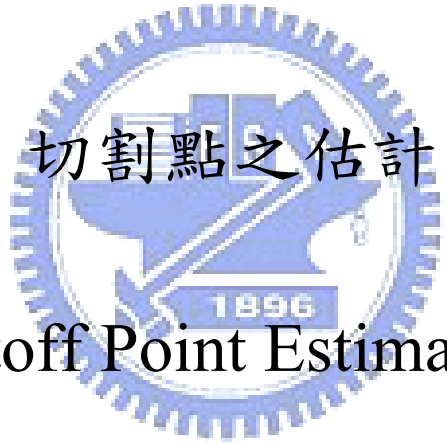
國 立 交 通 大 學

統計學研究所

碩 士 論 文



切割點之估計

Cutoff Point Estimation

研 究 生：侯智飛

指導教授：陳鄰安　博士

中 華 民 國 九 十 八 年 六 月

切割點之估計

Cutoff Point Estimation

研 究 生：侯智飛　　Student：Zhi-Fei Hou

指導教授：陳鄰安　　Advisor：Dr. Lin-An Chen

國 立 交 通 大 學

統計所研究所

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

中華民國九十八年六月

切割點之估計

研究生：侯智飛　　指導教授：陳鄰安 教授

國立交通大學統計學研究所

摘　要

切割點在建構使用基因影響表現分析的離群和或離群平均上扮演重要角色。我們在這篇論文中考慮了切割點的估計，其中樣本切割點估計量的近似分配，我們討論一種是根據經驗分位數來推導的，另一個是根據 Chen 和 Chiang(1996)發展的對稱分位數來推導的。在檢測離群值的近似變異數和檢定力顯示由對稱分位數估計的樣本切割點跟經驗分位數比較起來是非常有競爭力的。

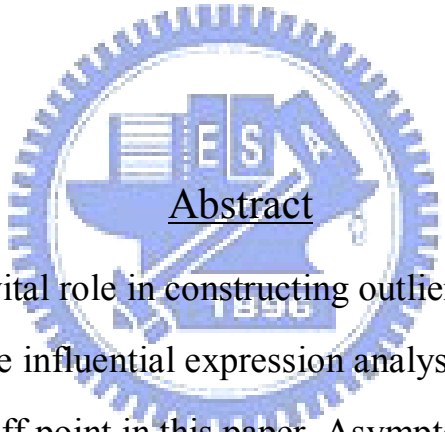
關鍵字：切割點；經驗分位數；檢定力比較；對稱分位數。

Cutoff Point Estimation

Student：Zhi-Fei Hou　　Advisor : Dr. Lin-An Chen

Institute of Statistics

National Chiao Tung University



Abstract

Cutoff point plays a vital role in constructing outlier sum or outlier mean which is used for gene influential expression analysis. We consider the estimation of the cutoff point in this paper. Asymptotic distributions of sample cutoff point estimates, one based on empirical quantiles and one based on symmetric quantiles of Chen and Chiang(1996), are developed. Comparisons of asymptotic variance and power for detecting outliers are performed showing that the version of sample cutoff point based on symmetric quantiles is very competitive with the one based on the empirical quantile.

Key words: Cutoff points; empirical quantile; power comparison; symmetric quantile.

誌謝

在研究所的這二年期間，真得非常感謝所上教授們的指導及照顧，所上的每個老師都很和藹可親，就像爸爸媽媽一樣，讓我覺得相處起來沒有壓力，使我可以在這期間可以順順利利地度過，也讓我在這期間學習了很多統計分析的技巧，更學習了許多統計相關軟體，讓我對統計有更深一層的認識，讓我帶著很多豐碩的知識離開學校。

也非常感謝我的指導教授-陳鄰安教授，對本來開始對寫論文非常害怕且懵懵懂懂的我，一步一步靠著老師的指導，讓我對寫論文不再害怕，並努力完成它，從老師身上學到很多東西，不僅是論文上的研究，還有一些為人處世的真理，老師都會不吝地教給我們，真的非常開心陳鄰安教授能當我的指導教授。

也非常慶幸在研究所時認識了一群不錯的同學，常常一起討論功課，研究作業，大考完還會一起約出去放鬆心情，是在課業壓力外的一個調劑，這群同學更能在我心情煩悶時給我安慰，在我開心時陪我大笑，在我沮喪時給我鼓勵，讓我的研究所生活過得多彩多姿。

最後，也感謝我的家人，常常鼓勵著我，讓我有一直往前的動力，而不會退縮不前進，讓我可以順利完成研究所的學業。

侯智飛　謹誌于

國立交通大學統計學研究所

中華民國九十八年六月

Contents

Cutoff Point Estimation

Abstract

Cutoff point plays a vital role in constructing outlier sum or outlier mean which is used for gene influential expression analysis. We consider the estimation of the cutoff point in this paper. Asymptotic distributions of sample cutoff point estimates, one based on empirical quantiles and one based on symmetric quantiles of Chen and Chiang (1996), are developed. Comparisons of asymptotic variance and power for detecting outliers are performed showing that the version of sample cutoff point based on symmetric quantiles is very competitive with the one based on the empirical quantile.

Key words: Cutoff point; empirical quantile; power comparison; symmetric quantile.

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)). 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.

Among statistical approaches proposed to identify those genes where only a subset of the sample genes has high expression, Tibshirani and Hastie (2007) and Wu (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 75% percentile and the interquartile range of the same gene. They also showed that the statistical test based on this outlier sum is noticeably more poweful in simulation. The distribution theory of an outlier mean,

modified from the outlier sum, has been studied by Chen, Chen and Chan (2008).

