國立交通大學

統計學研究所

碩士論文

在基因晶片中關鍵基因之選取方法

Gene Selection Methods

研究生：彭郃嵐

指導教授：洪慧念　博士

中華民國九十七年六月

在基因晶片中關鍵基因之選取方法

Gene Selection Methods

研究生：彭郃嵐　　Student：Ho-Lan Peng

指導教授：洪慧念　　Advisor：Hui-Nien Hung

國立交通大學

統計學研究所

碩士論文

A Thesis

Submitted to Institute of Statistics

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2008

Hsinchu, Taiwan, Republic of China

中華民國九十七年六月

在基因晶片中關鍵基因之選取方法

研究生：彭郤嵐　　　　　　　　　指導教授：洪慧念博士

國立交通大學統計學研究所

摘要

在分子生物學的領域上，利用統計方法分析基因晶片的資料已成為一種趨勢。若能因此發掘出造成疾病的關鍵基因，對人類會有重要的貢獻。本篇文章中，基於致病基因會在生病的群體中有異常的表現,我們提供一些統計方法能在眾多基因中找出可能致病的關鍵基因。這些方法包含了 WORT、WOS、PGM、TGM、QGM，以及 BRP。我們也將這些方法與過去曾經被發表的 T-statistic、OS、ORT 以及 COPA 等四個方法做比較。

關鍵字：基因選取、OS、ORT、COPA。

Gene Selection Methods

Student: Ho-Lan Peng Advisor: Dr. Hui-Nien Hung

Institute of Statistics

National Chiao Tung University

Abstract

It's a trend to use statistical methods in medical science. If the genes which cause the diseases could be found, it might be helpful to nowadays medical field. In this article, we proposed several methods to find the probable influential genes which are over- or down-expressed in some but not all samples in a disease group. Those methods include WORT (weight outlier robust t-statistic), WOS (weight outlier sum), PGM (the MLE of probability of Gaussian mixture model), TGM (T-statistic of Gaussian mixture model), QGM(Quantile of Gaussian mixture model), and Bayesian Rule P-value(BRP). Also we will compare those methods with four methods (T-statistic, OS, ORT, COPA) which have been proposed and published for detecting differentially expressed genes. Those new methods include improvements of ORT and OS methods, four methods related to Gaussian mixture model and Bayesian method.

Key words and phrases: *gene selection, OS, ORT, COPA.*

誌謝

首先，我要感謝洪慧念老師，在我推甄後、進入交大前給我的一些想法，讓我至今都認為進入交大是一個不悔的選擇，以及在學期間，亦師亦友的給予我鼓勵以及建議，幫助我順利地完成論文及學位，更在生活無助時有個依靠，肯定我對夢想的執著。謝謝陳鄰安老師、洪志真老師以及郭姊在我休學期間以及復學後的照顧，讓我把這裡當成家。謝謝班上同學們這一年半的陪伴，一年級時像家人般時常相聚的時光雖然在二年級較忙碌後不復見，但那片溫情會永遠在心裡。竹北夜市、阿里山低溫之旅、石門水庫之旅、台南行、泰國畢旅、南寮海邊大叫、九份雨中之旅還有大家的生日，所有的一切，都回是未來的愉悅回憶。謝謝可愛的室友兼好友，無論發生什麼事，我們總是可以一起度過，互相扶持。謝謝一起修課的學妹們，和我像同學般的相處，讓我在交大的第二學期可以過得不孤單，感覺有好多好多人和我一起往前，還有我兩個可愛的學妹，不管是共事、打球或生活上的相處，和你們在一起的時光一直充滿愉悅。我想要將這一切，獻給天上的媽媽，完成這一步，我會更加油地繼續往前。

