國立交通大學

統計學研究所

碩士論文

無母數離群平均之基因分析

Nonparametric Outlier Mean for Gene Expression Analysis

研究生：游雅芳

指導教授：陳鄰安　博士

中華民國九十八年六月

無母數離群平均之基因分析

Nonparametric Outlier Mean for Gene Expression Analysis

研 究 生：游雅芳	Student: Ya-Fang You
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 2009

Hsinchu, Taiwan, Republic of China

中華民國 九十八 年 六 月

無母數離群平均之基因分析

學生：游雅芳　　　　　　　　　　　　　　　　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

離群平均用於檢定整個分配的偏移時有不錯的檢定力，然而部分分配偏移時放大了離群平均值的變異數，導致檢定力大幅下降，而這部分分配偏移的情況在癌症的研究上頻繁可見。傳統的統計方法使用好的資料來做統計推論，而離群平均是利用離群值做統計推論，二者在觀念上有很大的不同。我們從兩個觀點來思考無母數離群平均值的研究，首先推導離群平均之漸進分配，建立α水準檢定與計算p值，接著針對離群值的判定原則，推論檢定力和漸進變異數之間的關係。

Nonparametric Outlier Mean for Gene Expression Analysis

Student: Ya-Fang You Advisor: Dr. Lin-An Chen

Institute of Statistics
National Chiao Tung University

ABSTRACT

The outlier mean has a reasonable power when the distribution is in a location shift, however, its power is remarkably reduced when he distribution is shifted on only a small fraction of observations, due to large asymptotic variances, while this happen frequently in the cancer study. We consider the study of the nonparametric outlier mean (outlier sum) in two aspects. First, the development of asymptotic distribution for establishing a level α test or computing p value is established. Second, concept of using outliers for statistical inferences may be treated differently from the classical statistical inferences that construct rules based on good data. We study the relation between powers and asymptotic variances of outliers means aiming at drawing principles for choosing outliers - based inference techniques.

致　　謝

兩年的碩士生活，十八年的學生生涯，即將在此劃上句點。

由衷地感謝我的指導教授　陳鄰安老師，有了老師細心、耐心的指導，不厭其煩地為我解決疑惑，這篇論文才能順利完成。謝謝口試委員黃冠華老師、蔡明田老師及吳柏林老師，老師們對此論文的指正與建議，使整體論文更加充實。

謝謝身邊的同學、朋友們，和你們一起成長的感覺真的很棒，情緒低落時有人分享，遇到問題時一起討論，因為你們，我不是孤軍奮戰，有你們在真好。

最後謝謝一直陪伴著我的家人們，有你們的支持，讓我求學的一路上沒有後顧之憂，讓我知道有一個溫暖港口隨時歡迎我停靠休憩，謝謝你們，我最愛的家人。

在此，將本論文獻給我的家人、朋友和師長們，致上我最誠摯的謝意，能和你們分享成果與喜悅是我最快樂的事。

雅芳　於交通大學統計學研究所
中華民國九十八年六月

Contents

Nonparametric Outlier Mean for Gene Expression Analysis

Ya-Fang You

Abstract

The outlier mean has a reasonable power when the distribution is in a location shift, however, its power is remarkably reduced when he distribution is shifted on only a small fraction of observations, due to large asymptotic variances, while this happen frequently in the cancer study. We consider the study of the nonparametric outlier mean (outlier sum) in two aspects. First, the development of asymptotic distribution for establishing a level α test or computing p value is established. Second, concept of using outliers for statistical inferences may be treated differently from the classical statistical inferences that construct rules based on good data. We study the relation between powers and asymptotic variances of outliers means aiming at drawing principles for choosing outliers - based inference techniques.

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)). For example, Sorlie et al., used gene expression to classify malignant breast tumors into five molecular subtypes (one basal-like, one ERBB2-overexpressing, two luminal-like, and one normal breast tissue-like subgroups) (Sorlie et al. (2003)). Alizadeh et al. reported that patients with 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 (Alizadeh et al. (2000)). Recently, microarray analysis has been advanced to disease classification by identifying outlier genes that are over-expressed only in a small number of disease samples (see, for example, Tibshirani and Hastie (2007); Tomlins et al. (2005)). To achieve this goal, 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) 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. Tibshirani and Hastie (2007) suggested use of an outlier sum that sums all the gene expression values in the disease group that are greater than the total of the 75th percentile and the interquartile range of the same gene. They also showed that the statistical test based on this outlier sum is noticeably more powerful than cancer outlier profile analysis in simulation. An alternative outlier sum-like statistic, called outlier robust t-statistic has been proposed by Wu (2007). Recently Chen, Chen and Chan (2008) has 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.

Although the large sample distribution of an outlier mean has been provided in Chen, Chen and Chan (2008), the nonparametric study of outlier mean is still very restricted so that its application in gene expression analysis is still limited. For specific, an outlier mean can be used to test a relation between distributions of normal group subjects and disease group subjects while this relation may be identity of these two distributions or minor relation such as only identity of two population outlier means. This is vital since different assumptions allows us to use it introducing different tests but tests for different hypotheses involves different scale estimates that may produce significant difference in their power performances. It is desired to have an advanced study of nonparametric outlier mean so that a principle for practitioner in choosing an appropriate, in terms of power performance, outlier mean test statistic is available. This is the aim that we want to achieve in this paper.

We define an outlier mean with cutoff point representing a specific form from a general class and develop its asymptotic representation and distribution. We also develop an asymptotic distribution for this outlier mean considering when the distributions of normal group subjects and disease group subjects are identical. This allows us to consider testing for hypothesis of equal distributions and hypothesis of equal population outlier means. Evaluation of power performances of these two tests are conducted and we have several interesting results. 1. If there is distributional shift in location only, then a test for hypothesis of population outlier mean is relatively more powerful than the other one. On the other hand, if there is shift in both location and scale, the two tests are very competitive. This provides important message for user when pattern of distributional shift may be observed from data. 2. The popularly used cutoff point with percentage $\alpha = 0.25$ is quite unsatisfactory in nonparametric power study for gene expression analysis while percentages $\alpha = 0.35$ or 0.45 for constructing cutoff point are satisfactory ones.

