements (in Chinese) ii

Contents

1. Introduction 1

2. Population Outlier variance 3

3. Sample Outlier Variance Based Gene Expression Analysis 5

4. Power Performance Evaluation 7

5. Outlier Variance With Quantile Based Cutoff Point 12

6. Simulation study 15

7. Appendix 18

References 19

List of Tables

Table 1. Table of parameter θ making outlier means equal........................ 4

Table 2. Ratio of population outlier variances.. 4

Table 3. Power performances of outlier mean and outlier variance tests.... 8

Table 4. Outlier mean ratio and outlier standard deviation ratio 9

Table 5. Power performances of outlier mean and outlier variance tests.. 10

Table 6. Power performance for outlier variance B test 11

Table 7. Power performances of new outlier mean and outlier variance tests 13

Table 8. Power performances of new outlier variance test....................... 13

Table 9. Power performances of new outlier variance test with implement of estimate 14

Table 10. Power performance comparison by simulation (Case I) 15

Table 11. Power performance comparison by simulation (Case II).......... 16

Outlier Variance for Gene Expression Analysis

SUMMARY

Discovering the existence of outliers in samples of influential genes 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 for the outlier data. We propose the outlier variance to measure the spreadness of the outlier data as an alternative tool for gene expression analysis. Large sample theory for this outlier variance is then developed and a test based this outlier variance is then compare with the outlier mean for their power performances. To avoid the inefficiency in estimating densities at tail quantiles for an estimate of asymptotic variance of the sample outlier variance, we further consider using the empirical quantile function as the sample cutoff point to propose an alternative outlier variance based 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); 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 outlier 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 (2009) proposed a new version of outlier sum and its corresponding outlier mean and developed its large sample theory that allows us to formulate the p value based on the asymptotic distribution. 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), gene expression analysis should consider to verify if the distributions of variables of disease group subjects and normal group subjects on the region excessing a specified cutoff point are identical. The outlier mean approach of Chen, Chen and Chiang (2010) can detect if the excessive means are different. We know that summarizing the outlier data by outlier mean or outlier sum may be efficient when the central tendencies of two outlier distributions on the excessive region are strongly different. However, it is known that it is not enough to detect just the shift in distributional mean when there exists of a distributional shift. So, it requires to measure other characteristics other than the central tendency of the outlier data as alternatives for detection of influential genes. Here, in this paper, we consider the measurement of outlier variance to detect the shift in distributional spread or dispersion as an alternative. Interestingly this study shows that using outlier variance in detection of influential genes is much more efficient than the outlier mean test.

In Section 2, we introduce the population outlier variance as a characteristic for detection of distributional shift. In Section 3, we study large sample property of the sample outlier variance and, in Section 4, we compare the power performances between the tests based on outlier mean and outlier variance. In Section 5, we propose an alternative outlier variance based test that avoids the estimation of densities and extreme quantiles for computing the test statistic.

2. Population Outlier Variances

Let X and Y be expression variables for group of normal subject and group of disease subject, respectively, with distribution functions F_X and F_Y. 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.

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, $\bar{Y}_{out} = (\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}))^{-1} \sum_{i=1}^{n_2} Y_i I(Y_i \geq \hat{\eta})$ for constructing a distribution based p value. Let η be the population counterpart of the sample cutoff point $\hat{\eta}$. Basically the idea in this series of study is to verify if the unknown population outlier means as follows

$$\mu_{X,out} = E(X|X \geq \eta) \text{ and } \mu_{Y,out} = E(Y|Y \geq \eta) \tag{2.1}$$

are the same. From now on, we suggest the population cutoff point of the form $\eta = 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)$. For stimulating the approach of outlier variance, we show that testing equality of outlier means are not sufficient for verifying equality of two distributions on excess region.

Consider the following distribution settings:

$$X \sim N(0,1), Y \sim 0.9N(\delta,1) + 0.1N(\theta,\sigma^2). \tag{2.2}$$

For given α and δ, we display, in the following table, the parameter values of θ that induces $\mu_{X,out} = \mu_{Y,out}$.

Table 1. Table of parameter θ making $\mu_{X,out} = \mu_{Y,out}$ when $\sigma = 0.5$

α	$\delta = -0.1$	-0.15	-0.20	-0.25	-0.5
$\alpha = 0.05$	3.952	3.952	3.952	3.952	3.952
$\alpha = 0.1$	3.182	3.182	3.182	3.182	3.182
$\alpha = 0.2$	2.278	2.280	2.280	2.281	2.280
$\alpha = 0.3$	1.674	1.682	1.688	1.692	1.693
$\alpha = 0.4$	1.227	1.254	1.275	1.291	1.321

The existence of equal population outlier means indicates that the outlier mean approach can not solve this problem that two distributions on regions exceeding η are definitely un-equal.