Basically an outlier is an observation that lies an essential distance from the mass of data in a random sample from a population and an outlier sum or outlier mean uses cutoff point $F^{-1}(0.5)+kIQR$ for some $k > 0$ to detect the upper-tail outliers where $IQR = F^{-1}(0.75) - F^{-1}(0.25)$ is the interquartile range. The cutoff point should be estimated from the observation when the distribution function F is unknown. With the fact that a cutoff point plays a vital role in detection of outliers, there are two concerns about estimation of the population cutoff point. First, from the point of estimation of an unknown parameter, we concern the estimator's efficiency in estimating the population cutoff point. Second, for its role of detecting influential genes, we concern the power in detecting outliers of an estimator when a distributional shift occurs. We consider a step on tackling these two concerns that help in advancing study of outlier mean for gene expression analysis.

The empirical quantile has long been the popular choice whenever estimation of population quantile is needed in constructing location and scale estimators. It is desired to see if there is competitive alternative choice of cutoff point estimator through other choice of estimation of the population quantile. This is the first step for the concerns. In order to improve the efficiency of a location estimator, the trimmed mean, Kim (1992) developed the metrically trimmed mean for a location model which, through comparison of asymptotic variances, was shown to be more efficient than the ordinary trimmed mean. Later, Chen and Chiang (1996) defined the symmetric quantile and used it to propose the symmetric trimmed mean as an extension of Kim's trimmed mean to the linear regression model. They observed that this symmetric trimmed mean of small trimming percentages can have asymptotic variances very close to the Crammer-Rao lower bounds when regression errors obey heavy tail distributions.

For solving our concerns, one interesting question is to see if the efficiency of symmetric trimmed mean can carry over to other quantile-based proposals. This is the topic that we want to investigate in this paper.

2. Symmetric and Classical Cutoff Points

In gene expression analysis, there are m genes to be concerned and for each gene there are two groups of subjects, one normal or healthy group and one cancer (disease) group. For a given gene, we assume that there are available n and m expression variables, respectively, for two groups forming as follows:

$$\begin{array}{cc} \text{Normal group} & \text{Cancer group} \\ X_1, ..., X_n & Y_1, ..., Y_m \end{array} \tag{2.1}$$

The test statistics been seen in literature to detect cancer genes is constructed based on an outlier sum of the form

$$\sum_{i=1}^{m} Y_i I(Y_i \geq \hat{C}),$$

when cancer genes are over-expressed and of the form

$$\sum_{i=1}^{m} Y_i I(Y_i \leq \hat{C})$$

when cancer genes are down-expressed where $\hat{C}$ is estimator of a cutoff point C, varying in over- and down-expressed cancer genes. Let us restrict on cutoff point with over-expressed cancer genes only. In Wu (2007) and Chen, Chen and Chang (2008), the cutoff point is $C = F_x^{-1}(0.75) + IQR$ with $IQR = F_x^{-1}(0.75) - F_x^{-1}(0.25)$, the interquartile range, constructed from the distribution function F_x of random variable X and Tibshirani and Hastie (2007) considered cutoff point constructed based on a combined disribution of random variables X and Y.

Given a population cutoff point, the efficiency of an outlier sum or outlier mean is then seriously dependent on the quality of the estimator of the unknown cutoff point. We raise this estimation question and consider two types of cutoff point estimators for comparison of asymptotic variances and powers.

Let us denote $(1-\alpha)$-central range $CR = F_x^{-1}(1-\frac{\alpha}{2}) - F_x^{-1}(\frac{\alpha}{2})$, the range of central $(1-\alpha)$ quantile interval $(F_x^{-1}(\frac{\alpha}{2}), F_x^{-1}(1-\frac{\alpha}{2}))$. The following two

formulations of population cutoff points are popularly used for identification of outlier observations:

$$C_a(1-\alpha) = F_x^{-1}(1-\frac{\alpha}{2}) + CR$$
$$= 2F_x^{-1}(1-\frac{\alpha}{2}) - F_x^{-1}(\frac{\alpha}{2})$$

and

$$C_b(1-\alpha) = F_x^{-1}(1-\frac{\alpha}{2}) + 1.5CR$$
$$= 2.5F_x^{-1}(1-\frac{\alpha}{2}) - 1.5F_x^{-1}(\frac{\alpha}{2}).$$

In case that $1-\alpha = 0.5$, the 0.5-CR is the interquartile range IQR. Let us call $C_a(1-\alpha)$ the type I cutoff point and $C_b(1-\alpha)$ the type II cutoff point. For estimation of cutoff points, we assume that there are a random sample $X_1, ..., X_n$ showing in (2.1) drawn from distribution F_x and we need to specify one estimator of $C_a(1-\alpha)$ or $C_b(1-\alpha)$.

3. Cutoff Points Estimators Based on Empirical Quantiles and Symmetric Quantiles

Classically the population quantile function F_x^{-1} is estimated by the empirical quantile F_n^{-1}. We call

$$\hat{C}_a(1-\alpha) = 2F_n^{-1}(1-\frac{\alpha}{2}) - F_n^{-1}(\frac{\alpha}{2})$$

the empirical quantile based type I cutoff point estimator and

$$\hat{C}_b(1-\alpha) = 2.5F_n^{-1}(1-\frac{\alpha}{2}) - 1.5F_n^{-1}(\frac{\alpha}{2})$$

the empirical quantile based type II cutoff point estimator.