彭郤嵐　　謹誌于

國立交通大學統計學研究所

中華民國九十七年六月

Abstract(in Chinese) . . . . **i**

Abstract(in English) . . . . **ii**

Acknowledgements(in Chinese) . . . . **iii**

Contents

1 Introduction **1**

2 Statistical Methods **2**

2.1 methods review 2

2.1.1 T-statistic 2

2.1.2 The Outlier Sum 2

2.1.3 The Outlier Robust T-statistic 3

2.1.4 Cancer Outlier Profile Analysis 4

2.2 New methods 5

2.2.1 The Weighted OS 5

2.2.2 The Weighted ORT 5

2.2.3 Methods related Gaussian mixture model 6

2.2.4 Bayesian Rule P-value 7

3 Simulation Study **9**

3.1 Comparison by mean, median, Q1, and Q3 9

3.2 Comparison by the true/false-positive rates plots 10

4 Real Data **20**

5 Appendix **21**

5.1 T-statistics 21

5.2 COPA 24

List of Tables

1 Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 1$ 11

2 Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 5$ 12

3 Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 10$ 13

4 Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 15$ 14

5 Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 20$ 15

6 Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 25$ 16

7 The number of the same choices for the ten methods 20

List of Figures

1 The true/false-positive rates plot as $\mu = 1$. 17

2 The true/false-positive rates plot as $\mu = 2$. 18

3 The true/false-positive rates plot as $\mu = 3$. 19

1 Introduction

In the past ten years, scientists discover the differentially expressed genes in human beings using statistical methods instead of traditional medical approach. They try to find some alternative schemes when there are more variables than observations. That is, the number of genes in the human is much larger than the number of sample size we observed. Then use those methods to classify people who would develop the disease. However, in the large sample theory, if we use all the genes to classify people, both the Fisher Rules and Independent Rules misclassify people with probability near 0.5, (Fan, J. and Fan, Y. (2007)) which means that the rules are no better than random guessing. So, these methods are not good for detecting disease. Therefore, when the number of sample size is small comparing to the number of genes, we can only use parts of genes to detect disease. We try to detect differentially expressed genes in stead of taking all genes in the experiment. In this thesis, we consider ten statistical methods to approach our target. For each method, we calculate an index for each gene. The index measures the significant difference between the disease group and the normal group. If the difference between the two group is significant, then the p-value of this index for disease gene would be very small. Repeat the experiment sufficient many times then we obtain the empirical distribution for the p-value of the disease gene. Therefore, by observing the plot of empirical distributions of p-values of those indexes, we can determine which method is better in different situations. Several indexes for detecting differential gene expression had been proposed, such as the traditional analytical method ”t-statistic”, ”cancer profile outlier analysis(COPA)” introduced by Tomlins and others(2005), ”the outlier sum(OS)” introduced by Tibshirani and Hastie(2006), and ”outlier robust t-statistic(ORT)” advanced by Wu(2007). The OS and COPA both use scale estimates and robust location of the gene expression values. The OS and ORT are similarly defined except using different baseline groups. Above four methods will be described in detail in the following section. In the thesis, we try to improve OS and ORT methods by assigning the data with different weights. It works better than the primitive one since it can avoid the abrupt augmentation of the p-value as just one data being added in. In addition, we use some indexes relative to Gaussian mixture model and Bayesian Rule to detect the differentially gene expression. Finally, we compare these methods from different point of view.

2 Statistical Methods

In our work, we consider a two-classes microarray data with p genes. For each gene, let x_i be the expression values for samples $i = 1, \ldots, n$, and we separate the samples fall into two groups, the normal group and the disease group. We assume n_1 samples in the normal group (the first group) and n_2 samples in the disease group (the second group) where $n_1+n_2 = n$. That is $x_1, \ldots, x_{n_1}$ come from normal group, and $x_{n_1+1}, \ldots, x_n$ come from disease group.

There are four methods for detecting differential genes are reviewed in section 2.1, and in section 2.2, we propose six new methods.

2.1 methods review

2.1.1 T-statistic

The two sample t statistic for the gene j is defined as

$$T = \frac{\overline{x}_2 - \overline{x}_1}{S}.$$

where $\overline{x}_j$ is the mean in group j for the gene, $j = 1, 2$, and S is the pooled within-group standard deviation for the gene, i.e.

$$\overline{x}_1 = \frac{\sum_{i=1}^{n_1} x_i}{n_1}, \ \overline{x}_2 = \frac{\sum_{i=n_1+1}^{n} x_i}{n_2}, \ S^2 = \frac{\sum_{i=1}^{n_1} (x_i - \overline{x}_1)^2 + \sum_{i=n_1+1}^{n} (x_i - \overline{x}_2)^2}{n-2}.$$

And we select genes with high value of t statistic. This t statistic is based on the assumptions of normal distributed of genes and the disease samples are over expressed in the important genes.

2.1.2 The Outlier Sum

The method is proposed by Robert Tibshirani and Trevor Hastie(2007). Before finding the outlier-sum statistic, they standardize each gene

$$x_i' = \frac{x_i - med}{mad}.$$

where *med* and *mad* are the median and the median absolute deviation of the expression values, i.e.

$$med = median\left(x_1, \cdots, x_n\right), mad = median\left(\left|x_i - med\right|\right).$$

The outlier-sum (OS) statistic is defined as

$$W = \sum_{i \in group2} x'_i I\left(x'_i > Q3' + IQR'\right) = \sum_{i=n_1+1}^{n} x'_i I\left(x'_i > Q3' + IQR'\right)$$

where $Q1'$ and $Q3'$ are the values of 25% quantile and 75% quantile for the standardized samples(i.e. $x'_1, ..., x'_n$) and $IQR' = Q3' - Q1'$ is the interquartile range.
We compare W for each gene. When W is large, it means there are many outliers in the second group and therefore this gene may cause disease.

2.1.3 The Outlier Robust T-statistic

The method is proposed by Baolin Wu (2007). Before finding the outlier robust t-statistic, they standardize each gene

$$x''_i = \frac{x_i - med_1}{mad}$$

where med_1 is the sample median for the normal group, i.e.

$$med_1 = median\left(x_1, \cdots, x_{n_1}\right).$$

and mad is an estimate for the median absolute deviation.

$$mad = median\left[|x_i - med_1|_{i \le n_1}, |x_i - med_2|_{i > n_1}\right].$$

where med_2 is the sample median for the disease group, i.e.

$$med_2 = median\left(x_{n_1+1}, \cdots, x_n\right).$$

The outlier robust t-statistic (ORT) is defined as

$$U = \sum_{i \in group2} x''_i I\left(x''_i > Q3'' + IQR''\right).$$

where $Q1''$ and $Q3''$ are the values of 25% quantile and 75% quantile for the standardized samples in group 1, and $IQR'' = Q3'' - Q1''$ is the interquartile range. We compare U for each gene.
The difference between OS and ORT are their chosen measured points which base on all samples or samples in normal group.

2.1.4 Cancer Outlier Profile Analysis

The method is proposed by Tomlins and others (2005). Before finding the COPA statistic, they standardize each gene

$$x'_i = \frac{x_i - med}{mad}.$$

where med and mad are the median and median absolute deviation of the expression values, i.e.

$$med = median\left(x_1, \cdots, x_n\right), mad = median\left(|x_i - med|\right).$$

The COPA statistic is defined as $Q = r\%$ quantile of standardized disease group where r could set $75, 90,$ or 95. Then compare Q for each gene.

