國立交通大學

統計學研究所

碩士論文

離群比例之基因表現分析

Outlier Proportion for Gene Expression Analysis

研 究 生: 徐國誠

指導教授: 陳鄰安 教授

中華民國一百零一年六月

離群比例之基因表現分析

Outlier Proportion for Gene Expression Analysis

研究生:徐國誠　　student：Guo-Cheng Shu

指導教授:陳鄰安 教授　　Advisor：Lin-An Chen

國 立 交 通 大 學 理 學 院

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 2012

Hsinchu, Taiwan, Republic of China

中 華 民 國 一 百 零 一 年 六 月

Outlier Proportion for Gene Expression Analysis

Student： Guo-Cheng Shu Adviser： Lin-An Chen

Institute of Statistics

National Chaio Tung University

Abstract

Discovering the influential genes through the detection of outliers in samples from disease group subjects is a very new and important approach for gene expression analysis. Extended the outlier mean of Chen. Chen and Chan(2010), we develop the asymptotic distribution of the outlier proportion for linear regression model. Power comparison shows that tests based on this outlier estimator is very competitive and promising in detecting a shift of parent tail distribution.

Key words： Cancer outlier profile analysis、 Gene expression、 Outlier mean、 Outlier proportion、 Regression model.

離群比例之基因表現分析

研究生:徐國誠　　指導教授:陳鄰安 教授

國立交通大學理學院

統計學研究所

摘要

從有病的樣本中透過離群值的檢測,發現有影響力的基因是目前一個很新也很重要的基因表現分析,延續 Chen. Chen and Chan(2010) 離群平均的想法,我們開發了以迴歸模型為基礎的離群比例的漸近分佈,比較其檢定力後發現,以這離群估計為基礎的測試是非常有競爭力,並有望檢測母體背後分佈的轉變。

關鍵詞: 癌症離群比例分析、基因表現、離群的平均、離群的比例、迴歸模型

誌謝

很榮幸能成為陳鄰安 教授的指導學生。在跟老師相處的一年多時間，教導我的不僅僅是論文上的研究。並且讓我學習到:如何將所學得之知識加以應用並解決問題。相信這對我對未來在面臨更大的挑戰時，一定能更有效率、邏輯地去思考解決。再來要感謝口試委員:許文郁 教授、蕭金福 教授、以及彭南夫 教授 在口試時對於我論文提出更佳之建議及修改方向。 使我此篇研究結果能更加完整。

在碩士班兩年的期間，我要感謝409 研究室的所有同學。當我在程式上遇到難題時，研究室的同學總會熱心地一起協助陪我渡過難關。才能使我論文如期完成。且在課餘閒暇之時，和研究室的同學一起去球場運動。讓我的課業壓力得以適當地釋放。讓我覺得在碩士生涯的兩年期間過的相當愉快。

最後，我要感謝最支持我的父母，及弟弟。在我忙碌、壓力大時，總會給我適時的鼓勵及安慰。讓我能全心全意地專注在研究上。離開學校後，自己會更加努力、專注去做好每一件事。

徐國誠 謹誌于

國立交通大學統計學研究所

中華民國一百零一年六月

Contents

List of Tables

1. Introduction.

Among the existing 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 and a limited number of subjects available. Tomlins et al. (2005) observed in a study of prostate cancer that differential genes are over expressed in a small number of disease samples. The problem of constructing statistical procedures based on outlier samples has been attracted considerable recent attention. 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 be applied. Recently Chen, Chen and Chan (2010) considered the outlier mean (average of outlier sum) and developed its large sample theory that allows us to formulate a distribution based p value. Simulation study and data analysis show desired efficiency for tests based on outlier mean.

Uncertainties of gene expressions also show causal effect upon one or some biological conditions (independent variables, see Jin, Si et al. (2006), Huang and Pan (2003), Rambow, Piton et al. (2008) and Muller, Chiou and Leng (2008)). From their observation, Tomlins et al. (2005), investigating and verifying the characteristics of the parent tail distribution of the disease group data in linear regression models through estimation and hypothesis testing is new but important topic to be explored.