Besides the two empirical quantile based cutoff point estimators, we also propose an alternative ones constructed by symmetric quantile of Chen and Chiang (1996). The so-called symmetric quantile is formulated based on a folded distribution function. Let μ_x be a constant, known or unknown, the folded cumulative function about μ_x for random variable X is defined as

$$F_s(a) = P(|X-\mu_x| \le a), a \ge 0.$$

Then the $1-\alpha$ symmetric quantile pair defined by Chen and Chiang (1996) is

$$(F_s^-(1-\alpha), F_s^+(1-\alpha)) = (\mu - F_s^{-1}(1-\alpha), \mu + F_s^{-1}(1-\alpha))$$

where $F_s^{-1}(1-\alpha) = \inf\{a : F_s(a) \geq 1-\alpha\}$. If F_x is continuous, the $1-\alpha$ symmetric quantile pair satisfies $1-\alpha = P(F_s^-(1-\alpha) \leq X \leq F_s^+(1-\alpha))$. If we further assume that F_x is symmetric at μ_x, it can be seen that

$$F_s^-(1-\alpha) = F_x^{-1}(\frac{\alpha}{2}) \text{ and } F_s^+(1-\alpha) = F_x^{-1}(1-\frac{\alpha}{2}), \tag{3.1}$$

the classical one and the symmetric one are identical.

Two symmetric type cutoff points are analogously defined as

$$\begin{aligned} C_a^s(1-\alpha) &= F_s^+(1-\alpha) + (F_s^+(1-\alpha) - F_s^-(1-\alpha)) \\ &= 2F_s^+(1-\alpha) - F_s^-(1-\alpha) \\ &= \mu_x + 3F_s^{-1}(1-\alpha) \end{aligned}$$

and

$$\begin{aligned} C_b^s(1-\alpha) &= F_s^+(1-\alpha) + 1.5(F_s^+(1-\alpha) - F_s^-(1-\alpha)) \\ &= 2.5F_s^+(1-\alpha) - 1.5F_s^-(1-\alpha) \\ &= \mu_x + 4F_s^{-1}(1-\alpha) \end{aligned}$$

Then, if F_x is continuous and symmetric, we have

$$C_a^s(1-\alpha) = C_a(1-\alpha) \text{ and } C_b^s(1-\alpha) = C_b(1-\alpha).$$

Let $\hat{\mu}_x$ be an estimate of μ_x. We may define the sample type $1-\alpha$ symmetric quantile pair as

$$(F_{sn}^-(1-\alpha), F_{sn}^+(1-\alpha)) = (\hat{\mu}_x - F_{sn}^{-1}(1-\alpha), \hat{\mu}_x + F_{sn}^{-1}(1-\alpha)) \tag{3.2}$$

where $F_{sn}(a) = \frac{1}{n}\sum_{i=1}^n I(|y_i - \hat{\mu}| \leq a)$ is the sample type folded cumulative distribution function and $F_{sn}^{-1}(1-\alpha) = \inf\{a : F_{sn}(a) \geq 1-\alpha\}$. The sample type symmetric cutoff points are as follows

$$\begin{aligned} \hat{C}_a^s(1-\alpha) &= 2F_{sn}^+(1-\alpha) - F_{sn}^-(1-\alpha) = \hat{\mu}_x + 3F_{sn}^{-1}(1-\alpha) \\ \hat{C}_b^s(1-\alpha) &= 2.5F_{sn}^+(1-\alpha) - 1.5F_{sn}^-(1-\alpha) = \hat{\mu}_x + 4F_{sn}^{-1}(1-\alpha) \end{aligned}$$

The equality of (3.1) does not hold when the underlying distribution F is not symmetric so that there is no fair criterion to compare their corresponding sample coverage intervals. Hence, we may set the case that F is symmetric to compare the precision of these two coverage intervals through the asymptotic variances of their sample type coverage intervals.

It is desired to give a simple example to describe the construction of these two cutoff point estimates and see how the symmetric type cutoff point estimate is worth to be introduced for outlier detection.

Example 1. Suppose that we have a set of 10 observations that are ordered as

$$-5, -3, -2, -1, -0.5, 0.5, 1, 3, 50, 100. \tag{3.3}$$

We want to construct $\alpha = 0.2$ empirical and symmetric type I cutoff point estimates for identification of outliers. With $F_n^{-1}(0.1) = -5$ and $F_n^{-1}(0.9) = 50$, the $\alpha = 0.2$ empirical type I cutoff point estimate is

$$\hat{C}_a(0.8) = 2F_n^{-1}(0.9) - F_n^{-1}(0.1) = 2 \times 50 - (-5) = 105$$

For construction of symmetric cutoff point estimate, we choose sample median as the estimate of μ_x. That is,

$$\hat{\mu}_x = F_n^{-1}(0.5) = \inf\{a : \frac{1}{10}\sum_{i=1}^{10} I(x_i \le a) \ge 0.5\} = -0.5.$$

Let's denote residuals $e_i = x_i - \hat{\mu}_x, i = 1, ..., 10$. The residuals are

$$-4.5, -2.5, -1.5, -0.5, 0, 1, 1.5, 3.5, 50.5, 100.5.$$

The sample type folded cumulative distribution function is

$$F_{sn}(a) = \frac{1}{10}\sum_{i=1}^{10} I(|e_i| \le a).$$

For examples, $F_{sn}(0) = \frac{1}{10}$, $F_{sn}(1) = \frac{1}{10}[I(|-0.5| \le 1) + I(|0| \le 1) + I(|1| \le 1)] = \frac{3}{10}$. Then we have

$$\begin{aligned} F_{sn}^{-1}(0.8) &= \inf\{a : \frac{1}{10}\sum_{i=1}^{10} I(|e_i| \le a) \ge 0.8\} \\ &= 4.5. \end{aligned}$$