In Section 2, we first introduce an outlier mean with cutoff point representing a specific form from a general class and develop the asymptotic representation and distribution. We then develop the asymptotic distribution in Section 3 for this outlier mean restricting on the assumption that the distribution of disease group subjects and the distribution of the normal group subjects are identical. This allows us to introduce several hypotheses defined on parameters involving in the asymptotic distribution and a test for each hypothesis may be determined through estimation of parameters used in this hypothesis. In Section 4, we perform a asymptotic variance comparison for this outlier mean with several distributions for normal group variable and disease group variable. This provides a guide for user to determine a hypothesis to test when the underlying distributions in this two group belongs to this specific type. In Section 5, we will make a power comparison for these tests. Finally, the proofs of theorems are displayed in Section 6.

2 Two Tests Based on Asymptotic Distribution of the Outlier Mean

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. Extending from Tibshirani and Hastie (2007), Wu (2007) and Chen, Chen and Chan (2008), a general type cutoff point used in gene expression analysis to detect outliers may be formulated as $\sum_{j=1}^{k} c_j F_X^{-1}(\alpha_j), 0 < \alpha_j < 1, j = 1, \ldots, k$. We now define population type outlier means.

Definition 2.1. If $\sum_{j=1}^{k} c_j F_X^{-1}(\alpha_j) > F_X^{-1}(0.5)$, we call

$$\lambda_{X,Y}^{p}(\alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{P\{Y \geq \sum_{j=1}^{k} c_j F_X^{-1}(\alpha_j)\}} E[YI(Y \geq \sum_{j=1}^{k} c_j F_X^{-1}(\alpha_j))]$$

a population outlier mean with positive outliers. On the other hand, if $\sum_{j=1}^{k} c_j F_X^{-1}(\gamma_j) < F_X^{-1}(0.5)$, we call

$$\lambda_{X,Y}^{n}(\gamma_1, \gamma_2, \ldots, \gamma_k) = \frac{1}{P\{Y \leq \sum_{j=1}^{k} c_j F_X^{-1}(\gamma_j)\}} E[YI(Y \leq \sum_{j=1}^{k} c_j F_X^{-1}(\gamma_j))]$$

a population outlier mean with negative outliers.

In the literature, the outlier sum of Wu (2007) and outlier mean of Chen, Chen and Chan (2008) are of this type that we list their corresponding coefficients in Table 1.

Table 1. Coefficients for some outlier means

Outlier Mean	$\{\alpha_1, \alpha_2, \alpha_3\}$	$\{c_1, c_2, c_3\}$
Wu (2007)	$\{0.25, 0.75, 0.75\}$	$\{-1, 1, 1\}$
Chen, Chen and Chan (2008)	$\{0.25, 0.5, 0.75\}$	$\{-\kappa, 1, \kappa\}$

where $\kappa > 0$

Invariance property is desired for any statistical function and then not every population outlier mean introduced above is interesting with this concern. Suppose that a random variable X has a quantile function $F_X^{-1}(\alpha)$. It is known that its quantile $F_X^{-1}(\alpha)$ has the following properties

$$F_{aX+b}^{-1}(\alpha) = \begin{cases} aF_X^{-1}(\alpha) + b & \text{if } a > 0 \\ aF_X^{-1}(1-\alpha) + b & \text{if } a < 0 \end{cases}$$

We may see the condition that a population outlier mean satisfies desired invariance properties.

Theorem 2.2. *Suppose that $c_j, j = 1, ..., k$ satisfy $\sum_{j=1}^{k} c_j = 1$. Then, the population outlier mean with positive outliers has the following properties*

$$\lambda^p_{aX+b,aY+b}(\alpha_1, \alpha_2, \ldots, \alpha_k) = \begin{cases} a\lambda^p_{X,Y}(\alpha_1, \alpha_2, \ldots, \alpha_k) + b & \textit{if } a > 0 \\ a\lambda^n_{X,Y}(1-\alpha_1, 1-\alpha_2, \ldots, 1-\alpha_k) + b & \textit{if } a < 0 \end{cases}$$

On the other hand, the population outlier mean with negative outliers has the following properties

$$\lambda^n_{aX+b,aY+b}(\gamma_1, \gamma_2, \ldots, \gamma_k) = \begin{cases} a\lambda^n_{X,Y}(\gamma_1, \gamma_2, \ldots, \gamma_k) + b & \textit{if } a > 0 \\ a\lambda^p_{X,Y}(1-\gamma_1, 1-\gamma_2, \ldots, 1-\gamma_k) + b & \textit{if } a < 0 \end{cases}$$

If outlier means $\lambda^p_{X,Y}(\alpha_1, \alpha_2, \ldots, \alpha_k)$ and $\lambda^n_{X,Y}(\gamma_1, \gamma_2, \ldots, \gamma_k)$ are formulated with $\sum_{j=1}^{k} c_j \neq 1$, we may see from the proof (see Section 6) of Theorem 2.2 that they are no longer to be equivalent like the quantile function.

We suggest the population cutoff point of the form $2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)$. Let $\hat{F}_X^{-1}$ be the empirical quantile function for estimating population quantile function F_X^{-1}. The sample outlier mean can be expressed as

$$\hat{\lambda} = \frac{\sum_{i=1}^{n_2} Y_i I(Y_i \geq 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha))}{\sum_{i=1}^{n_2} I(Y_i \geq 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha))}. \tag{2.1}$$

Implicitly this sample outlier means tries to estimate the following population outlier mean

$$\mu_\lambda = \frac{E[YI(Y \geq 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha))]}{P\{Y \geq 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)\}}.$$

For establishing large sample theory based p value, we consider the following location models,

$$\begin{aligned} X_i &= \mu_X + \epsilon_i, \quad i = 1, \ldots, n_1, \\ Y_i &= \mu_Y + \delta_i, \quad i = 1, \ldots, n_2, \end{aligned} \tag{2.2}$$

where ϵ_i's and δ_i's are finite sequences of independent and identically distributed random variables having distribution functions F_ϵ and F_δ and probability density functions f_ϵ and f_δ respectively. In addition, $E(\epsilon_i) = E(\delta_i) = 0$ and $Var(\epsilon_i) = \sigma_X^2$ and $Var(\delta_i) = \sigma_Y^2$. With this setup, $F_X(x) = F_\epsilon(x - \mu_X)$ and $F_Y(y) = F_\delta(y - \mu_Y)$. In terms of error distributions in (2.2), the population outlier mean is

$$\mu_\lambda = \mu_Y + \frac{\int_\eta^\infty \delta f_\delta(\delta) d\delta}{\beta}$$

where $\beta = P\{\delta \geq \eta\}$ with $\eta = 2F_\epsilon^{-1}(1-\alpha) - F_\epsilon^{-1}(\alpha) + \mu_X - \mu_Y$.