Known not enough to detect the difference in outlier means, a natural alternative is to infer population outlier variance

$$\sigma^2_{Y,out} = \beta_Y^{-1} E\{(Y - \mu_{Y,out})^2 I(Y \geq \eta)\} \tag{2.2}$$

for variable Y where $\beta_Y = P(Y \geq \eta)$. This is to measure the degree to which outlier observations are (or are not) clustered around the outlier mean $\mu_{Y,out}$. This measure of a truncated dispersion do not take every observation into account. The idea behind this approach is to verify if $\sigma^2_{Y,out}$ is different from the population outlier variance for variable X as

$$\sigma^2_{X,out} = \beta_X^{-1} E\{(X - \mu_{X,out})^2 I(X \geq \eta)\} \tag{2.3}$$

where $\beta_X = P(X \geq \eta)$.

We design situations that the population outlier means are identical for comparing their corresponding population outlier variances. For the following distribution setting:

$$X \sim N(0,1) \text{ and } Y \sim 0.9N(-0.1,1) + 0.1N(\theta, \sigma^2), \tag{2.4}$$

we choose parameters σ and θ so that their corresponding outlier means are identical and then to compute the ratios $\sigma^2_{X,out}/\sigma^2_{Y,out}$. The results of population outlier variances are listed in Table 2.

Table 2. Ratio of population outlier variances $\sigma^2_{X,out}/\sigma^2_{Y,out}$ when population outlier means are identical

α	$\sigma^2 = 0.01$	0.05	0.15
0.05	3.944 $(\theta = 5.115)$	1.710 $(\theta = 4.961)$	1.258 $(\theta = 4.468)$
0.1	5.301 $(\theta = 4.074)$	1.963 $(\theta = 3.971)$	1.349 $(\theta = 3.591)$
0.2	6.920 $(\theta = 2.845)$	2.387 $(\theta = 2.798)$	1.486 $(\theta = 2.557)$
0.3	3.016 $(\theta = 2.007)$	2.010 $(\theta = 1.991)$	1.414 $(\theta = 1.853)$
0.4	1.613 $(\theta = 1.375)$	1.467 $(\theta = 1.373)$	1.262 $(\theta = 1.320)$

When the ratio is 1, the population outlier variances are also identical and there is no chance to detect a distributional difference through outlier variance approach. Interestingly, the computed ratios in Table 2 for that their corresponding outlier means are identical are all larger than 1 indicating that a test based on outlier variance has an addtional chance for observing distributional difference.

3. Sample Outlier Variance Based Gene Expression Analysis

Let $\hat{F}_X^{-1}$ be the empirical quantile function for estimating population quantile function F_X^{-1} and we estimate the cutoff point η by $\hat{\eta} = 2\hat{F}_X^{-1}(1 - \alpha) - \hat{F}_X^{-1}(\alpha)$ for some $0 < \alpha < 0.5$. We propose a sample outlier variance as

$$\begin{aligned} S^2_{Y,out} =& (\sum_{i=1}^{n_2} I\{Y_i \geq 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha)\})^{-1} \\ & \sum_{i=1}^{n_2} (Y_i - \bar{Y}_{out})^2 I\{Y_i \geq 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha)\} \end{aligned} \tag{3.1}$$

This statistic using those observations from disease group exceeding the sample cutoff point does provide a concise summary of dispersion for the outlier data.

Let us now display the asymptotic properties of the outlier variance $S^2_{Y,out}$. A Bahadur representation of $S^2_{Y,out}$ and its asymptotic distribution are stated in the follwoing theorem.

Theorem 3.1. Suppose that assumptions (A_2) and (A_3) in the Appendix are true.

(a) A Bahadur representation of the outlier variance is

$$\begin{aligned}
&n_2^{1/2}(S^2_{Y,out} - \sigma^2_{Y,out}) \\
&\quad= [-(1-\alpha)af_X^{-1}\{F_X^{-1}(\alpha)\} + 2\alpha a f_X^{-1}\{F_X^{-1}(1-\alpha)\}]n_1^{-1/2}\sum_{i=1}^{n_1} I\{X_i \le F_X^{-1}(\alpha)\} \\
&\qquad + [\alpha a f_X^{-1}\{F_X^{-1}(\alpha)\} + 2\alpha a f_X^{-1}\{F_X^{-1}(1-\alpha)\}]n_1^{-1/2}\sum_{i=1}^{n_1} I\{F_X^{-1}(\alpha) \le X_i \le F_X^{-1}(1-\alpha)\} \\
&\qquad + [\alpha a f_X^{-1}\{F_X^{-1}(\alpha)\} - 2(1-\alpha)af_X^{-1}\{F_X^{-1}(1-\alpha)\}]n_1^{-1/2}\sum_{i=1}^{n_1} I\{X_i \ge F_X^{-1}(1-\alpha)\} \\
&\qquad + \beta_Y^{-1}n_2^{-1/2}\sum_{i=1}^{n_2}\{(Y_i - \mu_{Y,out})^2 - \sigma^2_{Y,out}\}I(Y_i \ge \eta) + o_p(1),
\end{aligned}$$

where

$$a = \beta_Y^{-1}\{(\eta - \mu_{Y,out})^2 - \sigma^2_{Y,out}\}f_Y(\eta)\gamma_{xy}^{1/2}$$