This indicates that the 80% symmetric coverage interval is

$$\hat{C}_a^s(0.8) = 2(\hat{\mu}_x + F_{sn}^{-1}(0.8)) - (\hat{\mu}_x - F_{sn}^{-1}(0.8))$$
$$= 2 \times (-0.5 + 4.5) - (-0.5 - 4.5) = 13.$$

We consider that the observations beyond the cutoff point estimate are classified as outliers. The empirical type I cutoff point estimate is $\hat{C}_a(0.8) = 105$ indicating that there is no observation to be classified as outlier. On the other hand, the symmetric cutoff point estimate is $\hat{C}_a^s(0.8) = 13$ indicating that there are observations $50, 100$. From the data in (3.3), the cutoff point estimate based on symmetric quantiles is quite satisfactory. □

The equality of (3.1) does not hold when the underlying distribution F_x is not symmetric so that it is not fair to compare, in any criterion, two types of sample coverage intervals. Hence, we may set the case that F_x is symmetric to compare the precision of these two coverage intervals through the asymptotic variances of their sample type coverage intervals.

4. Efficiency Comparisons for Cuoff Point Estimators

Two properties are desired to discover for two cutoff points. First, the asymptotic distributions of these two cutoff point nonparametric estimators are interesting to discover and a comparison for their asymptotic variances in estimation of the same population cutoff point is needed. Second, it is interesting to study the powers of these two cutoff points for their roles of identifying outliers. We study the first question in this section.

The following theorem introduced the asymptotic distributions of the two types of empirical cutoff point.

Theorem 4.1. (a) $n^{1/2}(\hat{C}_a(1-\alpha) - C_a(1-\alpha))$ is asymptotically normal $N(0, \sigma_{emp,a}^2)$ where

$$\sigma_{emp,a}^2 = \frac{\alpha}{2}[(\frac{\alpha}{f_x(F_x^{-1}(1-\frac{\alpha}{2}))} + \frac{\alpha-2}{2f_x(F_x^{-1}(\frac{\alpha}{2})}))^2 + (\frac{2-\alpha}{f_x(F_x^{-1}(1-\frac{\alpha}{2}))}$$
$$- \frac{\alpha}{2f_x(F_x^{-1}(\frac{\alpha}{2}))})^2] + (1-\alpha)(\frac{\alpha}{f_x(F_x^{-1}(1-\frac{\alpha}{2}))} + \frac{\alpha}{2f_x(F_x^{-1}(\frac{\alpha}{2}))})^2.$$

(b) $n^{1/2}(\hat{C}_b(1-\alpha)-C_b(1-\alpha))$ is asymptotically normal $N(0,\sigma^2_{emp,b})$ where

$$\sigma^2_{emp,b}=\frac{\alpha}{2}[(\frac{5\alpha}{4f_x(F_x^{-1}(1-\frac{\alpha}{2}))}+\frac{3(\alpha-2)}{4f_x(F_x^{-1}(\frac{\alpha}{2}))})^2+(\frac{5(2-\alpha)}{4f_x(F_x^{-1}(1-\frac{\alpha}{2}))}$$
$$-\frac{3\alpha}{4f_x(F_x^{-1}(\frac{\alpha}{2}))})^2]+(1-\alpha)(\frac{5\alpha}{4f_x(F_x^{-1}(1-\frac{\alpha}{2}))}+\frac{3\alpha}{4f_x(F_x^{-1}(\frac{\alpha}{2}))})^2.$$

To study the asymptotic distribution of the symmetric type cutoff points, we restrict to the following location models,

$$X_i=\mu_x+\epsilon_i, i=1,...,n \tag{4.1}$$

where ϵ_i's are independent and identically distributed (iid) random variables having distribution functions G_x with zero mean, variance σ_x^2 and probability density function g_x. For convenience of comparison, we also assume that G_x is symmetric at zero.

We consider that μ_x is the median parameter and let $\hat{\mu}_x$ be the sample median as

$$\hat{\mu}_x=\text{arginf}_{\mu_x\in R}\sum_{i=1}^{n}|X_i-\mu_x|.$$

Suppose that we assume that G_x is continuous and symmetric at 0. The asymptotic distributions of two symmetric cutoff points are stated in the following theorem.

Theorem 4.2. Assuming that distribution function G_x is symmetric at zero, we have the following asymptotic properties.
(a) $n^{1/2}(\hat{C}^s_a(1-\alpha)-C^s_a(1-\alpha))$ is asymptotically normal $N(0,\sigma^2_{sym,a})$ where

$$\sigma^2_{sym,a}=\frac{\alpha}{2}[(-\frac{1}{2g_x(0)}+\frac{3(1-\alpha)}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2+(\frac{1}{2g_x(0)}+\frac{3(1-\alpha)}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2]$$
$$+\frac{1}{2}(1-\alpha)[(\frac{1}{2g_x(0)}+\frac{3\alpha}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2+(\frac{1}{2g_x(0)}-\frac{3\alpha}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2]$$

(b) $n^{1/2}(\hat{C}^s_b(1-\alpha)-C^s_b(1-\alpha))$ is asymptotically normal $N(0,\sigma^2_{sym,b})$ where

$$\sigma^2_{sym,b}=\frac{\alpha}{2}[(-\frac{1}{2g_x(0)}+\frac{2(1-\alpha)}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2+(\frac{1}{2g_x(0)}+\frac{2(1-\alpha)}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2]$$
$$+\frac{1}{2}(1-\alpha)[(\frac{1}{2g_x(0)}+\frac{2\alpha}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2+(\frac{1}{2g_x(0)}-\frac{2\alpha}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))})^2]$$