2.2 New methods

2.2.1 The Weighted OS

Similar to the OS method, we standardize each gene. We change the original method (OS) from computing $\sum_{group2} x_i' I\left(x_i' > Q3' + IQR'\right)$ to $\sum_{group2} x_i' w_i$, where w_i is a weight function. In OS, w_i take values either 0 or 1, i.e.

$$w_i = \begin{cases} 0 & \text{if } x_i' < Q3' + IQR' \\ 1 & \text{if } x_i' \geqslant Q3' + IQR' \end{cases}$$

Therefore, it's not a robust statistics. In our method, we will choose w_i as a continuous function as follows.

$$w_i = \begin{cases} 0 & \text{if } x_i' < Q3' + \frac{1}{2}IQR' \\ \frac{x_i' - (Q3' + \frac{1}{2}IQR')}{IQR'} & \text{if } Q3' + \frac{1}{2}IQR' \leq x_i' \leq Q3' + \frac{3}{2}IQR' \\ 1 & \text{if } x_i' > Q3' + \frac{3}{2}IQR' \end{cases}$$

where $Q1'$ and $Q3'$ are the values of 25% quantile and 75% quantile for for the standardized samples and $IQR' = Q3' - Q1'$ is the interquartile range.
The weighted outlier-sum statistic (WOS) is defined as

$$W^* = \sum_{i \in group2} x_i' w_i$$

We compare W^* for each gene.

2.2.2 The Weighted ORT

The first step is to standardize each gene as ORT method, and to choose weight as

$$w_i = \begin{cases} 0 & \text{if } x_i'' < Q3'' + \frac{1}{2}IQR'' \\ \frac{x_i'' - (Q3'' + \frac{1}{2}IQR'')}{IQR''} & \text{if } Q3'' + \frac{1}{2}IQR'' \leq x_i'' \leq Q3'' + \frac{3}{2}IQR'' \\ 1 & \text{if } x_i'' > Q3'' + \frac{3}{2}IQR'' \end{cases}$$

where $Q1''$ and $Q3''$ are the values of 25% quantile and 75% quantile for the samples in group 1, and $IQR'' = Q3'' - Q1''$ is the interquartile range.
The weighted outlier robust t-statistic is defined as

$$U^* = \sum_{i \in group2} x_i'' w_i$$

where x_i'' is the data after standardization.
We compare U^* for each gene.

2.2.3 Methods related Gaussian mixture model

In disease group, the gene expression of some patients is no difference with the normal group. Under the normal assumption in the normal group and mixed normal assumption on the disease group. We use the EM algorithm to find the MLE of the parameters. Following three methods are related this MLE. Let

$$X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma^2) \text{ and } Y_1 = X_{n_1+1}, \ldots, Y_{n_2} = X_n \sim pN(\mu_1, \sigma^2) + qN(\mu_2, \sigma^2)$$

Let $\hat{\mu_1}, \hat{\mu_2}, \hat{\sigma^2}, \hat{p}, \hat{q}$ denote the MLE of $\mu_1, \mu_2, \sigma^2, p, q$ obtain by EM algorithm.
Let q_i be the probability that Y_i comes from group $N(\mu_2, \sigma^2)$ when we observe Y_i, that is $q_i = P(Y_i \in N(\mu_2, \sigma^2)|Y_i)$, then $q_i = \frac{qf_2(y_i)}{pf_1(y_i)+qf_2(y_i)}$, where f_1 and f_2 are the p.d.f. of $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively, and we can estimate q_i by $\hat{q}_i = \frac{\hat{q}\hat{f}_2(y_i)}{\hat{p}\hat{f}_1(y_i)+\hat{q}\hat{f}_2(y_i)}$

PGM method(the MLE of probability of Gaussian mixture model):

Let index for each gene be $\hat{q} = \frac{\sum_{i=1}^{n_2} \hat{q}_i}{n_2}$, where $\hat{q}$ is the MLE for q.

TGM method(T-statistic of Gaussian mixture model):

The index is defined as $\frac{\widehat{\mu}_2 - \widehat{\mu}_1}{\widehat{\sigma}}$, where

$$\widehat{\mu}_1 = \frac{\sum_{i=1}^{n_2} \widehat{p}_i y_i}{\sum_{i=1}^{n_2} \widehat{p}_i + n_1}, \widehat{\mu}_2 = \frac{\sum_{i=1}^{n_2} \widehat{q}_i y_i}{\sum_{i=1}^{n_2} \widehat{q}_i}$$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{i=1}^{n_2}(\widehat{p}_i(y_i - \mu_1)^2 + \widehat{q}_i(y_i - \mu_2)^2)}{n}$$

This index similar to the t-statistic. In the t-statistic, we only assume that the second group is normally distributed, and this index is an extension of the t-statistic by assuming the second group is a mixture model.

QGM method(Quantile of Gaussian mixture model):

Let $Y_{(1)}, \ldots, Y_{(n_2)}$ be the order statistic of $Y_1, \ldots, Y_{n_2}$, and $q_{(1)}, \ldots, q_{(n_2)}$ be the corresponding probability that $Y_{(i)}$ comes from group $N(\mu_2, \sigma^2)$.
Define the r-percent quantile of $Y_{(1)}, \ldots, Y_{(n_2)}$ in group $N(\mu_2, \sigma^2)$ by $y_{(l)}$ such that $\frac{\sum_{i=1}^{l} q_{(i)}}{\sum_{i=1}^{n_2} q_{(i)}} \geq r$ and $\frac{\sum_{i=l+1}^{n_2} q_{(i)}}{\sum_{i=1}^{n_2} q_{(i)}} \geq 1 - r$, for $r = 0.75, 0.90, 0.95 \ldots$.