Theorem 2.3. *Suppose that assumptions* (A_2), (A_3) *and* (A_4) *in the Appendix are true.*

(a) A Bahadur representation of the outlier mean is

$$\begin{aligned}\sqrt{n_2}(\hat{\lambda}-\mu_\lambda) =& ((1-\alpha)b_1-\alpha b_2)n_1^{-1/2}\sum_{i=1}^{n_1} I(\epsilon_i \le F_\epsilon^{-1}(\alpha)) \\ & -\alpha(b_1+b_2)n_1^{-1/2}\sum_{i=1}^{n_1} I(F_\epsilon^{-1}(\alpha) \le \epsilon_i \le F_\epsilon^{-1}(1-\alpha)) \\ & +(-\alpha b_1+(1-\alpha)b_2)n_1^{-1/2}\sum_{i=1}^{n_1} I(\epsilon_i \ge F_\epsilon^{-1}(1-\alpha)) \\ & +\frac{1}{\beta}n_2^{-1/2}\sum_{i=1}^{n_2}\{\delta_i I(\delta_i \ge \eta)-\int_\eta^\infty \delta f_\delta(\delta)d\delta\}+o_p(1)\end{aligned}$$

where

$$\begin{aligned}b_1 &= \frac{-1}{\beta}\eta f_\delta(\eta)\sqrt{h}f_\epsilon^{-1}(F_\epsilon^{-1}(\alpha)), \\ b_2 &= \frac{-2}{\beta}\eta f_\delta(\eta)\sqrt{h}f_\epsilon^{-1}(F_\epsilon^{-1}(1-\alpha)).\end{aligned}$$

(b) $\sqrt{n_2}(\hat{\lambda}-\mu_\lambda)$ *converges in distribution to* $N(0,\sigma_\lambda^2)$ *where*

$$\begin{aligned}\sigma_\lambda^2 =& \sigma^2(b_1,b_2,v) \\ =& \alpha(1-\alpha)((1-\alpha)b_1-\alpha b_2)^2+2(1-2\alpha)\alpha^3(b_1+b_2)^2 \\ & +\alpha(1-\alpha)(\alpha b_1-(1-\alpha)b_2)^2+v\end{aligned}$$

where

$$v=\frac{1}{\beta^2}[\int_\eta^\infty \delta^2 f_\delta(\delta)d\delta-(\int_\eta^\infty \delta f_\delta(\delta)d\delta)^2].$$

3 Outlier Mean Based Hypothesis Testings

The basic idea behind the use of the outlier mean or outlier sum in gene expression analysis is to see if the disease group subjects and the normal group subjects are similar in some sense. Asymptotic normality for the outlier mean allows us to develop tests for hypotheses dealing with all combinations of asymptotic mean μ_λ and asymptotic standard deviation σ_λ. However, it is not ready in introducing these tests without knowing the asymptotic properties of this outlier mean when the distributions for two groups of subjects are assumed to be identical as

$$H_0 : F_Y = F_X. \tag{3.1}$$

Under H_0, model (2.2) may be reformulated as the following model,

$$X_i = \mu_x + \epsilon_i, i = 1, \ldots, n_1 + n_2 \tag{3.2}$$

where $X_i, i = 1, \ldots, n_1$ belongs to normal group and $X_i, i = n_1+1, \ldots, n_1+n_2$ belongs to disease group and ϵ_i's are independent and identically distributed random variables having distribution as defined. Hence, when H_0 is true, the sample outlier mean of (2.1) may be reformulated as

$$\hat{\lambda} = \frac{\sum_{i=n_1+1}^{n_1+n_2} X_i I(X_i \geq 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha))}{\sum_{i=n_1+1}^{n_1+n_2} I(X_i \geq 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha))} \tag{3.3}$$

where quantile estimates $\hat{F}_X^{-1}(\alpha)$ and $\hat{F}_X^{-1}(1-\alpha)$ are constructed based on samples $X_1, \ldots, X_{n_1}$. The outlier mean of (3.3) tries to estimate the following parameter

$$\mu_{\lambda X} = \frac{E[XI(X \geq 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha))]}{P\{X \geq 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)\}}$$

which, in terms of error distribution, is

$$\mu_{\lambda X} = \mu_X + \frac{\int_{\eta_X}^{\infty} \epsilon f_\epsilon(\epsilon) d\epsilon}{\beta_X}$$

where $\beta_X = P\{\epsilon \geq \eta_X\}$ with $\eta_X = 2F_\epsilon^{-1}(1-\alpha) - F_\epsilon^{-1}(\alpha)$.

The following theorem states the asymptotic property for the outlier mean when the observations are drawn from model (3.2).

Theorem 3.1. *When H_0 is true,$\sqrt{n_2}(\hat{\lambda} - \mu_{\lambda X})$ converges in distribution to a normal random variable having distribution $N(0, \sigma_{\lambda X}^2)$ with*

$$\begin{aligned}\sigma_{\lambda X}^2 =& \sigma^2(b_{1X}, b_{2X}, v_X)\\ =& \alpha(1-\alpha)((1-\alpha)b_{1X} - \alpha b_{2X})^2 + 2(1-2\alpha)\alpha^3(b_{1X} + b_{2X})^2\\ &+ \alpha(1-\alpha)(\alpha b_{1X} - (1-\alpha)b_{2X})^2 + v_X\end{aligned}$$

where we denote

$$\begin{aligned} b_{1X} &= \frac{-1}{\beta_X}(\eta_X) f_\epsilon(\eta_X)\sqrt{h} f_\epsilon^{-1}(F_\epsilon^{-1}(\alpha)) \\ b_{2X} &= \frac{-2}{\beta_X}(\eta_X) f_\epsilon(\eta_X)\sqrt{h} f_\epsilon^{-1}(F_\epsilon^{-1}(1-\alpha)), \\ v_X &= \frac{1}{(\beta_X)^2}[\int_{\eta_X}^{\infty} \epsilon^2 f_\epsilon(\epsilon) d\epsilon - (\int_{\eta_X}^{\infty} \epsilon f_\epsilon(\epsilon) d\epsilon)^2]. \end{aligned}$$

Theorem 3.1 indicates that when H_0 is true, $\sqrt{n_2}(\frac{\hat{\lambda}-\mu_{\lambda X}}{\sigma_{\lambda X}})$ converges to the standard normal distribution and the distribution parameters when H_0 is true involved in the function are $\mu_{\lambda X}$ and $\sigma_{\lambda X}$. Then, joining Theorems 2.3 and 3.1, we have three choices of constructing test functions as follows:

$$\sqrt{n_2}(\frac{\hat{\lambda}-\mu_{\lambda X}}{\sigma_{\lambda X}}), \sqrt{n_2}(\frac{\hat{\lambda}-\mu_{\lambda X}}{\sigma_{\lambda}}), \text{ and } \sqrt{n_2}(\frac{\hat{\lambda}-\mu_{\lambda}}{\sigma_{\lambda X}}). \tag{3.4}$$

the first function considering testing hypothesis involving both asymptotic mean and standard deviation and the others consider only one of these two parameters. Then when we have appropriate estimates of the unknown parameters, test statistics are provided.