We consider the sample conditional quantile based on healthy group data as cutoff for determination of outliers and introduce the sample proportion computed from these outliers for monitoring the outlier distribution and present its asymptotic distribution. With simulation study, the outlier proportion based test is shown desirable in terms of powers.

2. Formalization of Outlier Proportion

We consider that there are two gene expression variables y_a and y_b, respectively, for normal (control) group population and disease group popu-

lation that follows linear regression models as

$$y_a = x'\beta_a + \epsilon \text{ and } y_b = x'\beta_b + \delta$$

where x is p-vector of covariates (biological conditions) with constant one on the first element. A key to the challenge of quantifying outlier information in model for gene response variable y_b is to reparametrize the regression parameters that characterizes the information contained in tail distribution of this variable.

By denoting the two distributions for y_a and y_b as $F_{y_a}(.|x)$ and $F_{y_b}(.|x)$ when vector x is given, the main objective for conduction of gene expression analysis is to perform a test for hypothesis of distributional equality as

$$H_F : F_{y_a}(.|x) = F_{y_b}(.|x), x \in R^+ \tag{2.1}$$

where $R^+ = \{(1, x_1')' : x_1 \in R^{p-1}\}$. The classical approach conducts this testing this hypothesis through verifying if there are equal conditional means as

$$x'\beta_b = x'\beta_a, x \in R^+ \tag{2.2}$$

which is equivalent to verify if $\beta_b = \beta_a$ is true. Following their observation, Tomlins et al. (2005) proposed to verify the parameters in outlier distribution instead of original distributions F_{y_a} and F_{y_b}. This requires a formalization of regression parameters that contains the information in tail distribution.

Given a fixed x, we consider the γ-th quantile of variable y_a as the cutoff for outlier detection threshold that may be written as $F_{y_a}^{-1}(\gamma) = x'\beta_a(\gamma)$ where $\beta_a(\gamma) = \beta_a + F_\epsilon^{-1}(\gamma)e$ is the population regression quantile of Koenker and Bassett (1978) and where p-vector $e = (1, 0, ..., 0)'$ and $F_\epsilon^{-1}(\gamma)$ is the γ-th quantile for error variable ϵ with distribution function F_ϵ. Observations of y_a and y_b over this quantile point are considered outliers. Our idea based on Tomlins et al.'s observation that is extended from outlier mean to regression model for testing hypothesis (2.1) is through a verification on variable y_b's conditional outlier proportion $\lambda_{b,out}(x) = P(y_b \geq x'\beta_a(\gamma))$.

We may see that

$$\lambda_{b,out}(x) = P(\delta \geq F_{\epsilon}^{-1}(\gamma) + x'(\beta_a - \beta_b)) \tag{2.3}$$

and the variable y_a's conditional outlier proportion $\lambda_{a,out}(x) = P(y_a \geq x'\beta_a(\gamma)) = 1 - \gamma$. For testing hypothesis (2.1), we propose to verify the following relation:

$$\lambda_{b,out}(x) = 1 - \gamma, x \in R^{+}. \tag{2.4}$$

By denoting the difference of two conditional outlier means as $D_{op}(x) = \gamma_{b,out}(x) - (1 - \gamma)$. For validation of this verification, considering the following model settings,

$$y_a = 1 + 2x + \epsilon \text{ and } y_b = 1.1 + 2.1x + \delta$$
$$\epsilon \sim N(0,1) \text{ and } \delta \sim \lambda N(0,1) + (1-\lambda)N(\mu,1)$$

and with sample size $n = 100$, we display the sizes $D_{op}(x)$ and efficiency $eff = D_{op}/\lambda_{a,out}$ in Table 1.