(b) $n_2^{1/2}(S^2_{Y,out} - \sigma^2_{Y,out})$ converges in distribution to $N(0, v_Y)$ where

$$\begin{aligned}
v_Y =& \alpha[-(1-\alpha)af_X^{-1}\{F_X^{-1}(\alpha)\} + 2\alpha a f_X^{-1}\{F_X^{-1}(1-\alpha)\}]^2 \\
&+ (1-2\alpha)[\alpha a f_X^{-1}\{F_X^{-1}(\alpha)\} + 2\alpha a f_X^{-1}\{F_X^{-1}(1-\alpha)\}]^2 \\
&+ \alpha[\alpha a f_X^{-1}\{F_X^{-1}(\alpha)\} - 2(1-\alpha)af_X^{-1}\{F_X^{-1}(1-\alpha)\}]^2 \\
&+ \beta_Y^{-2}E[\{(Y - \mu_{Y,out})^2 - \sigma^2_{Y,out}\}^2 I(Y \ge \eta)].
\end{aligned}$$

Following Theorem 3.1, the following variable

$$n_2^{1/2}v_Y^{-1/2}(S^2_{Y,out} - \sigma^2_{Y,out})$$

converge to $N(0,1)$ in distribution. For testing if the distributions of Y and X by outlier variance, we are testing this hypothesis by comparing two outlier variances $\sigma^2_{Y,out}$ and $\sigma^2_{X,out}$ and then we should choose $\hat{\sigma}^2_{X,out}$, an estimate of $\sigma^2_{X,out}$ and an estimate $\hat{v}$ to form a test statistic

$$n_2^{1/2}\hat{v}^{-1/2}(S^2_{Y,out} - \hat{\sigma}^2_{X,out}) \tag{3.2}$$

However, in literature, there are two choices for $\hat{v}$, it can be an estimate of v_Y or an estimate of v_X with

$$\begin{aligned} v_X =& \alpha[-(1-\alpha)af_X^{-1}\{F_X^{-1}(\alpha)\} + 2\alpha af_X^{-1}\{F_X^{-1}(1-\alpha)\}]^2 \\ &+ (1-2\alpha)[\alpha af_X^{-1}\{F_X^{-1}(\alpha)\} + 2\alpha af_X^{-1}\{F_X^{-1}(1-\alpha)\}]^2 \\ &+ \alpha[\alpha af_X^{-1}\{F_X^{-1}(\alpha)\} - 2(1-\alpha)af_X^{-1}\{F_X^{-1}(1-\alpha)\}]^2 \\ &+ \beta_X^{-2}E[\{(X-\mu_{X,out})^2 - \sigma^2_{X,out}\}^2 I(X \geq \eta)]. \end{aligned}$$

Hence there are two tests available based on outlier variance as

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}\hat{v}_Y^{-1/2}(S^2_{Y,out} - \hat{\sigma}^2_{X,out}) \geq z_\alpha. \tag{3.3}$$

and

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}\hat{v}_X^{-1/2}(S^2_{Y,out} - \hat{\sigma}^2_{X,out}) \geq z_\alpha. \tag{3.4}$$

where $\hat{v}_Y$ and $\hat{v}_X$ are, respectively, estimates of v_Y and v_X.

But how good are these two tests? An important part for an evaluation is to verify its power performance when there exists positive outlier in data of disease group.

4. Power Performance Evaluation

We evaluate the powers of test (3.3) for several distributional settings. An approximate power function for this test may be derived as follows:

$$\begin{aligned} p_{vy} &= P_{F_Y}\{n_2^{1/2}\hat{v}_Y^{-1/2}(S^2_{Y,out} - \hat{\sigma}^2_{X,out}) \geq z_\alpha\} \\ &= P_{F_Y}\{n_2^{1/2}v_Y^{-1/2}(S^2_{Y,out} - \sigma^2_{Y,out}) \geq v_Y^{-1/2}(z_\alpha \hat{v}_Y^{1/2} + n_2^{1/2}(\hat{\sigma}^2_{X,out} - \sigma^2_{Y,out}))\} \\ &\approx P\{Z \geq v_Y^{-1/2}(z_\alpha \hat{v}_Y^{1/2} + n_2^{1/2}(\hat{\sigma}^2_{X,out} - \sigma^2_{Y,out}))\} \\ &\approx P\{Z \geq z_\alpha + n_2^{1/2}v_Y^{-1/2}(\sigma^2_{X,out} - \sigma^2_{Y,out})\} \end{aligned} \tag{4.1}$$

Similarly, the test of (3.4) has an approximate power as

$$p_v \approx P\{Z \geq z_\alpha(\frac{v_X}{v_Y})^{1/2} + n_2^{1/2}v_Y^{-1/2}(\sigma^2_{X,out} - \sigma^2_{Y,out})\} \tag{4.2}$$

We are ready to study asymptotic powr for comparison with outlier mean where the following distributional settings

$$\text{Normal: } X \sim N(0,1), Y \sim N(\theta, \sigma^2)$$

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

$$\text{Mixed } \chi^2 : X \sim N(0,1), Y \sim 0.9N(0,1) + 0.1(\chi^2(10) + \theta)$$

are considered where p_m and p_v represent, respectively, the powers for outlier mean and outlier variance.tests.