Not all test functions are interesting in gene expression analysis since Tomlins et al. (2005) has observed that when outliers occurs in disease samples, they are either only over-expressed or down-expressed. Hence, without considering a location shift the resulted test function is not practical in gene expression analysis. The following procedures are designed for the first two test functions:

(I) Hypothesis for equality of distributions: $H_{\mu,\sigma} : \mu_\lambda = \mu_{\lambda X}, \sigma_\lambda^2 = \sigma_{\lambda X}^2$

(a) The rule for testing $H_{\mu,\sigma}$ is:

$$\text{rejecting } H_{\mu,\sigma} \text{ if } \sqrt{n_2}(\frac{\hat{\lambda}-\hat{\mu}_{\lambda X}}{\hat{\sigma}_{\lambda X}}) \geq z_{\alpha^*} \tag{3.5}$$

where $\hat{\mu}_{\lambda X}$ and $\hat{\sigma}_{\lambda X}$ are, respectively, estimators for parameters $\mu_{\lambda X}$ and $\sigma_{\lambda X}$.

(b) An approximate p value based on observations x_i's and y_i's is defined as

$$p = \int_{\sqrt{n_2}(\frac{\hat{\lambda}-\hat{\mu}_{\lambda X}}{\hat{\sigma}_{\lambda X}})}^{\infty} \phi(z) dz.$$

(II) Hypothesis for outlier variable's expectation: $H_\mu : \mu_\lambda = \mu_{\lambda X}$

(a) The rule for testing H_μ is:

$$\text{rejecting } H_\mu \text{ if } \sqrt{n_2}(\frac{\hat{\lambda}-\hat{\mu}_{\lambda X}}{\hat{\sigma}_\lambda}) \geq z_{\alpha^*} \tag{3.6}$$

where $\hat{\sigma}_\lambda$ is estimator of parameter σ_λ when $Y \sim F_Y$ has distribution F_Y.

(b) An approximate p value based on observations x_i's and y_i's is defined as

$$p = \int_{\sqrt{n_2}(\frac{\hat{\lambda}-\hat{\mu}_{\lambda X}}{\hat{\sigma}_\lambda})}^{\infty} \phi(z)dz.$$

The determination of test selection now relies on (i) power performance and (ii) choice of parameters estimates that will be studied in subsequent sections.

4 Comparison of Outlier Coverages and Asymptotic Variances of Outlier Mean

Tests (3.5) and (3.6) use the same critical point z_{α^*} and the same estimate $\hat{\mu}_{\lambda X}$ for the outlier variable's expectation. When we consider to choose one from hypotheses $H_{\mu,\sigma}$ or H_μ, the right choice is the one that has smaller asymptotic variance ($\sigma_{\lambda X}$ or σ_λ). We will see that the size of this asymptotic variance has a relation with the outlier coverage β. We compute the outlier coverage probabilities β_X and β and asymptotic variances $\sigma_{\lambda X}$ and σ_λ with the following distribution setting:

$$F_X = N(0,1) \text{ and } F_Y = N(\theta,1). \tag{4.1}$$

Table 2. Coverage probabilities and asymptotic variances when there is distributional shift

θ	α	β_X	β	$\sigma^2_{\lambda X}$	σ^2_λ
1	0.45	0.3531	0.7333	3.1485	1.6203
	0.35	0.1238	0.4380	32.379	5.4481
	0.25	0.0215	0.1530	369.41	48.11
	0.15	0.0009	0.0174	13068.76	877.44
	0.05	4.0e-7	4.2e-5	6.5e+7	6.5e+5
3	0.45	0.3531	0.9956	3.1485	1.0123
	0.35	0.1238	0.9674	32.379	1.2644
	0.25	0.0215	0.8356	369.41	3.2525
	0.15	0.0009	0.4565	13068.76	18.63
	0.05	4.0e-7	0.0265	6.5e+7	1416.67
10	0.45	0.3531	1	3.1485	1
	0.35	0.1238	1	32.379	1
	0.25	0.0215	1	369.41	1
	0.15	0.0009	1	13068.76	1
	0.05	4.0e-7	1	6.5e+7	1

We have several comments drawn from Table 2:

1. It is seen that $\beta_X < \beta$ for all cases of θ and α. This indicates that the outlier interval $[2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha), \infty)$ covers space of Y more probable than space of X. This size of the difference could be huge. For example, when $\theta = 10$, β_X's are all very small but β is or nearly 1 indicating that outlier interval contains almost whole probable space of variable Y.

2. The differences in coverage probabilities strongly affect the asymptotic variances in the way that $\sigma^2_{\lambda X} > \sigma^2_\lambda$ for all cases of θ and α where the asymptotic variance under hypothesis $H_{\mu,\sigma}$ could be hundred or thousand times it under hypothesis H_μ.

3. When $\theta = 10$, the asymptotic variances under hypothesis concerning population outlier mean are vales nearly 1's. This indicating that the asymptotic variance under this hypothesis is the variance of the random variable Y.

We may be more interesting in the comparison for the following contaminated alternative one:

$$F_X = N(0,1) \text{ and } F_Y = (1-\gamma)N(0,1) + \gamma N(\theta,1) \tag{4.2}$$

where $\theta > 0$. This alternative hypothesis assumes that Y has a location model with positive mean $\gamma\theta$ and contaminated error variable. Table 3 provides β, β_X, $\sigma^2_{\lambda X}$ and σ^2_λ for this underlying distribution.