By the way, we get a theorem according to QGM.

Theorem.
The $y_{(l)}$ in the QGM converges to the r-percent quantile of the group $N(\mu_2, \sigma^2)$.

Proof.
The data in disease group comes from the distribution of $pf_1 + qf_2$ where f_1 is the distribution of $N(\mu_1, \sigma^2)$ and f_2 is the distribution of $N(\mu_2, \sigma^2)$.
Given data y,

$$q_i = \frac{\widehat{q}f_2(y)}{\widehat{p}f_1(y) + \widehat{q}f_2(y)} \rightarrow \frac{qf_2(y)}{pf_1(y) + qf_2(y)} \text{ as } n_2 \rightarrow \infty$$

Let $Y_{(1)}, \ldots, Y_{(n_2)}$ be the order statistic, and $q_{(1)}, \ldots, q_{(n_2)}$ be the corresponding probability, where $q_{(i)} = \frac{qf_2(y_{(i)})}{pf_1(y_{(i)})+qf_2(y_{(i)})}$
We have

$$\begin{aligned}\frac{\sum_{i=1}^{n_2} q_i}{n_2} \rightarrow E(q_i) &= \int_{-\infty}^{\infty} \frac{qf_2(y)}{pf_1(y) + qf_2(y)}[pf_1(y) + qf_2(y)]dy \\ &= \int_{-\infty}^{\infty} qf_2(y) = q\end{aligned}$$

That is $\sum_{i=1}^{n_2} q_i \approx n_2 q$
So,

$$\begin{aligned}\frac{\sum_{i=1}^{l} q_i}{n_2} &\approx \frac{rn_2q}{n_2} = rq \\ &\approx \int_{-\infty}^{y_{(l)}} \frac{qf_2(y)}{pf_1(y) + qf_2(y)}[pf_1(y) + qf_2(y)]dy = \int_{-\infty}^{y_{(l)}} qf_2(y)dy\end{aligned}$$

That is $r \approx \int_{-\infty}^{y_{(l)}} f_2(y)dy$. □

2.2.4 Bayesian Rule P-value

There are many statistican using Bayesian Rule to solve their problems in biology, P. Baldi and A.D. Long. (2001) and E. Kristiansson and A. Sjogren (2006). Here, we will try using Bayesian Rule in our problem. Let

$$X_1, \ldots, X_{n_1}|\mu_1, \sigma \sim N(\mu_1, \sigma^2), X_{n_1+1}, \ldots, X_n|\mu_2, \sigma \sim N(\mu_2, \sigma^2).$$

where μ_1 and μ_2 comes from uniform distribution, and σ^2 comes from Inverse Gaussian distribution with the mean one and the shape parameter one.
We know that

$$X_1, \ldots, X_{n_1}|\mu_1, \sigma \sim f_1(x_1, \ldots, x_n|\mu_1, \sigma) = (\frac{1}{\sqrt{2\pi}\sigma})^{n_1} e^{-\frac{\sum_{i=1}^{n}(x_i - \mu_1)^2}{2\sigma^2}}$$

$$X_{n+1}, \ldots, X_n|\mu_2, \sigma \sim f_2(x_{n_1+1}, \ldots, x_n|\mu_2, \sigma) = (\frac{1}{\sqrt{2\pi}\sigma})^{n_2} e^{-\frac{\sum_{i=n_1+1}^{n}(x_i - \mu_2)^2}{2\sigma^2}}$$

$$\sigma^2 \sim f_3(\sigma) = \frac{1}{\sigma^4}e^{-\frac{1}{\sigma^2}}$$

Then

$$\mu_1, \mu_2, \sigma^2 | x_1, \ldots, x_n \sim c f_1(x_1, \ldots, x_{n_1} | \mu_1, \sigma) f_2(x_{n_1+1}, \ldots, x_n | \mu_2, \sigma) f_3(\sigma)$$

where

$$\begin{aligned} c^{-1} &= \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty f_1(x_1, \ldots, x_{n_1} | \mu_1, \sigma) f_2(x_{n_1+1}, \ldots, x_n | \mu_2, \sigma) f_3(\sigma) d\mu_1 d\mu_2 d\sigma^2 \\ &= \frac{2}{\sqrt{n_1 n_2}} (\frac{1}{\sqrt{\pi}})^{n-2} \Gamma(\frac{n}{2}) (\sum_{i=1}^{n_1} (x_i - \overline{x}_1)^2 + \sum_{i=n_1+1}^{n} (x_i - \overline{x}_2)^2 + 2)^{-\frac{n}{2}} \end{aligned}$$

Then, the distribution of $\mu_1 - \mu_2$ is

$$\frac{1}{2} c f_1(x_1, \ldots, x_{n_1} | \frac{u+v}{2}, \sigma) f_2(x_{n_1+1}, \ldots, x_n | \frac{v-u}{2}, \sigma) f_3(\sigma)$$

if we let $u = \mu_1 - \mu_2$ and $v = \mu_1 + \mu_2$.
So, we get

$$\begin{aligned} & f_4(u | x_1, \ldots, x_n) \\ &= \pi^{-\frac{1}{2}} \sqrt{\frac{n_1 n_2}{n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{[\sum_{i=1}^{n_1} (x_i - \overline{x}_1)^2 + \sum_{i=n_1+1}^{n} (x_i - \overline{x}_2)^2 + 2]^{\frac{n}{2}}}{[\frac{(u-(\overline{x}_1 - \overline{x}_2))^2}{\frac{n}{n_1 n_2}} + \sum_{i=1}^{n_1} (x_i - \overline{x}_1)^2 + \sum_{i=n_1+1}^{n} (x_i - \overline{x}_2)^2 + 2]^{\frac{n+1}{2}}} \end{aligned}$$

is a p.d.f. of $u | x_1, \ldots, x_n$.
The index is defined to be $P(u > 0 | x_1, \ldots, x_n)$.