Table 1. Outlier proportion differences $\begin{pmatrix} D_{op} \\ eff \end{pmatrix}$

	$\gamma = 0.1$	$\gamma = 0.3$	$\gamma = 0.6$	$\gamma = 0.9$
$\mu = 1$				
$x = 1$	0.037 (0.041)	0.084 (0.121)	0.113 (0.284)	0.072 (0.725)
$x = 2$	0.048 (0.053)	0.112 (0.160)	0.152 (0.380)	0.098 (0.976)
$x = 3$	0.058 (0.064)	0.137 (0.196)	0.190 (0.478)	0.125 (1.248)
$x = 4$	0.066 (0.073)	0.160 (0.229)	0.227 (0.568)	0.154 (1.542)
$\mu = 5$				
$x = 1$	0.037 (0.042)	0.089 (0.127)	0.131 (0.327)	0.126 (1.257)
$x = 2$	0.049 (0.054)	0.116 (0.165)	0.167 (0.417)	0.147 (1.468)
$x = 3$	0.058 (0.065)	0.140 (0.200)	0.202 (0.506)	0.170 (1.701)
$x = 4$	0.066 (0.074)	0.163 (0.232)	0.238 (0.594)	0.196 (1.955)

Significant differences in size between $D_{om}(x)$ and $D_m(x)$ showing in this table reveals that detection of difference in conditional outlier means may be better in terms of power than detection of difference in non-outlier conditional means.

We have observed the positive sign of using variable y_b's outlier proportion for gene expression analysis. However, consistent estimator of outlier proportion is too complicated since it requires to build up a consistent estimator of x-related parameter function $\lambda_{b,out}(x)$ in (2.3). The difficulty can be solved if a reformalization of regression parameters can be done. By letting $\beta_b = \begin{pmatrix} \beta_{b0} \\ \beta_{b1} \end{pmatrix}$ and $\beta_a = \begin{pmatrix} \beta_{a0} \\ \beta_{a1} \end{pmatrix}$ with β_{b0} and β_{a0} the intercept parameters and β_{b1} and β_{a1} being, respectively, vectors of slope parameters, we then set the following restriction

$$\beta_{a1} = \beta_{b1} \tag{2.5}$$

which is true when H_0 of (2.1) is true. This allows us to write outlier proportion as

$$\lambda_{b,out} = P(\delta \geq F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0}). \tag{2.6}$$

The next objective is establishing an estimator for the outlier proportion $\lambda_{b,out}$ of (2.6) and developing its distributional theory for construction of a test for hypothesis (2.1).

3. Outlier Proportion Estimator and Its Large Sample Theory

For this gene expression study, we assume that there are 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. The gene expressions for normal group subject have the regression model

$$y_{ai} = x_i'\beta_a + \epsilon_i, i = 1, ..., n_1 \tag{3.1}$$

where ϵ_i's are iid error variables with distribution function F_ϵ and the disease group subject have the regression model

$$y_{bi} = x_i'\beta_b + \delta_i, i = 1, ..., n_2 \tag{3.2}$$

where δ_i's are iid error variables with distribution function F_δ.

We let the sample threshold be $\hat{F}^{-1}_{y_a}(\gamma) = x'\hat{\beta}_a(\gamma)$ where $\hat{\beta}_a(\gamma)$ is the sample regression quantile of Koenker and Bassett (1978) that solves

$$\text{Min}_{b\in R^p} \sum_{i=1}^{n_1} (y_{ai} - x_i'b)(\gamma - I(y_{ai} \leq x_i'b)).$$

The estimator of the outlier proportion is

$$\hat{\lambda}_{b,out} = n_2^{-1} \sum_{i=1}^{n_2} I(y_{bi} \geq x_i'\hat{\beta}_a(\gamma)). \tag{3.3}$$

We consider a simulation study for efficiency of sample outlier proportion with the following models:

$$\begin{aligned} y_{ai} =& 1 + 2x_i + \epsilon_i, i = 1, ..., n_1 \text{ where } \epsilon_i's \text{ are iid } N(0,1), \text{ and} \\ y_{bi} =& \beta_0 + 2x_i + \delta_i, i = 1, ..., n_2 \text{ where } \delta_i's \text{ are iid } \sim F_\delta. \end{aligned} \tag{3.4}$$

Under $F_\delta = 0.9N(0,1) + 0.1N(1,1)$, we perform replications $m = 1,000$ with $n = n_1 = n_2$ from the above models and display the MSE's in Table 2 while their corresponding true outlier proportions are listed in ( )'s with $n = 50$.