Table 3. Coverage probabilities and asymptotic variances when there small proportion ($\gamma = 0.1$) of distributional shift

θ	α	β_X	β	$\sigma^2_{\lambda X}$	σ^2_{λ}
1	0.45	0.3531	0.3911	3.1485	3.0160
	0.35	0.1238	0.1552	32.379	25.0101
	0.25	0.0215	0.0346	369.41	239.5184
	0.15	0.0009	0.0025	13068.76	5095.878
	0.05	4.0e-7	4.5e-6	6.5e+7	5.9e+6
3	0.45	0.3531	0.4173	3.1485	5.9860
	0.35	0.1238	0.2082	32.379	27.118
	0.25	0.0215	0.1029	369.41	97.7885
	0.15	0.0009	0.0465	13068.76	359.8992
	0.05	4.0e-7	0.0003	6.5e+7	12821.53
10	0.45	0.3531	0.4178	3.1485	50.6809
	0.35	0.1238	0.2115	32.379	201.8219
	0.25	0.0215	0.1194	369.41	640.0708
	0.15	0.0009	0.1008	13068.76	895.2527
	0.05	4.0e-7	0.1000	6.5e+7	909.9943

We have several comments for interpreting the results in Table 3:

1. Setting F_Y as a contaminated normal distribution of (4.2) indicating that response variable for disease gene has large proportion of observations from the distribution F_X but with a small part of observations shifted to the right. The variance of the contaminated distribution is $1+\gamma(1-\gamma)\theta^2$. Both the contamination and variance enlargement affect the coverage probability β, smaller than those in Table 2. This results in the outlier mean asymptotic variance σ^2_λ, larger than those in Table 2.

2. For mild shifts ($\theta = 1$ or 3), the test for hypothesis H_μ has asymptotic variances σ^2_λ's almost smaller (except $(\theta, \alpha) = (3, 0.45)$) than those for hypothesis $H_{\mu,\sigma}$. When there is significant shift $\theta = 10$, $\sigma^2_{\lambda X}$'s are smaller than σ^2_λ's for $\alpha \in \{0.25, 0.35, 0.45\}$.

We now consider the case that random variable Y has a mixed distribution with shift not only the mean but also the variance as follows:

$$F_X = N(0,1) \text{ and } F_Y = 0.9N(0,1) + 0.1N(\theta, \sigma^2)$$

where $\theta > 0$. For cutting percentage α and true values θ and σ, we compute the asymptotic variance and display the comparison in table 4.

Table 4. Asymptotic variances comparison when there small proportion ($\gamma = 0.1$) of distributional shift

σ	θ	$\sigma^2_\lambda < \sigma^2_{\lambda X}$	$\sigma^2_\lambda > \sigma^2_{\lambda X}$
1	1	0.45, 0.35, 0.25, 0.15, 0.05	none
	3	0.35, 0.25, 0.15, 0.05	0.45
	10	0.15, 0.05	0.45, 0.35, 0.25
3	1	0.25, 0.15, 0.05	0.45, 0.35
	3	0.25, 0.15, 0.05	0.45, 0.35
	10	0.15, 0.05	0.45, 0.35, 0.25
5	1	0.15, 0.05	0.45, 0.35, 0.25
	3	0.15, 0.05	0.45, 0.35, 0.25
	10	0.15, 0.05	0.45, 0.35, 0.25
10	1	0.15, 0.05	0.45, 0.35, 0.25
	3	0.15, 0.05	0.45, 0.35, 0.25
	10	0.15, 0.05	0.45, 0.35, 0.25

In this case that both contaminated mean and variance are shifted, it shows $\sigma^2_\lambda < \sigma^2_{\lambda X}$ for most of smaller $\alpha \in \{0.05, 0.15\}$ and $\sigma^2_\lambda > \sigma^2_{\lambda X}$ for most larger $\alpha \in \{0.25, 0.35, 0.45\}$. This provides a guide to choose hypothesis for testing when percentage α is already decided.

5 Power Studies with Tests Based on Outlier Mean

Consider the power function for testing equal distributions hypothesis $H_{\mu,\sigma}$. By letting $\mu_{\lambda Y}$ and $\sigma_{\lambda Y}$, respectively, as parameters of μ_λ and σ_λ when $Y \sim F_Y$ is true, an approximate power with significant level α^* based on test (3.5) may be derived as bellows

$$\begin{aligned}\ell_{H_{\mu,\sigma}} &= P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\lambda}-\hat{\mu}_{\lambda X}}{\hat{\sigma}_{\lambda X}}) \geq z_{\alpha^*}\} \\ &= P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\lambda}-\mu_{\lambda Y}}{\sigma_{\lambda Y}}) \geq \frac{z_{\alpha^*}\hat{\sigma}_{\lambda X}+\sqrt{n_2}(\hat{\mu}_{\lambda X}-\mu_{\lambda Y})}{\sigma_{\lambda Y}}\} \\ &\approx P\{Z \geq \frac{z_{\alpha^*}\hat{\sigma}_{\lambda X}+\sqrt{n_2}(\hat{\mu}_{\lambda X}-\mu_{\lambda Y})}{\sigma_{\lambda Y}}\}. \end{aligned} \tag{5.1}$$

This is the power function when we test for hypothesis of equal distributions.

On the other hand, the power function for testing equal outlier means hypothesis H_μ with significant level α^* may be derived as bellows

$$\begin{aligned}\ell_{H_{\mu}} &= P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\lambda}-\hat{\mu}_{\lambda X}}{\hat{\sigma}_{\lambda Y}}) \geq z_{\alpha^*}\} \\ &= P_{F_Y}\{\sqrt{n_2}(\frac{\hat{\lambda}-\mu_{\lambda Y}}{\sigma_{\lambda Y}}) \geq \frac{z_{\alpha^*}\hat{\sigma}_{\lambda Y}+\sqrt{n_2}(\hat{\mu}_{\lambda X}-\mu_{\lambda Y})}{\sigma_{\lambda Y}}\} \\ &\approx P\{Z \geq \frac{z_{\alpha^*}\hat{\sigma}_{\lambda Y}+\sqrt{n_2}(\hat{\mu}_{\lambda X}-\mu_{\lambda Y})}{\sigma_{\lambda Y}}\}. \end{aligned} \tag{5.2}$$

From (5.1) and (5.2), the performance of these two tests rely on several elements describing in the following:

n_2 : the larger the sample size for the disease gene, the larger the powers. Due to the fact that $\hat{\mu}_{\lambda X} < \mu_{\lambda Y}$ when there are outliers in Y.