Table 3. Power performances of outlier mean and outlier variance tests

α	$\theta = 2$	$\theta = 4$	$\theta = 6$
Normal			
$\alpha = 0.1, p_m$	0.700	0.968	1
p_v	0.044	0.161	0.517
$\alpha = 0.2, p_m$	0.926	0.999	1
p_v	0.918	0.987	0.998
$\alpha = 0.3, p_m$	0.984	1	1
p_v	0.980	0.998	0.999
$\alpha = 0.4, p_m$	0.996	1	1
p_v	0.992	0.999	0.999
Mixed normal			
$\alpha = 0.1, p_m$	0.232	0.421	0.670
p_v	0.291	0.372	0.506
$\alpha = 0.2, p_m$	0.282	0.492	0.726
p_v	0.681	0.780	0.852
$\alpha = 0.3, p_m$	0.212	0.324	0.434
p_v	0.755	0.864	0.937
$\alpha = 0.4, p_m$	0.177	0.253	0.316
p_v	0.762	0.860	0.925
Mixed χ^2			
$\alpha = 0.1, p_m$	0.924	0.985	0.998
p_v	0.767	0.768	0.768
$\alpha = 0.2, p_m$	0.929	0.975	0.991
p_v	0.838	0.831	0.818
$\alpha = 0.25, p_m$	0.773	0.823	0.854
p_v	0.881	0.880	0.874
$\alpha = 0.4, p_m$	0.383	0.397	0.407
p_v	0.945	0.965	0.978

We have several comments for the results in this table:

(a). The power increases as location parameter θ increses indicating that when there are more wide outliers the outlier means and the outlier variance are more efficient in detection the existence of distributional difference.

(b). Consider the location shift models (Normal, Laplace and t distributions). The outlier means and outlier variances with cutoff point of larger percentage α are relatively more powerful. Hence, choosing smaller cutoff point (larger α) is advisable for application when there is a difference in location parameter. However, in this distributional settings, the outlier variance with smaller α (0.1) is not a poweful one.

(c). For a distributional difference of only a small proportion of sample points (Mixed normal), the outlier mean with all percentages are inefficient with small powers. However, the outlier variances are relatively more powerful especially for larger $\alpha's$.

(d). In an over all comparison, since there is specific distribution being known in nonparametric hypothesis testing and it is supposed to have only a small proportion of outliers in the influential genes, the outlier variance with cutoff point of α larger than 0.25 is recommended.

For verification of the above conclusions, we consider the mixed normal distribution case with $\sigma = 3$ to compute the following ratios

$$\pi_m = \mu^{-1}_{X,out}\mu_{Y,out}, \pi_v = \sigma^{-1}_{X,out}\sigma_{Y,out}.$$

Table 4. Outlier mean ratio and outlier standard deviation ratio

α	$\theta = 2$	$\theta = 4$	$\theta = 6$
$\alpha = 0.05, \pi_m$	1.273	1.370	1.514
π_v	7.369	9.003	10.98
$\alpha = 0.1, \pi_m$	1.391	1.543	1.767
π_v	6.744	8.251	9.972
$\alpha = 0.2, \pi_m$	1.593	1.880	2.278
π_v	5.821	7.178	8.622
$\alpha = 0.3, \pi_m$	1.549	1.931	2.420
π_v	4.600	6.163	7.912
$\alpha = 0.4, \pi_m$	1.430	1.770	2.192
π_v	3.102	4.379	5.867

The ratios of population outlier variance are much larger than the corresponding ratios of population outlier means that provides a message for the efficiencies in powers obtained from the outlier variance test.

We further consider the following distributional settings for comparison:

Model I: $X \sim Laplace(0,1), Y \sim 0.9Laplace(0,1) + 0.1Laplace(\theta, 10)$

Model II: $X \sim t(10), Y \sim 0.9t(10) + 0.1(\chi^2(10) + \theta)$

and the results are listed in Table 5.