Table 2. Mean square error for outlier proportion estimation

γ	$\beta_0 = 1.1$	$\beta_0 = 1.3$	$\beta_0 = 1.5$
$n = 50$			
0.7	0.0099 (0.3738)	0.0110 (0.4482)	0.0111 (0.5247)
0.8	0.0081 (0.2664)	0.0101 (0.3323)	0.0115 (0.4041)
0.9	0.0056 (0.1496)	0.0075 (0.1975)	0.0097 (0.2541)
$n = 100$			
0.7	0.0049	0.0054	0.0053
0.8	0.0044	0.0051	0.0055
0.9	0.0029	0.0039	0.0052
$n = 200$			
0.7	0.0026	0.0027	0.0028
0.8	0.0025	0.0027	0.0029
0.9	0.0019	0.0025	0.0029
$n = 500$			
0.7	0.0012	0.0011	0.0012
0.8	0.0015	0.0014	0.0013
0.9	0.0014	0.0016	0.0016

Small MSE's shows that outlier proportion is a parameter appropriate for statistical inferences of hypothesis (2.1).

For the asymptotic properties for the outlier proportion, we need the following assumptions:

(a) Assumption 2: $lim_{n_2,n_1\to\infty}\frac{n_2}{n_1} = \ell_{yx}$, a fixed constant.

(b) Assumption 3: $lim_{n_2\to\infty} n_2^{-1}\sum_{i=1}^{n_2} x_i = \theta_x$ which is a fixed p-vector.

From Asumptions (a) and (b), we see that $lim_{n_1\to\infty} n_1^{-1}\sum_{i=1}^{n_1} x_i = \ell_{yx}\theta_x$ Let us further denote f_ϵ and f_δ as probability densitity functions, respectively, for F_ϵ and F_δ. For the rest of this section, we assume that condition (2.4) and Assumptions (a)-(d) are true where (c)-(d) are listed in Appendix.

Theorem 3.1. (a) The outlier proportion $\hat{\lambda}_{b,out}$ has the following representation

$$\begin{aligned} n_2^{1/2}(\hat{\lambda}_{b,out} - \lambda_{b,out}) = & - f_\delta(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0})\ell_{yx}^{1/2} f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma))\theta_x' Q_x^{-1} \\ & n_1^{-1/2} \sum_{i=1}^{n_1} x_i(\gamma - I(\epsilon_i \le F_\epsilon^{-1}(\gamma))) \\ & + n_2^{-1/2} \sum_{i=1}^{n_2} [I(\delta_i \ge F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0}) - \lambda_{b,out}] + o_p(1). \end{aligned}$$

(b) $n_2^{1/2}(\hat{\lambda}_{b,out} - \lambda_{b,out})$ converges in distribution to a normal random variable with distribution $N(0, v_{b,out})$ where

$$\begin{aligned} v_{b,out} = & \gamma(1-\gamma)\ell_{yx}[f_\delta(F_\epsilon^{-1}(\gamma) + \beta_{a0} - \beta_{b0}) f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma))]^2 \\ & \theta_x' Q_x^{-1}\theta_x + \lambda_{b,out}(1 - \lambda_{b,out}). \end{aligned} \tag{3.5}$$

By letting $X \sim N(0,1)$ and $Y \sim N(m,1)$, we compute $v_{b,out}$ for comparison.

Table 3. Variance $v_{b,out}$ comparison

F_Y	$m=0$	$m=1$	$m=3$	$m=5$
$N(m,1)$				
$\gamma = 0.7$	0.42	0.437	0.007	$3.8E-6$
$\gamma = 0.8$	0.32	0.562	0.018	$1.6E-5$
$\gamma = 0.9$	0.18	0.667	0.065	0.000
$\lambda = 0.1$				
$\gamma = 0.7$	0.42	0.434	0.405	0.403
$\gamma = 0.8$	0.32	0.353	0.334	0.331
$\gamma = 0.9$	0.18	0.224	0.232	0.226
$\lambda = 0.2$				
$\gamma = 0.7$	0.420	0.447	0.384	0.381
$\gamma = 0.8$	0.320	0.385	0.339	0.333
$\gamma = 0.9$	0.180	0.271	0.271	0.259
$\lambda = 0.3$				
$\gamma = 0.7$	0.420	0.456	0.358	0.353
$\gamma = 0.8$	0.320	0.415	0.334	0.325
$\gamma = 0.9$	0.180	0.317	0.296	0.277

Larger percentage γ gives the outlier proportion estimator $\hat{\lambda}_{b,out}$ the smaller asymptotic variance. Hence, larger percentage γ's may also give $\hat{\lambda}_{b,out}$ better power performance. However, this requires further investigation.