$\sigma^2_{\lambda X}$: the larger the asymptotic variance, the smaller the power for testing hypothesis $H_{\mu,\sigma}$

$\sigma^2_{\lambda Y}$: the larger the asymptotic variance, the smaller the power for testing hypothesis H_μ

We also note that when cutoff point percentage α decreases, the outlier mean asymptotic variances $\sigma^2_{\lambda X}$ and $\sigma^2_{\lambda Y}$ are both increase.

We now consider the design of distributional shift of (4.1) and compute the approximate powers for testing hypotheses $H_{\mu,\sigma}$ and H_μ. The results are displayed in Tables 5.

Table 5. Approximate powers $(\ell_{H_{\mu,\sigma}}, \ell_{H_\mu})$ of outlier mean when there is distributional shift

n_2	α	$\theta = 1$	$\theta = 3$	$\theta = 5$	$\theta = 10$
30	0.45	(0.2775, 0.5230)	(1, 1)	(1, 1)	(1, 1)
	0.35	(3.0e-4, 0.1441)	(0.0826, 1)	(1, 1)	(1, 1)
	0.25	(4.5e-6, 0.0634)	(0, 0.8634)	(0, 1)	(1, 1)
	0.15	(1.2e-10, 0.0517)	(0, 0.1513)	(0, 1)	(0, 1)
	0.05	(0, 0.0500)	(0, 0.0530)	(0, 0.1544)	(0, 1)
50	0.45	(0.4622, 0.7099)	(1, 1)	(1, 1)	(1, 1)
	0.35	(5.6e-4, 0.1861)	(0.7358, 1)	(1, 1)	(1, 1)
	0.25	(5.3e-6, 0.0679)	(0, 0.9708)	(0, 1)	(1, 1)
	0.15	(1.3e-10, 0.0521)	(0, 0.1971)	(0, 1)	(0, 1)
	0.05	(0, 0.0500)	(0, 0.0538)	(0, 0.2018)	(0, 1)

We have comments drawn from results showing in Tables 5:

1. Testing hypotheses $H_{\mu,\sigma}$ and H_μ have small powers for mild shifts $\theta = 1, 3$ unless we choose large proportions α ($\alpha = 0.35$ and 0.45). If there is significant shifting in location ($\theta = 10$), most of these two tests are satisfactory. The percentage $\alpha = 0.25$ is the recommended popularly in literature (see Hoaglin et al. (1983)).

2. A comparison of approximate powers between these two tests shows that the test for hypothesis H_μ seems to be the right choice. To test hypothesis $H_{\mu,\sigma}$ gives unsatisfactory powers besides cases of strong distributional shift such as $\theta = 5$ or 10 with choosing percentage α as large as 0.35 or 0.45.

3. The effects of these two tests exist in sample size. Basically the larger the sample size generates larger power for either one test.

Next, we consider that the assumption for distributions of X and Y is that Y has a case of contaminated normal in (4.2) as

$$F_X = N(0,1) \text{ and } F_Y = 0.9N(0,1) + 0.1N(\theta,1)$$

where $\theta > 0$. The computed approximate powers for testing hypotheses $H_{\mu,\sigma}$ and H_μ are displayed in Table 6.

Table 6. Approximate powers $(\ell_{H_{\mu,\sigma}}, \ell_{H_\mu})$ of outlier mean when there is small fraction distributional shift

n_2	α	$\theta = 1$	$\theta = 3$	$\theta = 5$	$\theta = 10$
30	0.45	(0.0740, 0.0791)	(0.4420, 0.2750)	(0.7307, 0.4072)	(0.8921, 0.5011)
	0.35	(0.0363, 0.0584)	(0.1353, 0.1713)	(0.4635, 0.3136)	(0.8060, 0.4512)
	0.25	(0.0217, 0.0525)	(0.0026, 0.1077)	(0.0669, 0.2326)	(0.5517, 0.3951)
	0.15	(0.0043, 0.0505)	(0.0000, 0.0658)	(0.0000, 0.1444)	(1.9e-7, 0.3285)
	0.05	(2.1e-8, 0.0500)	(0.0000, 0.0510)	(0.0000, 0.0638)	(0.0000, 0.2238)
50	0.45	(0.0840, 0.0897)	(0.5631, 0.3847)	(0.8474, 0.5698)	(0.9570, 0.6852)
	0.35	(0.0382, 0.0611)	(0.1843, 0.2277)	(0.5970, 0.4411)	(0.9043, 0.6256)
	0.25	(0.0221, 0.0532)	(0.0038, 0.1311)	(0.1087, 0.3213)	(0.7023, 0.5537)
	0.15	(0.0043, 0.0506)	(0.0000, 0.0711)	(0.0000, 0.1866)	(1.1e-6, 0.4623)
	0.05	(2.1e-8, 0.0500)	(0.0000, 0.0512)	(0.0000, 0.0684)	(0.0000, 0.3079)

We have comments drawn from results showing in Tables 6:

1. Basically the contaminated distribution F_Y reduces the powers of two tests due to enlarging the asymptotic outlier mean asymptotic variances $\sigma^2_{\lambda X}$ and σ^2_λ due to contamination and increasing the variance of distribution F_Y.

2. If we specify cutoff point percentage α to be 0.35 or more, the test for hypothesis $H_{\mu,\sigma}$ seems to be the right choice. On the other hand, if we specify cutoff point percentage α to be smaller than 0.25, the test for hypothesis H_μ seems to be the right choice. For $\alpha = 0.25$, the test for hypothesis H_μ is better unless the location parameter θ in contaminated distribution is as large 10.

Table 7. Approximate powers $(\ell_{H_{\mu,\sigma}}, \ell_{H_\mu})$ of outlier mean when there is small fraction distributional shift ($n_2 = 50$)