With asymptotic distributions of two types of cutoff point estimators developed, we may consider several distributions for error variable ϵ for computation of their asymptotic variances to compare their efficiencies for estimating the unknown cutoff points. The distributions considered here include standard normal distribution $N(0,1)$, t-distribution $t(r)$ where r is the degrees of freedom, Cauchy distribution ($Cauchy(s), s>0$) with pdf

$$g_x(\epsilon) = \frac{1}{\pi}\frac{s}{\epsilon^2+s^2}, \epsilon \in R$$

and the Laplace distribution ($Lap(b)$) with pdf

$$g_x(\epsilon) = \frac{1}{2b}e^{-\frac{|\epsilon|}{b}}, \epsilon \in R.$$

We display the computed efficiencies in Table 1.

Table 1. Asymptotic variances for two quantile-based cutoff point estimations

α	$\sigma^2_{sym,a}$	$\sigma^2_{emp,a}$	$\sigma^2_{sym,b}$	$\sigma^2_{emp,b}$
$N(0,1)$				
0.05*	32.85	34.94	57.19	59.28
0.15*	15.88	16.18	27.02	27.32
0.25	11.52	11.43	19.26	19.17
0.35	9.274	9.020	15.26	15.01
0.45	7.761	7.441	12.57	12.25
$t(1)$				
0.05*	27838	31091	49488	52741
0.15*	955.8	1077	1697	1819
0.25*	196.6	222.9	347.6	373.9
0.35*	70.25	79.37	122.9	132.0
0.45*	33.36	37.13	57.39	61.16
$Cauchy(3)$				
0.05*	250543.8	279822.4	445394.0	474672.5
0.15*	8602.33	9701.71	15275.7	16375.1
0.25*	1769.50	2006.15	3128.51	3365.17
0.35*	632.25	714.33	1106.74	1188.81
0.45*	300.25	334.22	516.51	550.48
$Lap(1)$				
0.05*	172.0	191.0	305.0	324.0
0.15*	52.00	57.66	91.66	97.33
0.25*	28.00	31.00	49.00	52.00
0.35*	17.71	19.57	30.71	32.57
0.45*	12.00	13.22	20.55	21.77

$*$: Symmetric type cutoff points have smaller asymptotic variance than it of empirical quantile based cutoff points

We may draw several conclusions from the results in Table 1:

1. In few cases (normal distribution with $\alpha = 0.25, 0.35$ and 0.45), it is relatively more efficient estimating the cutoff point by the empirical quantiles. This indicates that when we want to estimate the population cutoff point and we know that the underlying distribution is normal the version estimated by empirical quantiles is appropriate. However, we note that although the differences between these two versions are not significant.
2. For the distributions other than the normal one, the estimate based on symmetric quantiles is simultaneously more efficient than it based on empirical quantiles. In an overall comparison, we may say that the cutoff point estimate based on symmetric quantile is a robust one.
3. In this consideration of nonparametric estimation, we may expect that any cutoff point estimator based on symmetric quantiles is a robust one.

In gene influential analysis, a common situation is that there are only few observations lie beyond the main trend of the model. Hence, to study the large sample properties of cutoff estimators for these distributions is desired. We consider the following contaminated normal distribution

$$\epsilon = (1-\delta)H + \delta N(h, 1) \tag{4.2}$$

which ensures that a large proportion of observations drawn from the same distribution under H_0 and a small proportion δ of observations are outliers. We compute the efficiencies of two cutoff point estimates defined as the followings:

$$eff_{sym} = \frac{\min\{\sigma^2_{sym,a}, \sigma^2_{emp,a}\}}{\sigma^2_{sym,a}}, eff_{emp} = \frac{\min\{\sigma^2_{sym,a}, \sigma^2_{emp,a}\}}{\sigma^2_{emp,a}}$$

A simulation results are displayed in Table 2.

Table 2. Efficiencies $\begin{pmatrix} eff_{sym} \\ eff_{emp} \end{pmatrix}$ of type I cutoff point estimates by symmetric quantile and empirical quantile ($h = 1$)

H	$\alpha = 0.05$	0.15	0.25	0.35	0.45
$H = N(0,1)$					
$\delta = 0.05$	$\begin{pmatrix} 1 \\ 0.955 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.989 \end{pmatrix}$	$\begin{pmatrix} 0.987 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.971 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.959 \\ 1 \end{pmatrix}$
$\delta = 0.1$	$\begin{pmatrix} 1 \\ 0.963 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.994 \end{pmatrix}$	$\begin{pmatrix} 0.984 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.970 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.959 \\ 1 \end{pmatrix}$
$H = t(1)$					
$\delta = 0.05$	$\begin{pmatrix} 1 \\ 0.895 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.883 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.857 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.880 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.906 \end{pmatrix}$
$\delta = 0.1$	$\begin{pmatrix} 1 \\ 0.895 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.874 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.835 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.878 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.915 \end{pmatrix}$
$H = Lab(1)$					
$\delta = 0.05$	$\begin{pmatrix} 1 \\ 0.879 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.889 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.906 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.917 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.922 \end{pmatrix}$
$\delta = 0.1$	$\begin{pmatrix} 1 \\ 0.859 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.878 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.909 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.928 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.935 \end{pmatrix}$

Several comments may be drawn from the results in Table 2:

1. The effiicencies of the symmetric type cutoff point estimate although has efficiencies smaller than one on situations that H is normal and $\alpha \geq 0.25$, however, they are at least more than 0.95.