The above asymptotic distribution allows us to consider an outlier proportion based asymptotic pivotal quantity as

$$\sqrt{n_2}(\frac{\hat{\lambda}_{b,out} - (1-\gamma)}{\sqrt{\hat{v}_{out}}}$$

where $\hat{v}_{out}$ is estimator of $v_{b,out}$. However, it is unpleasant for this quantity being involved with densities f_ϵ and f_δ so that their estimations when they are unknown could be very in-efficient for not enough sample sizes. Hence, when (2.1) is true with

$$v_{a,out} = \gamma(1-\gamma)\ell_{yx}\theta_x' Q_x^{-1}\theta_x + \lambda_{b,out}(1-\lambda_{b,out}).$$

Let $\hat{v}_{out}$ be estimate of v_{out}. We then define an outlier proportion based test for hypothesis (2.1) as

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}(\frac{\hat{\lambda}_{b,out} - (1-\gamma)}{\sqrt{\hat{v}_{a,out}}}) \geq z_\alpha \tag{3.6}$$

where z_α is the $(1-\alpha)$th quantile of the standard normal distribution. This is an extension of the classical proportion p test to this outlier gene problem.

4. Power Performance for a Test based on Outlier Proportion

We consider the following design:

$$\begin{aligned} y_{ai} =& 1 + 2x_i + \epsilon_i, i = 1, ..., n_1 \text{ where } \epsilon_i's \text{ are iid } F_\epsilon, \text{ and} \\ y_{bi} =& \beta_0 + hx_i + \delta_i, i = 1, ..., n_2 \text{ where } \delta_i's \text{ are iid } \sim F_\delta. \end{aligned} \tag{4.1}$$

for evaluation of the test of (3.6). Additional to the outlier proportion $\hat{\lambda}_{b,out}$ and regression quantile $\hat{\beta}_a(\gamma)$, some estimates are defined as follows:

$$\begin{aligned} \hat{\theta}_x &= \frac{1}{n_2}\sum_{i=1}^{n_2} x_{bi}, \ \hat{\ell}_{xy} = \frac{n_2}{n_1} \\ \hat{Q}_x &= \frac{1}{n_2}\sum_{i=1}^{n_2} x_{bi}x_{bi}' \\ \hat{v}_{a,out} &= \gamma(1-\gamma)\hat{\ell}_{xy}\hat{\theta}_x'\hat{Q}_x^{-1}\hat{\theta}_x + \hat{\lambda}_{b,out}(1-\hat{\lambda}_{b,out}). \end{aligned}$$

With $\alpha = 0.05$, we evaluate the following approximate power

$$\pi = \frac{1}{m}\sum_{j=1}^{m} I(n_2^{1/2}(\frac{\hat{\lambda}_{b,out} - (1-\gamma)}{\sqrt{\hat{v}_{a,out}}}) \geq 1.645). \tag{4.2}$$

By letting $h = 2$ and $F = F_\epsilon = F_\delta$, the first aim is to measure the type I error probabilities of this test. We consider several distributions F and the simulated sizes under this setting are displayed in Table 4.

Table 4. Evaluation of probability of type I error

F	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
$N(0,1)$	0.049	0.049	0.045	0.041	0.057
$t(3)$	0.051	0.052	0.046	0.05	0.063
$t(5)$	0.047	0.05	0.047	0.05	0.057
$t(10)$	0.048	0.056	0.044	0.051	0.053
$Lapalce(0,1)$	0.048	0.049	0.046	0.057	0.049
$Lapalce(1,1)$	0.049	0.052	0.054	0.058	0.057

It is seen that the outlier proportion based test is quite robust with simulated sizes all closer to the specified significance level (0.05).