γ	α	$\theta = 5$	$\theta = 10$	$\theta = 20$
0.05	0.45	(0.5902, 0.3196)	(0.8241, 0.4207)	(0.9008, 0.4606)
	0.35	(0.3782, 0.2445)	(0.7326, 0.3746)	(0.8706, 0.4383)
	0.25	(0.1323, 0.1961)	(0.5828, 0.3339)	(0.8200, 0.4130)
	0.15	(2.4e-15, 0.1301)	(0.0005, 0.2816)	(0.1986, 0.3824)
	0.05	(0.0000, 0.0632)	(0.0000, 0.1955)	(0.0000, 0.3301)
0.20	0.45	(0.9818, 0.8759)	(0.9977, 0.9385)	(0.9992, 0.9550)
	0.35	(0.8295, 0.7529)	(0.9881, 0.9064)	(0.9981, 0.9455)
	0.25	(0.0614, 0.5591)	(0.8336, 0.8492)	(0.9879, 0.9289)
	0.15	(0.0000, 0.3026)	(8.6e-13, 0.7516)	(0.0376, 0.9011)
	0.05	(0.0000, 0.0758)	(0.0000, 0.5274)	(0.0000, 0.8367)
0.30	0.45	(0.9985, 0.9785)	(1.0000, 0.9939)	(1.0000, 0.9965)
	0.35	(0.9338, 0.9227)	(0.9990, 0.9871)	(0.9999, 0.9952)
	0.25	(0.0302, 0.7601)	(0.9101, 0.9687)	(0.9986, 0.9924)
	0.15	(0.0000, 0.4305)	(0.0000, 0.9187)	(0.0102, 0.9859)
	0.05	(0.0000, 0.0825)	(0.0000, 0.7246)	(0.0000, 0.9635)
0.50	0.45	(1.0000, 0.9968)	(1.0000, 1.0000)	(1.0000, 1.0000)
	0.35	(0.9952, 0.9111)	(1.0000, 1.0000)	(1.0000, 1.0000)
	0.25	(0.0038, 0.4660)	(0.9836, 0.9999)	(1.0000, 1.0000)
	0.15	(0.0000, 0.1104)	(0.0000, 0.9986)	(0.0002, 1.0000)
	0.05	(0.0000, 0.0525)	(0.0000, 0.9616)	(0.0000, 0.9998)

We have several comments on the results in Table 7:

1. Although the more the contamination (γ) makes the variance of the response variable Y, however, it is easier in detection of existence of outliers so that the powers of two tests increase. The powers for $\gamma = 0.5$ are very close to the performance of location shift in Table 5.

2. Even the large contamination ($\gamma = 0.5$), the test for hypothesis $H_{\mu,\sigma}$ with low α's ($\alpha = 0.05$ and 0.05) is still very poor in power performance.

3. Combining the discussions for the results in Tables 5-7, the test for hypothesis H_μ is relatively more robust.

Besides the cases of normal or mixed normal distributions, we may consider the cases that X and Y draw from the following two cases:

Case 1: $F_X = Laplace(0,1)$ and $F_Y = Laplace(\theta,1)$
Case 2: $F_X = t(5)$ and $F_Y = t(5)+\theta$.

Table 8. Approximate powers $(\ell_{H_{\mu,\sigma}}, \ell_{H_\mu})$ for hypothesis H_μ and $H_{\mu,\sigma}$ when X and Y are with Laplace or t distribution $(n_2 = 30)$

8.(a) $Y \sim Laplace(\theta, 1)$

α	$\theta = 1$	$\theta = 3$	$\theta = 5$	$\theta = 10$
0.45	(0.0193, 0.2899)	(1.0000, 1.0000)	(1, 1)	(1, 1)
0.35	(5.2e-7, 0.0500)	(0.0134, 0.9997)	(1, 1)	(1, 1)
0.25	(7.1e-7, 0.0500)	(0, 0.4908)	(0, 1)	(1, 1)
0.15	(6.7e-4, 0.0500)	(0, 0.0500)	(0, 0.8694)	(0, 1)

8.(b) $Y \sim t(5) + \theta$

α	$\theta = 1$	$\theta = 3$	$\theta = 5$	$\theta = 10$
0.45	(0.0465, 0.4128)	(1, 1)	(1, 1)	(1, 1)
0.35	(8.2e-8, 0.0777)	(0.0002, 0.9999)	(1, 1)	(1, 1)
0.25	(4.0e-6, 0.0433)	(0, 0.5184)	(0, 1)	(0.9838, 1)
0.15	(0.0001, 0.0460)	(0, 0.0167)	(0, 0.8794)	(0, 1)

From the displayed results, it seems that two tests are quite satisfactory when there are significant location shifts. However, the test for hypothesis H_μ is uniformly better than it for hypothesis $H_{\mu,\sigma}$. The test for hypothesis H_μ is very satisfactory for small percentage α when there is location shift is as large as 5 or more.

We now consider the case that random variable Y has a mixed distribution with shift not only the mean but also the variance as follows:

$$F_X = N(0,1) \text{ and } F_Y = 0.9N(0,1) + 0.1N(\theta, \sigma^2)$$

where $\theta > 0$. For sample size $n_2 = 30$, cutting percentage α and true values θ and σ, we compute the approximate powers, $\ell_{H_{\mu,\sigma}}$ and ℓ_{H_μ}. With, $\alpha = 0.05, 0.15, 0.25, 0.35, 0.45$, $\theta = 1, 3, 10$, we display a comparison of two approximate powers in the following table.

Table 9. Comparison of approximate powers

σ	θ	$\ell_{H_\mu} > \ell_{H_{\mu,\sigma}}$	$\ell_{H_\mu} < \ell_{H_{\mu,\sigma}}$
1	1	0.45, 0.35, 0.25, 0.15, 0.05	none
	3	0.35, 0.25, 0.15, 0.05	0.45
	10	0.15, 0.05	0.45, 0.35, 0.25
3	1	0.25, 0.15, 0.05	0.45, 0.35
	3	0.25, 0.15, 0.05	0.45, 0.35
	10	0.15, 0.05	0.45, 0.35, 0.25
5	1	0.15, 0.05	0.45, 0.35, 0.25
	3	0.15, 0.05	0.45, 0.35, 0.25
	10	0.15, 0.05	0.45, 0.35, 0.25
10	1	0.15, 0.05	0.45, 0.35, 0.25
	3	0.15, 0.05	0.45, 0.35, 0.25
	10	0.15, 0.05	0.45, 0.35, 0.25

In general, we test hypothesis $H_{\mu,\sigma}$ is more powerful than to test hypothesis $H_{\mu,\sigma}$ when we choose percentage α as large as 0.35 or 0.45 and test hypothesis H_μ when we choose percentage α as small as 0.15 or 0.25.

6 Appendix

To investigate Theorem 3.2, let's establish a more general theory for outlier mean Π. The following assumptions are needed.

(A_1) The limit $h = lim_{n_1,n_2\to\infty}\frac{n_2}{n_1}$ exists.

(A_2) Suppose that there is constant C such that $\sqrt{n}_1(\hat{C}-C) = O_p(1)$ where C depends generally on distribution of X.