3 Simulation Study

Now, we try to compare above methods. Theoretically, we could derive the distribution functions of the p-value of the indexes for a gene. Let the distribution of the index of normal group and disease group be F_1 and F_2 respectively. Suppose for some gene, the distribution of the index follows the distribution of F_2, $F_2 = F_1$ if the gene comes from the normal group. If we observe the index value v, then $1 - F_1(v)$ is the proportion of the normal genes with the index statistics greater than this index value, that is $1 - F_1(*)$ is the p-value of this gene. Therefore, $1 - F_2(F_1^{-1}(1-*))$[1] is the cumulative distribution function of the p-value for the gene. And we could obtain the mean, median, $Q1$, $Q3$ and the plot of the distribution of this p-value(i.e. the true/false-positive rates plot) and then use these statistics to compare all methods. The distribution of T and Q in the t-statistic method and COPA method could be obtained by analytically, which will be seen in Appendix in detail. Else, we use simulation to find mean, median, standard deviation, $Q1$, $Q3$, and the empirical cumulative distribution function plot. In the simulation study, we let $n_1 = n_2 = 25$ samples in normal and disease group. Set one disease gene which contains $k = 1, 5, 10, 15, 20, 25$ outlier disease samples from the normal distribution with $\mu = 1, 2, 3$ and $\sigma^2 = 1$, and the other $n_2 - k$ genes and the 999 normal genes coming from the standard normal distribution. That is $X_1, \ldots, X_{n_1}, X_{n_1+k+1}, \ldots, X_n \sim N(0,1)$, $X_{n_1+1}, \ldots, X_{n_1+k} \sim N(\mu, 1)$ where $\mu = 1, 2, 3$ and $k = 1, 5, 10, 15, 20, 25$.

3.1 Comparison by mean, median, Q1, and Q3

We use simulation to compare these methods by checking mean, median, Q1, and Q3 of the p-value for the disease gene (See the tables in Appendix. The blue marked numbers are the smallest one of each row, and the light blue numbers are a little bit bigger than the red ones. There are no big differences between them.) We compare all methods by two ways, by fixing k and μ.

First, we fix k to see the behavior of the p-value when μ increases. When k is small, such as $k = 1, 5$, from table 2 and table 3, we can see that no matter what μ is, QGM is the best choice for finding the over-expressed gene. Subsequently, when k is a little bit larger ($k = 10$), from table 4, it's shown that QGM and BRP are good choices. Besides,

[1] We want to find the distribution of $1 - F_1(v)$:

$$F(\zeta) = P(1 - F_1(v) \leq \zeta) = P(1 - \zeta \leq F_{(v)}) = P(F_1^{-1}(1-\zeta) \leq v) = 1 - F_2(F_1^{-1}(1-\zeta))$$

as μ increases, ORT and WORT could be considered to be good indexs. When k is large, from table 5, 6, and 7, T-statistics and BRP are acceptable. And as μ increases, similarly as $k = 10$, ORT and WORT could be taken into consideration. By the way, when k equals to the number of patients in disease group, i.e. $k = 25$, T-statistics, ORT, WORT, PGM, BRP can be considered.

Second, we fix μ to observe the behavior of the p-value when k increases. When μ is small, that is, the difference between normal persons' and patients' genes is small, we can change our choice from QGM to T-statistics and BRP if k is increasing. When μ is bigger ($\mu = 2, 3$), QGM may be good as k is small, but we have more choices such like T-statistics, ORT, WORT, PGM, and BRP if k increases.

3.2 Comparison by the true/false-positive rates plots

We have checked the empirical cumulative distribution function of the p-value for the disease gene to compare all methods. In a word, we could obtain an index for each gene in every method, and we then sort and rank the index for all genes. By finding out the ranking of the testing gene, we could get the p-value for the disease gene and plot its empirical cumulative distribution function after repeating 1000 times. Following are the empirical cumulative distribution function plots for the disease gene, i.e. the true/false-positive rates plots

In Figure 1, when $\mu = 1$ and $k = 1$, no methods perform significant results. As k increases, the performances of t-statistic, PGM, and BRP become better than other methods. T-statistic is based on the assumption that all disease samples are over-expressed. That is, t-statistic would be a good choice as k=25. BRP performs perfect as k is not too small, and OS, COPA, WOS are suitable when k is small. When μ is larger, from Figure 2 and 3, we can see that OS, WOS, and COPA are not good choices. And PGM could be used only when k is large. All methods could be used in different situations, depending on someone's need.

Table 1: Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 1$

	k=1									
	t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
mean	0.503425	0.332726	0.382031	0.475478	0.460767	0.456226	0.515849	0.463279	0.444288	0.461666
	0.488659	0.291153	0.327696	0.437280	0.365272	0.372576	0.552766	0.359449	0.336933	0.425518
	0.498380	0.202792	0.242448	0.393305	0.210031	0.253342	0.592947	0.187595	0.160489	0.400017
median	0.496121	0.438157	0.445672	0.473869	0.444869	0.450985	0.520469	0.446672	0.422981	0.443521
	0.488611	0.294744	0.318148	0.412339	0.302780	0.312552	0.557383	0.282517	0.254169	0.401678
	0.486773	0.144665	0.212457	0.349542	0.132408	0.194902	0.654231	0.075434	0.033146	0.364375
Q1	0.243397	0.202927	0.215220	0.237498	0.213725	0.212589	0.266356	0.193261	0.178892	0.217912
	0.229477	0.140786	0.144221	0.171781	0.128670	0.142797	0.259048	0.093499	0.057813	0.172720
	0.250230	0.062242	0.095908	0.137837	0.053121	0.085610	0.321078	0.006838	0.003157	0.160044
Q3	0.759316	0.452763	0.561711	0.712427	0.703074	0.688448	0.779123	0.701897	0.685438	0.695382
	0.744498	0.462319	0.567777	0.683190	0.579120	0.586721	0.864087	0.611211	0.574247	0.660862
	0.751982	0.288952	0.336103	0.606989	0.284704	0.351789	0.892738	0.265613	0.216195	0.602832