Next we evaluate the power performance for the outlier proportion based test. With the same design of experimment besides h and consider the distribution $F_\epsilon = N(0,1)$ and some F_δ's, the simulated powers are displayed in Table 5.

We consider mixed distribution for variable y_b with $F_\epsilon = N(0,1)$ and $F_\delta = 0.9N(0,1) + 0.1N(\theta,1)$. The simulated powers are displayed in Table 6.

Table 5. Power performance of outlier proportion test under mixed normal distribution

γ	$h = 2.1$	$h = 2.3$	$h = 2.5$
$F_\delta = N(\theta, 1)$			
$\theta = 1$			
$\gamma = 0.5$	1	1	1
$\gamma = 0.7$	1	1	1
$\gamma = 0.8$	1	1	1
$\gamma = 0.9$	0.999	1	1
$\theta = 2$			
$\gamma = 0.5$	1	1	1
$\gamma = 0.7$	1	1	1
$\gamma = 0.8$	1	1	1
$\gamma = 0.9$	1	1	1
$F_\delta = \chi^2(3) - 2.5$			
$\gamma = 0.7$	0.542	0.933	0.998
$\gamma = 0.8$	0.875	0.993	1
$\gamma = 0.9$	0.995	1	1
$F_\delta = t(10) + 0.5$			
$\gamma = 0.7$	0.993	1	1
$\gamma = 0.8$	0.980	1	1
$\gamma = 0.9$	0.955	1	1

Table 6. Power performance of outlier proportion test

γ	$h = 2.1$	$h = 2.3$	$h = 2.5$
$\theta = 1$			
$\gamma = 0.5$	0.552	0.991	1
$\gamma = 0.7$	0.455	0.987	1
$\gamma = 0.8$	0.368	0.964	1
$\gamma = 0.9$	0.283	0.896	0.999
$\theta = 2$			
$\gamma = 0.5$	0.785	0.999	1
$\gamma = 0.7$	0.672	0.998	1
$\gamma = 0.8$	0.564	0.992	1
$\gamma = 0.9$	0.423	0.966	1

The powers displayed in these two tables show that the outlier proportion is quite satisfactory for gene expression analysis.

Lai, Chen and Chen (2012, unpublished) proposed a least squares estimator as

$$\hat{\beta}_{b,ls} = (X_b' A_b X_b)^{-1} X_b' A_b y_b. \tag{4.3}$$

where we denote $X_b = (x_{b1}, x_{b2}, ..., x_{bn_2})'$, trimming matrix $A_b = \text{diag}\{a_{ii} =$

$I(y_{bi} \geq x'_{bi}\hat{\beta}_a(\gamma)), i = 1, ..., n_2\}$ and $y_b = (y_{b1}, ..., y_{bn_2})'$. Two tests (denoted by OL1 and OL2) based on this estimator are also introduced. With two outlier LSE based tests available, it is desired to verify if these two tests are competitive. With the same design of experiment besides various h and the error distributional settings: $F_\epsilon = N(0, 1)$ and several distributions F_δ. The simulated powers are displayed in Table 7.

Table 7. Power comparison of two outlier LSE based tests

γ	$h = 2.1$	$h = 2.3$	$h = 2.5$
$F_\delta = N(\theta, 1), \theta = 1$			
$\gamma = 0.8, OL1$	0.213	0.31	0.397
$OL2$	0.634	0.799	0.871
OP	1	1	1
$\gamma = 0.85, OL1$	0.157	0.186	0.258
$OL2$	0.573	0.706	0.819
OP	1	1	1
$\theta = 2$			
$\gamma = 0.8, OL1$	0.681	0.755	0.828
$OL2$	0.985	0.997	0.998
OP	1	1	1
$\gamma = 0.85, OL1$	0.444	0.502	0.589
$OL2$	0.97	0.983	0.989
OP	1	1	1
$F_\delta = \chi^2(3) - 2.5$			
$\gamma = 0.8, OL1$	0.899	0.932	0.928
$OL2$	0.960	0.982	0.984
OP	0.875	0.993	1
$\gamma = 0.85, OL1$	0.823	0.821	0.845
$OL2$	0.957	0.967	0.979
OP	0.962	1	1
$F_\delta = t(10) + 0.5$			
$\gamma = 0.8, OL1$	0.899	0.932	0.928
$OL2$	0.960	0.982	0.984
OP	0.98	1	1
$\gamma = 0.85, OL1$	0.823	0.821	0.845
$OL2$	0.957	0.967	0.979
OP	0.97	1	1

5. Appendix

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

(c). 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.