(A_3) Probability density function f_δ of distribution F_δ is bounded away from zero in a neighborhood of quantity $C - \mu_Y$.

(A_4) Probability density function f_ϵ is bounded away from zero in the neighborhood of $F_\epsilon^{-1}(\alpha)$ for $\alpha \in (0,1)$.

Proof of Theorem 2.2

Proof. If $a > 0$,

$$\begin{aligned}
&\lambda^p_{aX+b,aY+b}(\alpha_1,\alpha_2,\ldots,\alpha_k)\\
=&\frac{E[(aY+b)I(aY+b \geq \sum_{j=1}^k c_j F^{-1}_{aX+b}(\alpha_j))]}{P\{aY+b \geq \sum_{j=1}^k c_j F^{-1}_{aX+b}(\alpha_j)\}}\\
=&\frac{E[(aY+b)I(aY+b \geq \sum_{j=1}^k c_j (aF^{-1}_X(\alpha_j)+b))]}{P\{aY+b \geq \sum_{j=1}^k c_j (aF^{-1}_X(\alpha_j)+b)\}}\\
=&\frac{E[(aY+b)I(Y \geq \sum_{j=1}^k c_j F^{-1}_X(\alpha_j))]}{P\{Y \geq \sum_{j=1}^k c_j F^{-1}_X(\alpha_j)\}}\\
=&a\frac{E[YI(Y \geq \sum_{j=1}^k c_j F^{-1}_X(\alpha_j))]}{P\{Y \geq \sum_{j=1}^k c_j F^{-1}_X(\alpha_j)\}}+b\\
=&a\lambda^p_{X,Y}(\alpha_1,\alpha_2,\ldots,\alpha_k)+b.
\end{aligned}$$

On the other hand, if $a < 0$,

$$
\begin{aligned}
&\lambda^p_{aX+b,aY+b}(\alpha_1,\alpha_2,\ldots,\alpha_k)\\
=&\frac{E[(aY+b)I(aY+b \geq \sum_{j=1}^k c_j(aF_X^{-1}(1-\alpha_j)+b))]}{P\{aY+b \geq \sum_{j=1}^k c_j(aF_X^{-1}(1-\alpha_j)+b)\}}\\
=&\frac{E[(aY+b)I(Y \leq \sum_{j=1}^k c_jF_X^{-1}(1-\alpha_j))]}{P\{Y \leq \sum_{j=1}^k c_jF_X^{-1}(1-\alpha_j)\}}\\
=&a\frac{E[YI(Y \leq \sum_{j=1}^k c_jF_X^{-1}(1-\alpha_j))]}{P\{Y \leq \sum_{j=1}^k c_jF_X^{-1}(1-\alpha_j)\}}+b\\
=&a\lambda^n_{X,Y}(1-\alpha_1,1-\alpha_2,\ldots,1-\alpha_k)+b.
\end{aligned}
$$

The proof of transformation on outlier mean with negative outliers may be similarly proved and it is skipped. □

Proof of Theorem 2.3

Proof. Let $C = 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)$ and $\hat{C} = 2\hat{F}_X^{-1}(1-\alpha) - \hat{F}_X^{-1}(\alpha)$. From model (2.2) and the expression of $\hat{\lambda}_X$ in (2.1), we have

$$\hat{\lambda} = \mu_Y + \frac{\sum_{i=1}^{n_2}\delta_i I(\delta_i > C - \mu_y + n_1^{-1/2}T)}{\sum_{i=1}^{n_2} I(Y_i > \hat{C})}$$

where $T = \sqrt{n_1}(\hat{C} - C)$.

This implies that

$$\sqrt{n_2}(\hat{\lambda} - \mu_Y) = \frac{n_2^{-1/2}\sum_{i=1}^{n_2}\delta_i I(\delta_i > C - \mu_y + n_1^{-1/2}T)}{n_2^{-1}\sum_{i=1}^{n_2} I(Y_i > \hat{C})}. \tag{6.1}$$

With assumption (A_4), the key in this proof is that

$$
\begin{aligned}
&n_2^{-1/2}\sum_{i=1}^{n_2}\delta_i[I(\delta_i > C - \mu_X + n_1^{-1/2}T) - I(\delta_i > C - \mu_Y)]\\
=&-n_2^{-1/2}\sum_{i=1}^{n_2}\delta_i[I(\delta_i \leq C - \mu_Y + n_1^{-1/2}T) - I(\delta_i \leq C - \mu_Y)]\\
=&-(C - \mu_X)g_y(C - \mu_Y)\sqrt{h}T + o_p(1) \qquad (6.2)
\end{aligned}
$$

which may seen in Ruppert and Carroll (1980) and Chen and Chiang (1996). The Bahadur representation of the outlier mean $\hat{\lambda}$ may be formulated from

Assumption (A_1), equation (6.1), (6.2) and the following representation of empirical quantile

$$\begin{aligned}
&\sqrt{n_1}(\hat{F}_{\epsilon}^{-1}(\alpha) - F_{\epsilon}^{-1}(\alpha)) \\
=&f_{\epsilon}^{-1}(F_{\epsilon}^{-1}(\alpha))n_1^{-1/2}\sum_{i=1}^{n_1}[\alpha - I(\epsilon_i \leq F_{\epsilon}^{-1}(\alpha))] + o_p(1) \qquad (6.3)
\end{aligned}$$

see, for example, Ruppert and Carroll (1980). The asymptotic distribution in (b) of Theorem 2.3 is induced from the Central Limit Theorem. □

The proof of Theorem 3.1 is exactly identical to it of Theorem 2.3 with replacing μ_Y by μ_X, δ_i by ϵ_i and Y_i by X_i. Hence, it is skipped.

Reference

1. Agrawal, D., Chen, T., Irby, R., et al. (2002). Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *J. Natl. Cancer Inst.*, **94**, 513-521.

2. Alizadeh, A. A., Eisen, M. B., Davis, R. E., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

3. Chen, L.-A., Chen, Dung-Tsa and Chan, Wenyaw. (2008). The p Value for the Outlier Sum in Differential Gene Expression Analysis. Submitted to *Biometrika* for publication (In revision).

4. 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.

5. Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*, Wiley: New York.

6. Ohki, R., Yamamoto, K., Ueno, S., et al. (2005). Gene expression profiling of human atrial myocardium with atrial fibrillation by DNA microarray analysis. *Int. J. Cardiol.*, **102**, 233-238.

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

8. Sorlie, T., Tibshirani, R., Parker, J., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 8418-8423.

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

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

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