2. The effiicencies of the empirical type cutoff point estimate are with efficiencies smaller than one on all distributions other than normal one and it can be as small as 0.835. In comparison for this contaminated distribution, the symmetric type cutoff point estimate is also a robust proposal.

3. Since our study of cutoff point estimation is primarily motivated from the gene influential analysis and, in this analysis, it often faces few extreme outliers in the treatment group that is a type of contaminated distribution, this comparison shows that the symmetric type cutoff point estimate is an appropriate choice for outlier detection of this analysis.

We also consider the two estimators of the type II cutoff point that we display their efficiencies in Table 2.

Table 3. Efficiencies $\begin{pmatrix} eff_{sym} \\ eff_{emp} \end{pmatrix}$ of type II cutoff point estimates by symmetric quantile and empirical quantile ($h = 1$)

H	$\alpha = 0.05$	0.15	0.25	0.35	0.45
$H = N(0,1)$					
$\delta = 0.05$	$\begin{pmatrix} 1 \\ 0.988 \end{pmatrix}$	$\begin{pmatrix} 0.996 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.985 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.977 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.971 \\ 1 \end{pmatrix}$
$\delta = 0.1$	$\begin{pmatrix} 0.998 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.986 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.978 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.972 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.968 \\ 1 \end{pmatrix}$
$H = t(1)$					
$\delta = 0.05$	$\begin{pmatrix} 1 \\ 0.938 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.927 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.890 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.917 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.946 \end{pmatrix}$
$\delta = 0.1$	$\begin{pmatrix} 1 \\ 0.938 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.914 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.854 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.906 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.952 \end{pmatrix}$
$H = Lab(1)$					
$\delta = 0.05$	$\begin{pmatrix} 1 \\ 0.911 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.920 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.942 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.955 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.962 \end{pmatrix}$
$\delta = 0.1$	$\begin{pmatrix} 1 \\ 0.881 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.899 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.939 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.964 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.977 \end{pmatrix}$

From the results showing in Table 2, the efficiencies performed by two quantile methods are quite similar to those showing in Table 1.

5. Power Comparisons for Cutoff Point Estimators

With asymptotic distributions of two sample cutoff point estimates, it is desired to study a cutoff point for its ability to detecting an observation drawn from an alternative distribution. Consider that we have the following sample location model

$$X_i = \mu_x + \epsilon_i, i = 1, .., n$$

and we have a random variable Y drawn from an alternative distribution. Let C be a cutoff point and $\hat{C}$ be an estimator of C constructed from the sample of variable X. The power of detection of outlier Y is

$$\pi = P\{Y \geq \hat{C}\}. \tag{5.1}$$

Suppose that $\sqrt{n}(\hat{C} - C)$ converges, in distribution to a normal distribution $N(0, \sigma_c^2)$. An approximate power is

$$\begin{aligned} \pi &= P\{Y \geq \hat{C}\} \\ &\approx P\{\sqrt{n}Y - \sqrt{n}(\hat{C} - C) \geq \sqrt{n}C\} \\ &= P\{\sqrt{n}Y - N_c \geq \sqrt{n}C\} \end{aligned}$$

where N_c is a random variable with distribution $N(0, \sigma_c^2)$. If a distribution of the combination $\sqrt{n}Y - N_c$ is available, the approximate power then can be computed. It is then interesting to compare the powers for cutoff points estimated from empirical quantile and symmetric quantile. We further compute the powers of symmetric cutoff point and empirical cutoff point, respectively denoted by π_{sym} and π_{emp}. We note that π_{sym} and π_{emp} are very close values in all the cases. However, it is still worthy to compare their sizes.

In the following two tables, we display the comparisons of π_{sym} and π_{emp} for type I and type II symmetric and empirical cutoff point estimators when the sample is drawn from several distributions of interest.

Table 4. Power comparison of type I symmetric and empirical cutoff point estimators ($n = 30$)

	$\pi_{sym} > \pi_{emp}$	$\pi_{sym} < \pi_{emp}$
$X \sim t(r_1)$,		
$Y \sim t(r_2) + \mu$		
$r_1 = r_2 = 1$		
$\alpha = 0.2$	$\mu = 3, ..., 10$	$\mu = 0.5, 1, 2$
$\alpha = 0.5$	$\mu = 3, ..., 10$	$\mu = 0.5, 1, 2$
$r_1 = 3, r_2 = 1$		
$\alpha = 0.2$	$\mu = 5, ..., 10$	$\mu = 0.5, 1, 2, 3, 4$
$\alpha = 0.5$	$\mu = 3, 4, .., 10$	$\mu = 0.5, 1, 2$
$X \sim N(0, 1)$,		
$Y \sim N(0, 1) + \mu$		
$\alpha = 0.1$	$\mu = 5, 6, 7, 8, 9, 10$	$\mu = 0.5, 1, 2, 3, 4$
$\alpha = 0.2$	$\mu = 4, 5, 6, 7, 8, 9, 10$	$\mu = 0.5, 1, 2, 3$
$\alpha = 0.5$	$\mu = 0.5, 1, 2$	$\mu = 3, 4, 5, 6, 7$
$X \sim N(0, 1)$,		
$Y \sim (1-\gamma)N(0, 1) + \gamma N(\mu, 1)$		
$\gamma = 0.05, \alpha = 0.5$	$\mu = 0.5, 1, ..., 10$	none
$\gamma = 0.05, \alpha = 0.2$	none	$\mu = 0.5, 1, ..., 10$
$\gamma = 0.1, \alpha = 0.5$	$\mu = 0.5, 1, ..., 10$	none
$\gamma = 0.1, \alpha = 0.2$	none	$\mu = 0.5, 1, ..., 10$
$\gamma = 0.2, \alpha = 0.5$	$\mu = 0.5, 1, ..., 10$	none
$\gamma = 0.2, \alpha = 0.2$	none	$\mu = 0.5, 1, ..., 10$