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

Proof of Theorem 3.1.

From the expression of outlier proportion of (3.3) and linear regression model of (3.2), we have

$$\begin{aligned}\hat{\lambda}_{b,out} =& n_2^{-1}\sum_{i=1}^{n_2}[I(\delta_i \geq F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0}+n_1^{-1/2}x_i'T_a) - I(\delta_i \geq F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0})] \\ &+ n_2^{-1}\sum_{i=1}^{n_2} I(\delta_i \geq F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0}) \qquad (5.1)\end{aligned}$$

where $T_a = n_1^{1/2}(\hat{\beta}_a(\gamma)-\beta_a(\gamma))$.

With (2.5), the first term on the right hand side of (5.1) can be shown (Ruppert and Carroll (1980) and Chen and Chiang (1996) as

$$\begin{aligned}& n_2^{-1/2}\sum_{i=1}^{n_2}(I(\delta_i \geq F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0}+n_1^{-1/2}x_i'T_n) - I(\delta_i \geq F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0})) \\ &\quad = -\ell_{yx}^{1/2} f_\delta(F_\epsilon^{-1}(\gamma)+\beta_{a0}-\beta_{b0})\theta_x' T_n + o_p(1) \qquad (5.2)\end{aligned}$$

for any sequence $T_n = O_p(1)$. Also, a representation of regression quantile $\hat{\beta}_a(\gamma)$ may be formulated as

$$\begin{aligned}&\sqrt{n_1}(\hat{\beta}_a(\gamma)-\beta_a(\gamma)) \\ &\quad = f_\epsilon^{-1}(F_\epsilon^{-1}(\gamma))Q_x^{-1}n_1^{-1/2}\sum_{i=1}^{n_1} x_i[\gamma - I(\epsilon_i \leq F_\epsilon^{-1}(\gamma))] + o_p(1), \qquad (5.3)\end{aligned}$$

see, for example, Ruppert and Carroll (1980). By letting $T_n = T_a$ and combining the results in (5.1)-(5.3), the theorem is followed. □

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., Chen, D.-T. and Chan, W.. (2010). 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.

Cheng, S, W., and Thaga, K. (2006). On single variable control charts: an overview. *Quality and Reliability Engineering International*, **22**, 811-820.

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

Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics* **19**, 2072-2078.

Jin, R, Si L, Srivastava S, Li Z, Chan, C. A knowledge driven regression model for gene expression and microarray analysis. *Conference Proceedings - IEEE Engineering in Medicine and Biology Society* **1**, 5326-9.

Koenker, R. and Bassett, G.J. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

Luan, H. L. Y. (2003). Kernel Cox regression models for liking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, **8**, 65-76.

Muller, H.-G., Chiou, J.-M. and Leng, X. (2008). Inferring gene expression dynamics via functional regression analysis. *BMC Bioinformatics*, **9**,

60.

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.

Rambow, F., Piton, G., Bouet, S., Leplat, J.-J., Baulande, S., Marrau, A., Stam, M., Horak, V., Vincent-Naulleau, S. (2008). Gene expression signature for spontaneous cancer regression in Melanoma pigs. *Neoplasia* **10**, 714-726.

Rong, J., Luo, S., Srivastava, S., Zheng, L. and Chan, C. (2006). A knowledge driven regression model for gene expression and microarray analysis. *EMBS 06, 28th Annual International Conference of the IEEE*, 5326-5329.

Ruppert, D. and 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. Vinciotti, V. and Yu, K. (2009). M-quantile regression analysis of temporal gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 8 (1) : 41

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