Table 5. Power performances of outlier mean and outlier variance tests

α	$\theta = 2$	$\theta = 4$	$\theta = 6$
Model I			
$\alpha = 0.1, p_m$	0.2289	0.260	0.3009
p_v	0.6293	0.6418	0.6549
$\alpha = 0.2, p_m$	0.2172	0.2511	0.2939
p_v	0.6645	0.6799	0.6958
$\alpha = 0.25, p_m$	0.2088	0.2403	0.277
p_v	0.673	0.6899	0.7079
$\alpha = 0.4, p_m$	0.1971	0.2222	0.2485
p_v	0.6851	0.7044	0.7251
Model II			
$\alpha = 0.1, p_m$	0.872	0.968	0.995
p_v	0.527	0.536	0.542
$\alpha = 0.2, p_m$	0.863	0.928	0.96
p_v	0.823	0.828	0.824
$\alpha = 0.25, p_m$	0.719	0.771	0.803
p_v	0.887	0.902	0.908
$\alpha = 0.3, p_m$	0.542	0.571	0.591
p_v	0.938	0.963	0.978
$\alpha = 0.4, p_m$	0.385	0.401	0.412
p_v	0.942	0.963	0.977

We have several comments drawing from the results in the above table:

(a) On Model I, the two methods are both not very powerful in detection of influential genes. However, the outlier variance seems to be much more better.

(b) On Model II, the two methods are more powerful in the purpose. The outlier means show better in smaller $\alpha's$ and the outlier variances show better in larger $\alpha's$. This provides a guidence for users in choosing appropriate outlier mean and outlier variance.

(c) In overall evaluation, the outlier variance method seems to be quite robust of with powers larger than 0.5 in all situations. This observation indicates that the outlier variance approach for gene expression analysis seems to be desirable in application since, in gene expressions, the underlying distribution is generally non-normal.

Suppose that an outlier mean test is conducted and it results in acceptance of equal outlier means. From Table 2, it is seen that larger values of $\sigma^2_{Y,out}$ than $\sigma^2_{X,out}$ shows that an left handed one sided outlier variance based test in this situation is appropriate. We propose the following left handed outlier variance test:

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}\hat{v_Y}^{-1/2}(S^2_{Y,out} - \hat{\sigma}^2_{X,out}) \leq -z_\alpha$$

An approximate power function may be derived as follows:

$$p_{vB} \approx P\{Z \leq -z_\alpha + n_2^{1/2} v_Y^{-1/2}(\sigma^2_{X,out} - \sigma^2_{Y,out})\}.$$

We list the computed powers for distributions of (2.4) with $\sigma = 1$ in Table 6.

Table 6. Power performance for outlier variance B test when outlier means are identical

α	$\delta = -0.01$	-0.1	-0.5	-1.0	-1.5
p_m	0.05	0.05	0.05	0.05	0.05
p_{vB}					
0.05	0.352	0.352	0.352	0.352	0.352
0.1	0.970	0.971	0.973	0.974	0.974
0.2	0.820	0.917	1	1	1
0.3	0.233	0.284	0.686	0.999	1
0.4	0.135	0.160	0.302	0.668	0.997

This observation shows that when we accept the null hypothesis of equal outlier means through the outlier mean based test it suggests to further test the outlier variance by left hand one sided test.

5. Outlier Variance With Quantile Based Cutoff Point

The test based on outlier variance $S^2_{Y,out}$ requires to estimate density points $f_X\{F_X^{-1}(\alpha)\}$ and $f_X\{F_X^{-1}(1-\alpha)\}$ ((b) of Theorem 3.1). There is generally no satisfactory solution for this estimation unless the sample sizes are large enough. Here we consider an alternative design of the cutoff point for a new outlier variance. We let $\eta = F_X^{-1}(\gamma)$ and $\hat{\eta} = \hat{F}_X^{-1}(\gamma)$. In the following, we state the large sample theory for this outlier variance.

Theorem 5.1. Suppose that assumptions (A_2) and (A_3) in the Appendix are true.

(a) A Bahadur representation of the outlier variance is

$$\begin{aligned} n_2^{1/2}&(S^2_{Y,out}-\sigma^2_{Y,out}) \\ =& -\beta_Y^{-1}\{(F_X^{-1}(\gamma)-\mu_{Y,out})^2-\sigma^2_{Y,out}\}f_Y(F_X^{-1}(\gamma))f_X^{-1}(F_X^{-1}(\gamma))(\gamma_{xy})^{1/2}n_1^{-1/2}\sum_{i=1}^{n_1}(\gamma - I(X_i \\ &\le F_X^{-1}(\gamma))) + \beta_Y^{-1}n_2^{-1/2}\sum_{i=1}^{n_2}[(Y_i-\mu_{Y,out})^2-\sigma^2_{Y,out}]I(Y_i \ge F_X^{-1}(\gamma)) + o_p(1). \end{aligned}$$