Table 5. Power comparison of type II symmetric and empirical cutoff point estimators ($n = 30$)

	$\pi_{sym} > \pi_{emp}$	$\pi_{sym} < \pi_{emp}$
$X \sim t(r_1)$,		
$Y \sim t(r_2) + \mu$		
$r_1 = 3, r_2 = 1$		
$\alpha = 0.2$	$\mu = 7, ..., 10$	$\mu = 0.5, 1, 2, 3, ..., 6$
$\alpha = 0.5$	$\mu = 4, ..., 10$	$\mu = 0.5, 1, 2, 3$
$r_1 = 3, r_2 = 3$		
$\alpha = 0.2$	$\mu = 7, ..., 10$	$\mu = 0.5, 1, 2, ..., 6$
$\alpha = 0.5$	$\mu = 4, .., 10$	$\mu = 0.5, 1, 2, 3$
$X \sim N(0, 1)$,		
$Y \sim (1-\gamma)N(0,1) + \gamma N(\mu, 1)$		
$\gamma = 0.05, \alpha = 0.5$	$\mu = 0.5, 1, ..., 10$	none
$\gamma = 0.05, \alpha = 0.2$	none	$\mu = 0.5, 1, ..., 10$
$\gamma = 0.1, \alpha = 0.5$	$\mu = 0.5, 1, ..., 10$	none
$\gamma = 0.1, \alpha = 0.2$	none	$\mu = 0.5, 1, ..., 10$
$\gamma = 0.2, \alpha = 0.5$	$\mu = 0.5, 1, ..., 10$	none
$\gamma = 0.2, \alpha = 0.2$	none	$\mu = 0.5, 1, ..., 10$

From the results showing in Tables 4 and 5, the two cutoff point estimators are quite competitive. Some are better with symmetric type estimation and some are better with empirical estimation. However, from the robustness consideration, we prefer to use the symmetric type cutoff point estimatior since its estimation is more reliable with smaller asymptotic variances that showed in Section 4.

It is desired to study the power performance for these two cutoff point estimators when the outliers not only shift in both location and scale. We further consider the following contaminated normal distribution

$$(1-\gamma)N(0,1) + \gamma N(\mu, \sigma^2)$$

Table 6. Power comparison of types I and II symmetric and empirical cutoff point estimators ($n = 30$)

	$\pi_{sym} > \pi_{emp}$	$\pi_{sym} < \pi_{emp}$
$X \sim N(0,1)$, $Y \sim (1-\gamma)N(0,1)+\gamma N(\mu,\sigma^2)$		
$\gamma = 0.1, \sigma = 2, \alpha = 0.5$	$\mu = 0.5, ..., 10$	none
$\gamma = 0.1, \sigma = 2, \alpha = 0.2$	none	$\mu = 0.5, ..., 10$
$\gamma = 0.1, \sigma = 5, \alpha = 0.5$	$\mu = 0.5, ..., 10$	none
$\gamma = 0.1, \sigma = 5, \alpha = 0.2$	none	$\mu = 0.5, ..., 10$
$\gamma = 0.1, \sigma = 10, \alpha = 0.5$	$\mu = 0.5, ..., 10$	none
$\gamma = 0.1, \sigma = 10, \alpha = 0.2$	none	$\mu = 0.5, ..., 10$

6. Appendix

Proof of Theorem 4.1. From Ruppert and Carroll (1980), we have a representation of the empirical quantile $\hat{F}_x^{-1}(\alpha)$ as

$$n^{1/2}(\hat{F}_x^{-1}(\alpha)-F_x^{-1}(\alpha)) = \frac{1}{f_x(F_x^{-1}(\alpha))} n^{-1/2}\sum_{i=1}^{n}(\alpha - I(X_i \le F_x^{-1}(\alpha))) + o_p(1). \tag{6.1}$$

Since the emprical quantile based cutoff point estimates $\hat{C}_a(1-\alpha)$ and $\hat{C}_b(1-\alpha)$ are both linear functions of empirical quantiles $\hat{F}_x^{-1}(1-\alpha)$ and $\hat{F}_x^{-1}(\alpha)$, careful arrangements of the corresponding representations of these two empirical quantiles lead to the following representations.