Table 2: Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 5$

	k=5									
	t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
mean	0.447036	0.298661	0.307272	0.394098	0.382450	0.345894	0.502511	0.336772	0.279038	0.323577
	0.303036	0.162421	0.132597	0.180655	0.155544	0.130587	0.462746	0.132812	0.076709	0.169944
	0.188841	0.028930	0.029282	0.036730	0.024285	0.026961	0.228617	0.034379	0.006004	0.095127
median	0.429920	0.365749	0.285679	0.343618	0.332446	0.280452	0.471238	0.262588	0.187220	0.268012
	0.209407	0.081171	0.064656	0.097691	0.071080	0.063439	0.432612	0.079287	0.026803	0.103590
	0.105044	0.003455	0.009717	0.007353	0.003086	0.008975	0.221750	0.018798	0.000545	0.050595
Q1	0.181776	0.140741	0.108292	0.149100	0.135458	0.107519	0.210109	0.096130	0.064136	0.100438
	0.069651	0.019475	0.017439	0.027777	0.017111	0.016985	0.183690	0.033560	0.004359	0.033525
	0.036040	0.000465	0.002717	0.000995	0.000419	0.002576	0.155852	0.006939	0.000030	0.016717
Q3	0.698392	0.448697	0.536373	0.618438	0.606539	0.561833	0.813603	0.529939	0.423733	0.499015
	0.485495	0.243280	0.181018	0.274684	0.215245	0.181837	0.753184	0.170049	0.103403	0.251543
	0.274795	0.019262	0.030732	0.031288	0.015737	0.029217	0.283391	0.044429	0.002646	0.134837

Table 3: Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 10$

	k=10									
	t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
mean	0.277613	0.301088	0.248365	0.328232	0.322601	0.232136	0.401751	0.199308	0.144069	0.163045
	0.078849	0.178209	0.057111	0.134010	0.155642	0.048754	0.190313	0.085287	0.033592	0.038436
	0.019822	0.057664	0.005101	0.028508	0.032571	0.004020	0.059814	0.026887	0.002183	0.009327
median	0.164221	0.287583	0.159554	0.279836	0.260442	0.150776	0.275487	0.126878	0.091423	0.083322
	0.025747	0.090231	0.012798	0.073883	0.068151	0.012646	0.131403	0.063797	0.016843	0.012121
	0.006217	0.003343	0.000788	0.006192	0.002455	0.000808	0.056863	0.016783	0.000263	0.002742
Q1	0.047873	0.111438	0.045464	0.121610	0.103974	0.042717	0.102282	0.054191	0.038995	0.023293
	0.005379	0.016874	0.002687	0.019868	0.014601	0.002636	0.064232	0.032762	0.003207	0.002586
	0.001399	0.000162	0.000182	0.000990	0.000121	0.000162	0.039585	0.007126	0.000010	0.000677
Q3	0.447162	0.536146	0.396834	0.495091	0.502056	0.365532	0.737407	0.273401	0.194352	0.234791
	0.091080	0.315487	0.053661	0.182483	0.234493	0.048803	0.250058	0.112918	0.047050	0.044141
	0.017399	0.036106	0.002444	0.025803	0.023419	0.002662	0.073949	0.032378	0.001470	0.008187

Table 4: Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 15$

	k=15									
	t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
mean	0.143662	0.301874	0.193259	0.291045	0.326352	0.176118	0.279331	0.138400	0.097999	0.076572
	0.009025	0.273737	0.026894	0.175177	0.264060	0.020461	0.047634	0.075346	0.026318	0.004332
	0.000537	0.308976	0.001091	0.084341	0.296479	0.000774	0.012662	0.026649	0.001131	0.000244
median	0.048691	0.299936	0.089176	0.224169	0.260674	0.085686	0.133155	0.094661	0.065378	0.023919
	0.001030	0.239634	0.001460	0.111428	0.173109	0.001333	0.031929	0.060893	0.014768	0.000626
	0.000045	0.422915	0.000061	0.044186	0.237038	0.000051	0.011283	0.018101	0.000263	0.000030
Q1	0.010353	0.092731	0.018899	0.083130	0.085954	0.016793	0.034121	0.044353	0.030424	0.004702
	0.000101	0.051671	0.000121	0.038495	0.040368	0.000152	0.013030	0.035101	0.004808	0.000121
	0.000010	0.100206	0.000010	0.012697	0.055090	0.000010	0.007136	0.008364	0.000030	0.000030
Q3	0.201862	0.532338	0.311860	0.466475	0.537651	0.263310	0.449031	0.180988	0.135817	0.101236
	0.005606	0.534732	0.012449	0.259028	0.446172	0.010581	0.062358	0.099433	0.035949	0.002495
	0.000182	0.494652	0.000197	0.109479	0.480626	0.000182	0.016207	0.035484	0.001010	0.000091