(b) $n_2^{1/2}(S^2_{Y,out}-\sigma^2_{Y,out})$ converges in distribution to $N(0, v_Y)$ where

$$\begin{aligned} v_Y =& \gamma(1-\gamma)\beta_Y^{-2}\{(F_X^{-1}(\gamma)-\mu_{Y,out})^2-\sigma^2_{Y,out}\}^2(f_Y(F_X^{-1}(\gamma))f_X^{-1}(F_X^{-1}(\gamma)))^2\gamma_{xy} \\ &+ \beta_Y^{-2}E[\{(Y-\mu_{Y,out})^2-\sigma^2_{Y,out}\}^2 I(Y \ge F_X^{-1}(\gamma))]. \end{aligned}$$

We may consider asymptotic variance v_Y under the assumption that Y and X have the same distribution setting by

$$v_X = \beta_X^{-2}[\gamma(1-\gamma)\{(F_X^{-1}(\gamma)-\mu_{X,out})^2-\sigma^2_{X,out}\}^2\gamma_{xy}+E[\{(X-\mu_{X,out})^2-\sigma^2_{X,out}\}^2 I(X \ge F_X^{-1}(\gamma))]].$$

This variance save the effort in estimating unknown density points. An outlier variance based test may be stated as

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}\hat{v}_X^{-1/2}(S^2_{Y,out}-\hat{\sigma}^2_{X,out}) \ge z_\alpha. \tag{5.1}$$

It is interesting to study the power performance of this outlier variance based nonparametric test for models with only a small proportion of the data in disease group been shifted. Observed from Tomblins et al. (2005), this happen in some regular cancer genes. We first consider the mixed normal distribution.

Table 7. Power performances of new outlier mean and outlier variance test for mixed normal distribution

	$\theta = 2$	$\theta = 4$	$\theta = 6$
$\rho = 0.1$			
$\gamma = 0.8, p_m$	0.523	0.701	0.805
p_v	0.764	0.864	0.930
$\gamma = 0.85, p_m$	0.541	0.709	0.809
p_v	0.764	0.866	0.936
$\gamma = 0.9, p_m$	0.557	0.716	0.812
p_v	0.762	0.868	0.941
$\gamma = 0.95, p_m$	0.563	0.711	0.810
p_v	0.754	0.861	0.934
$\rho = 0.2$			
$\gamma = 0.8, p_m$	0.710	0.866	0.936
p_v	0.871	0.957	0.991
$\gamma = 0.85, p_m$	0.710	0.863	0.935
p_v	0.867	0.955	0.990
$\gamma = 0.9, p_m$	0.705	0.857	0.933
p_v	0.860	0.950	0.986
$\gamma = 0.95, p_m$	0.679	0.837	0.924
p_v	0.840	0.932	0.971

The use of quantile to construct the cutoff point still shows the advantage better performance by the outlier variance approach. Also, by comparing with the results in Table 3, the use of quantile for constructing the cutoff point is competitive with using quantile combination for constructing the cutoff point.

We further consider the following distribution settings:

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

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

for investigation and the results are displayed in Table 8.

Table 8. Power performances of new outlier variance test

	$\theta = 2$	$\theta = 4$	$\theta = 6$
Case I			
$\gamma = 0.8, p_m$	0.895	0.907	0.916
p_v	0.950	0.971	0.983
$\gamma = 0.85, p_m$	0.896	0.908	0.916
p_v	0.955	0.976	0.989
$\gamma = 0.9p_m$	0.897	0.909	0.917
p_v	0.957	0.979	0.991
$\gamma = 0.95, p_m$	0.899	0.911	0.919
p_v	0.935	0.953	0.964
Case II			
$\gamma = 0.8, p_m$	0.881	0.896	0.907
p_v	0.946	0.968	0.982
$\gamma = 0.85, p_m$	0.880	0.896	0.906
p_v	0.950	0.973	0.987
$\gamma = 0.9p_m$	0.879	0.895	0.905
p_v	0.950	0.975	0.989
$\gamma = 0.95, p_m$	0.873	0.892	0.903
p_v	0.921	0.943	0.957

The outlier mean and outlier variance are performed quite well in models Case I and Case II. This support the observation by Tomlins et al. (2005) that when outliers exist in influential genes the gene expression techniques should take the outliers into more consideration. The following table is to display the results with designing the use of v_Y for constructing the quantile based outlier variance.

Table 9. Power performances of new outlier variance test with implemet of estimate v_Y $(\rho = 0.1)$

	$\theta = 2$	$\theta = 4$	$\theta = 6$
Case I			
$\gamma = 0.8$	0.507	0.602	0.687
$\gamma = 0.85$	0.527	0.637	0.739
$\gamma = 0.9$	0.535	0.656	0.767
$\gamma = 0.95$	0.455	0.519	0.566
Case II			
$\gamma = 0.8$	0.500	0.595	0.682
$\gamma = 0.85$	0.516	0.627	0.731
$\gamma = 0.9$	0.518	0.639	0.752
$\gamma = 0.95$	0.436	0.499	0.548

The results showed above are less powerful than the implement of estimate v_X. The test (5.1) that is based on v_Y is not with benefit of avoiding the estimation of density points.