(a) A large sample representation for cutoff point estimator $\hat{C}_a(1-\alpha)$ is as follows:

$$n^{1/2}(\hat{C}_a(1-\alpha) - C_a(1-\alpha)) = n^{-1/2}\sum_{i=1}^{n}\{[-\frac{\alpha}{f_x(F_x^{-1}(1-\frac{\alpha}{2}))} - \frac{\alpha-2}{2f_x(F_x^{-1}(\frac{\alpha}{2})}]$$
$$I(X_i \le F_x^{-1}(\frac{\alpha}{2})) + [-\frac{\alpha}{f_x(F_x^{-1}(1-\frac{\alpha}{2}))} - \frac{\alpha}{2f_x(F_x^{-1}(\frac{\alpha}{2}))}]$$
$$I(F_x^{-1}(\frac{\alpha}{2}) \le X_i \le F_x^{-1}(1-\frac{\alpha}{2})) + [\frac{2-\alpha}{f_x(F_x^{-1}(1-\frac{\alpha}{2}))} - \frac{\alpha}{2f_x(F_x^{-1}(\frac{\alpha}{2}))}]$$
$$I(X_i \ge F_x^{-1}(1-\frac{\alpha}{2}))\} + o_p(1)$$

(b) A large sample representation for cutoff point estimator $\hat{C}_b(1-\alpha)$ is as

follows:

$$n^{1/2}(\hat{C}_b(1-\alpha) - C_b(1-\alpha)) = n^{-1/2}\sum_{i=1}^{n}\{[-\frac{5\alpha}{4f_x(F_x^{-1}(1-\frac{\alpha}{2}))} - \frac{3(\alpha-2)}{4f_x(F_x^{-1}(\frac{\alpha}{2})}]$$
$$I(X_i \le F_x^{-1}(\frac{\alpha}{2})) + [-\frac{5\alpha}{4f_x(F_x^{-1}(1-\frac{\alpha}{2}))} - \frac{3\alpha}{4f_x(F_x^{-1}(\frac{\alpha}{2}))}]$$
$$I(F_x^{-1}(\frac{\alpha}{2}) \le X_i \le F_x^{-1}(1-\frac{\alpha}{2})) + [\frac{5(2-\alpha)}{4f_x(F_x^{-1}(1-\frac{\alpha}{2}))} - \frac{3\alpha}{4f_x(F_x^{-1}(\frac{\alpha}{2}))}]$$
$$I(X_i \ge F_x^{-1}(1-\frac{\alpha}{2}))\} + o_p(1).$$

The theorem is induced from the central limit theorem. □

Proof of Theorem 4.2. Again, from Ruppert and Carroll (1980), we have a representation for this sample median as

$$n^{1/2}(\hat{\mu}_x - \mu_x) = n^{-1/2}\frac{1}{g_x(0)}\sum_{i=1}^{n}(0.5 - I(\epsilon_i \le 0)) + o_p(1). \tag{6.2}$$

On the other hand, a Barhadur representation for $F_{sn}^{-1}(1-\alpha)$ developed by Chen and Chiang (1996) is

$$n^{1/2}(F_{sn}^{-1}(1-\alpha) - (F_x^{-1}(1-\frac{\alpha}{2}) - \mu_x)) = \frac{1}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))}n^{-1/2}\sum_{i=1}^{n}\{1-\alpha$$
$$- I(G_x^{-1}(\frac{\alpha}{2}) \le \epsilon_i \le G_x^{-1}(1-\frac{\alpha}{2})\} + o_p(1). \tag{6.3}$$

With, again, careful arrangements of representations of (6.3), we can derive representations of symmetric type cutoff point estimates $\hat{C}_a^s(1-\alpha)$ and $\hat{C}_b^s(1-\alpha)$.

A large sample representation for outlier mean $\hat{C}_a^s(1-\alpha)$ is as follows:

$$n^{1/2}(\hat{C}_a^s(1-\alpha) - C_a^s(1-\alpha)) = n^{-1/2}\sum_{i=1}^{n}\{[-\frac{1}{2g_x(0)} + \frac{3(1-\alpha)}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]$$
$$I(\epsilon_i \le G_x^{-1}(\frac{\alpha}{2})) + [-\frac{1}{2g_x(0)} - \frac{3\alpha}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]I(G_x^{-1}(\frac{\alpha}{2}) \le \epsilon_i \le 0)$$
$$+ [\frac{1}{2g_x(0)} - \frac{3\alpha}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]I(0 \le \epsilon_i \le G_x^{-1}(1-\frac{\alpha}{2})) + [\frac{1}{2g_x(0)}$$
$$+ \frac{3(1-\alpha)}{2g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]I(\epsilon_i \ge G_x^{-1}(1-\frac{\alpha}{2})) + o_p(1)$$

A large sample representation for outlier mean $\hat{C}_b^s(1-\alpha)$ is as follows:

$$n^{1/2}(\hat{C}_b^s(1-\alpha) - C_b^s(1-\alpha)) = n^{-1/2}\sum_{i=1}^{n}\{[-\frac{1}{2g_x(0)} + \frac{2(1-\alpha)}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]$$
$$I(\epsilon_i \le G_x^{-1}(\frac{\alpha}{2})) + [-\frac{1}{2g_x(0)} - \frac{2\alpha}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]I(G_x^{-1}(\frac{\alpha}{2}) \le \epsilon_i \le 0)$$
$$+ [\frac{1}{2g_x(0)} - \frac{2\alpha}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]I(0 \le \epsilon_i \le G_x^{-1}(1-\frac{\alpha}{2})) + [\frac{1}{2g_x(0)}$$
$$+ \frac{2(1-\alpha)}{g_x(G_x^{-1}(1-\frac{\alpha}{2}))}]I(\epsilon_i \ge G_x^{-1}(1-\frac{\alpha}{2})) + o_p(1).$$

The theorem is induced from the central limit theorem for the above two representations, respectively, for $\hat{C}_a^s(1-\alpha)$ and $\hat{C}_b^s(1-\alpha)$. □

References

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.

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.

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, D.-T. and Chan, W. (2008). p value for outlier sum in differential gene expression analysis. Submitted to *Biometrika* for publication (In revision).

Kim, S. J. (1992). The metrically trimmed means as a robust estimator of location, *Annals of Statistics.* 20, 1534-1547.

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.

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

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

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.