Table 5: Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 20$

	k=20									
	t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
mean	0.049171	0.346446	0.131156	0.285948	0.322618	0.121705	0.152151	0.094703	0.072484	0.024368
	0.000414	0.336430	0.009894	0.228798	0.361276	0.005684	0.009837	0.068837	0.019152	0.000224
	0.000012	0.444797	0.000163	0.247474	0.601847	0.000058	0.001361	0.025075	0.000749	0.000053
median	0.007848	0.331456	0.044186	0.238134	0.281553	0.039065	0.039575	0.074772	0.053227	0.003621
	0.000020	0.415127	0.000091	0.181943	0.325320	0.000121	0.004773	0.060530	0.011540	0.000020
	0.000010	0.482046	0.000010	0.218579	0.638165	0.000010	0.001050	0.017646	0.000141	0.000051
Q1	0.001293	0.105817	0.005495	0.091120	0.083327	0.004510	0.008848	0.038550	0.025621	0.000646
	0.000010	0.184629	0.000020	0.076085	0.130135	0.000020	0.001651	0.035358	0.004348	0.000010
	0.000010	0.473318	0.000010	0.102640	0.420925	0.000010	0.000657	0.008667	0.000010	0.000010
Q3	0.040883	0.623761	0.190145	0.422152	0.512919	0.160716	0.162791	0.124928	0.098272	0.020030
	0.000131	0.490354	0.001212	0.343476	0.556757	0.001091	0.011611	0.093317	0.026545	0.000101
	0.000010	0.486672	0.000010	0.358491	0.816123	0.000010	0.001788	0.034747	0.000854	0.000081

Table 6: Results of simulation study mean, median, standard deviation, Q1,Q3 of p-values for gene 1,over 50 simulations. The three numbers in one area stand for $\mu = 1, 2, 3$ as $k = 25$

	k=25									
	t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
mean	0.011452	0.349903	0.110370	0.293863	0.337343	0.086801	0.055586	0.080016	0.063338	0.005677
	0.000011	0.439489	0.006235	0.343225	0.466619	0.003109	0.000553	0.062147	0.017812	0.000048
	0.000010	0.498207	0.000074	0.480043	0.715808	0.000031	0.000025	0.022028	0.000603	0.000046
median	0.001071	0.348193	0.019954	0.239205	0.293037	0.017646	0.007656	0.067343	0.049904	0.000495
	0.000010	0.573307	0.000010	0.330189	0.469046	0.000010	0.000010	0.056878	0.013091	0.000030
	0.000010	0.490500	0.000010	0.476465	0.769118	0.000010	0.000010	0.017879	0.000242	0.000020
Q1	0.000192	0.108529	0.002096	0.090661	0.100762	0.002071	0.000929	0.040717	0.028813	0.000091
	0.000010	0.308472	0.000010	0.175761	0.239644	0.000010	0.000010	0.038995	0.006040	0.000030
	0.000010	0.452864	0.000010	0.335396	0.597186	0.000010	0.000010	0.008480	0.000040	0.000020
Q3	0.006323	0.603342	0.131357	0.462763	0.531757	0.097711	0.040126	0.104514	0.081120	0.002965
	0.000010	0.573913	0.000111	0.484894	0.686766	0.000091	0.000253	0.080671	0.023848	0.000071
	0.000010	0.556131	0.000010	0.622974	0.873385	0.000010	0.000010	0.029550	0.000838	0.000111