6. Simulation study

We consider a simulation study in the comparison of the quantile based outlier variance with the outlier mean and classical two sample t test. Defining estimates

$$\hat{\beta}_X = \frac{1}{n_1}\sum_{i=1}^{n_1} I(X_i \geq \hat{F}_X^{-1}(\gamma)), \hat{\mu}_{X,out} = \frac{\sum_{i=1}^{n_1} X_i I(X_i \geq \hat{F}_X^{-1}(\gamma))}{\sum_{i=1}^{n_1} I(X_i \geq \hat{F}_X^{-1}(\gamma))},$$

$$\hat{\sigma}^2_{X,out} = \frac{\sum_{i=1}^{n_1}(X_i - \hat{\mu}_{X,out})^2 I(X_i \geq \hat{F}_X^{-1}(\gamma))}{\sum_{i=1}^{n_1} I(X_i \geq \hat{F}_X^{-1}(\gamma))},$$

Suppose that we have test statistic $T = n_2^{1/2}\hat{v}_X^{-1/2}(S^2_{Y,out} - \hat{\sigma}^2_{X,out})$ and its observation at ith replication is T_i. We search constant c such that

$$0.05 \approx \frac{1}{m}\sum_{i=1}^{m} I(T_i \geq c | H_0 : F_Y = F_X) \qquad (6.1)$$

and then apply this constant as the cutoff point to evaluate the following power

$$\frac{1}{m}\sum_{i=1}^{m} I(T_i \geq c | H_1).$$

In the follwoing tables, we list the simulated probability under H_0 at the simulated constant c and the simulated powers under distributions Case I and Case II.

Table 10. Power performance comparison by simulation (Case I, $n_1 = n_2 = 30$)

	H_0	$\theta = 2$	$\theta = 4$	$\theta = 6$
p_t	0.049	0.459	0.482	0.504
$\gamma = 0.5$				
$p_m(c = 2.16)$	0.0527	0.9109	0.9303	0.9419
$p_v(c = 4.48)$	0.051	0.9569	0.9597	0.9597
$\gamma = 0.55$				
$p_m(c = 2.23)$	0.0501	0.9167	0.9332	0.9443
$p_v(c = 4.98)$	0.0508	0.9588	0.959	0.9592
$\gamma = 0.6$				
$p_m(c = 2.28)$	0.0504	0.9192	0.9355	0.9443
$p_v(c = 5.35)$	0.0506	0.9581	0.9596	0.9596
$\gamma = 0.65$				
$p_m(c = 2.37)$	0.0523	0.9227	0.9394	0.9474
$p_v(c = 6.1)$	0.0508	0.9582	0.9602	0.9599
$\gamma = 0.7$				
$p_m(c = 2.48)$	0.0513	0.9227	0.9387	0.9469
$p_v(c = 6.68)$	0.0503	0.9581	0.9599	0.9493
$\gamma = 0.75$				
$p_m(c = 2.74)$	0.0511	0.9225	0.9388	0.9493
$p_v(c = 8.23)$	0.0492	0.9562	0.9589	0.9606
$\gamma = 0.8$				
$p_m(c = 2.96)$	0.0526	0.9243	0.9388	0.9486
$p_v(c = 9.5)$	0.0498	0.9559	0.9576	0.9589
$\gamma = 0.85$				
$p_m(c = 3.8)$	0.0508	0.9169	0.9332	0.942
$p_v(c = 13.9)$	0.0519	0.9496	0.952	0.9528
$\gamma = 0.9$				
$p_m(c = 4.81)$	0.051	0.9034	0.926	0.9368
$p_v(c = 21.3)$	0.0507	0.9388	0.944	0.9444
$\gamma = 0.95$				
$p_m(c = 20.8)$	0.0502	0.6608	0.7208	0.7659
$p_v(c = 200)$	0.2932	0.9296	0.9315	0.9293

Table 11. Power performance comparison by simulation (Case II, $n_1 = n_2 = 30$)

	H_0	$\theta = 2$	$\theta = 4$	$\theta = 6$
p_t	0.049	0.448	0.473	0.492
$\gamma = 0.5$				
$p_m(c = 2.35)$	0.0497	0.8881	0.9119	0.9281
$p_v(c = 6.35)$	0.05	0.9506	0.9562	0.9582
$\gamma = 0.55$				
$p_m(c = 2.42)$	0.0501	0.8932	0.9166	0.9304
$p_v(c = 7.18)$	0.0501	0.9496	0.9565	0.9587
$\gamma = 0.6$				
$p_m(c = 2.47)$	0.0508	0.8918	0.9159	0.9336
$p_v(c = 7.64)$	0.0509	0.9483	0.9552	0.9596
$\gamma = 0.65$				
$p_m(c = 2.65)$	0.0492	0.8925	0.9167	0.9316
$p_v(c = 8.83)$	0.0507	0.9467	0.9545	0.9582
$\gamma = 0.7$				
$p_m(c = 2.75)$	0.051	0.8956	0.917	0.9344
$p_v(c = 9.8)$	0.0497	0.9465	0.9541	0.9592
$\gamma = 0.75$				
$p_m(c = 3.05)$	0.05	0.8924	0.9168	0.9308
$p_v(c = 11.87)$	0.0508	0.9429	0.9532	0.9572
$\gamma = 0.8$				
$p_m(c = 3.36)$	0.0494	0.8847	0.9109	0.9288
$p_v(c = 13.67)$	0.0496	0.9378	0.95	0.9548
$\gamma = 0.85$				
$p_m(c = 4.25)$	0.0503	0.868	0.9001	0.9185
$p_v(c = 20.5)$	0.0505	0.9221	0.9366	0.9439
$\gamma = 0.9$				
$p_m(c = 5.45)$	0.0505	0.8366	0.8775	0.9019
$p_v(c = 31)$	0.0506	0.8935	0.9152	0.9258
$\gamma = 0.95$				
$p_m(c = 23)$	0.0502	0.5262	0.588	0.6364
$p_v(c = 200)$	0.3004	0.9289	0.9281	0.927

We have several comments on these simulated results:

(a) The outlier mean and outlier variance techniques are both more powerful than the two samples t test showing that applying all data for inferences is not appropriate.

(b) More interestingly the outlier variance is the most efficient method in this comparison.

7. Appendix

Three assumptions for the two sample outlier variance test are as follows.

ASSUMPTION 1: *The limit* $\gamma = lim_{n_1,n_2\to\infty} n_1^{-1}n_2$ *exists.*

ASSUMPTION 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*.*

ASSUMPTION 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: With Assumption 2, a representation of $\hat{F}_X^{-1}(\alpha)$ such as

$$n_1^{1/2}\{\hat{F}_X^{-1}(\alpha) - F_X^{-1}(\alpha)\} = f_X^{-1}\{F_X^{-1}(\alpha)\}n_1^{-1/2}\sum_{i=1}^{n_1}[\alpha - I\{X_i \leq F_X^{-1}(\alpha)\}] + o_p(1), \tag{7.1}$$

implies that $\hat{\eta} = 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha)$ satisfies $T = n_1^{1/2}(\hat{\eta} - \eta) = O_p(1)$ (Ruppert & Carroll, 1980). First, we can rewrite the sample outlier variance as

$$S^2_{Y,out} = (\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}))^{-1}\sum_{i=1}^{n_2}[(Y_i - \mu_{Y,out})^2 + (\bar{Y}_{out} - \mu_{Y,out})^2]. \tag{7.2}$$

A Bahadur representation of $\bar{Y}_{out}$ in Chen, Chen and Chan (2009) indicates that $n_2^{1/2}(\bar{Y}_{out} - \mu_{Y,out}) = O_p(1)$ which leads to the fact that $n_2^{1/2}(\bar{Y}_{out} - \mu_{Y,out})^2 = o_p(1)$ and we may write the sample outlier variance in the following

$$\begin{aligned} & n_2^{1/2}(S^2_{Y,out} - \sigma^2_{Y,out}) \\ =& n_2^{1/2}(\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}))^{-1}\sum_{i=1}^{n_2}[(Y_i - \mu_{Y,out})^2 - \sigma^2_{Y,out}][I(Y_i \geq \eta + n_2^{-1/2}T) - I(Y_i \geq \eta)] \\ & + n_2^{1/2}(\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}))^{-1}\sum_{i=1}^{n_2}[(Y_i - \mu_{Y,out})^2 - \sigma^2_{Y,out}]I(Y_i \geq \eta) + o_p(1). \end{aligned} \tag{7.3}$$

With (A1), Assumptions 1 and 3, and techniques from Ruppert & Carroll (1980) and Chen & Chiang (1996), we may see that

$$\begin{aligned} & n_2^{-1/2}\sum_{i=1}^{n_2}[(Y_i - \mu_{Y,out})^2 - \sigma^2_{Y,out}][I(Y_i \geq \eta + n_2^{-1/2}T) - I(Y_i \geq \eta)] \\ & = -\{(\eta - \mu_{Y,out})^2 - \sigma^2_{Y,out}\}f_Y(\eta)T + o_p(1). \end{aligned} \tag{7.4}$$

The first term on the right hand side of (7.2) may be formulated as

$$n_2^{-1}\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}) = n_2^{-1}\sum_{i=1}^{n_2} I(Y_i \geq \eta) + o_p(1). \tag{7.5}$$

Plugging (6.1) into (7.4), the theorem is followed from (7.3)-(7.5). □
The proof of Theorem 5.1 is quite similar to the above one so tat we skip it.

References

Chen, L.-A., Chen, Dung-Tsa and Chan, Wenyaw. (2008). The p Value for the Outlier Sum in Differential Gene Expression Analysis. *Biometrika*, 97, 246-253.

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.

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.