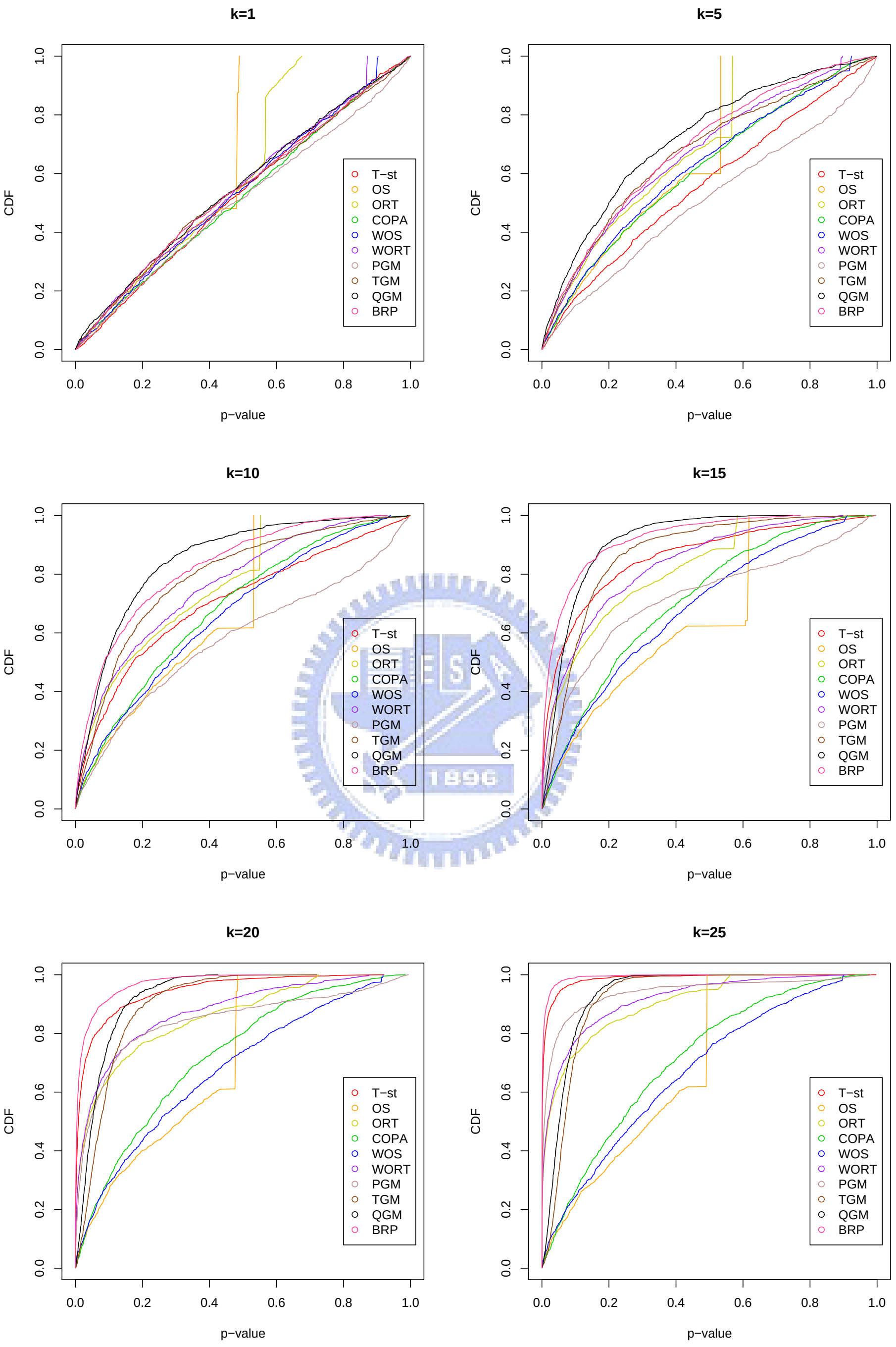

Figure 1: The true/false-positive rates plot as $\mu = 1$.

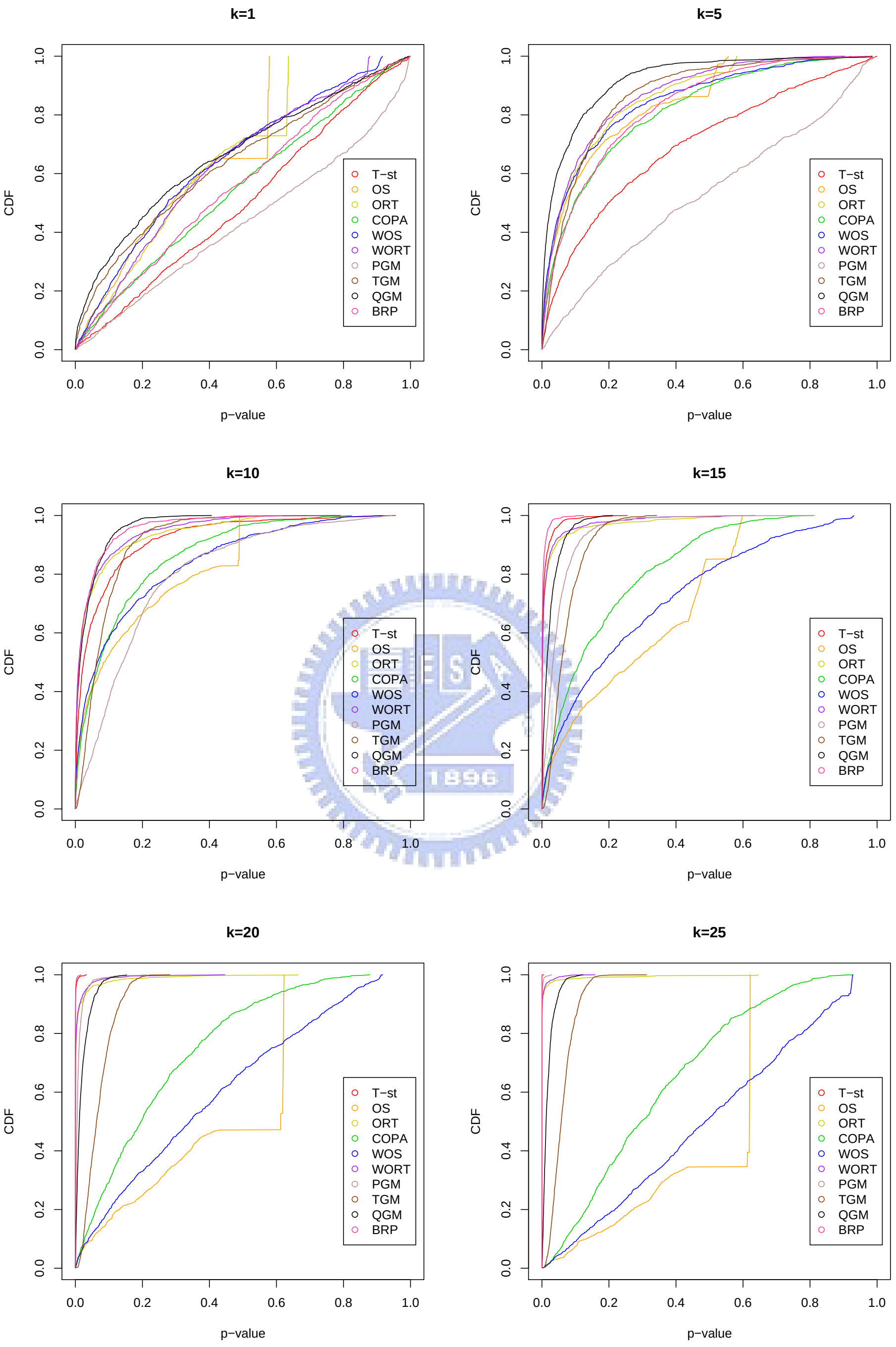

Figure 2: The true/false-positive rates plot as $\mu = 2$.

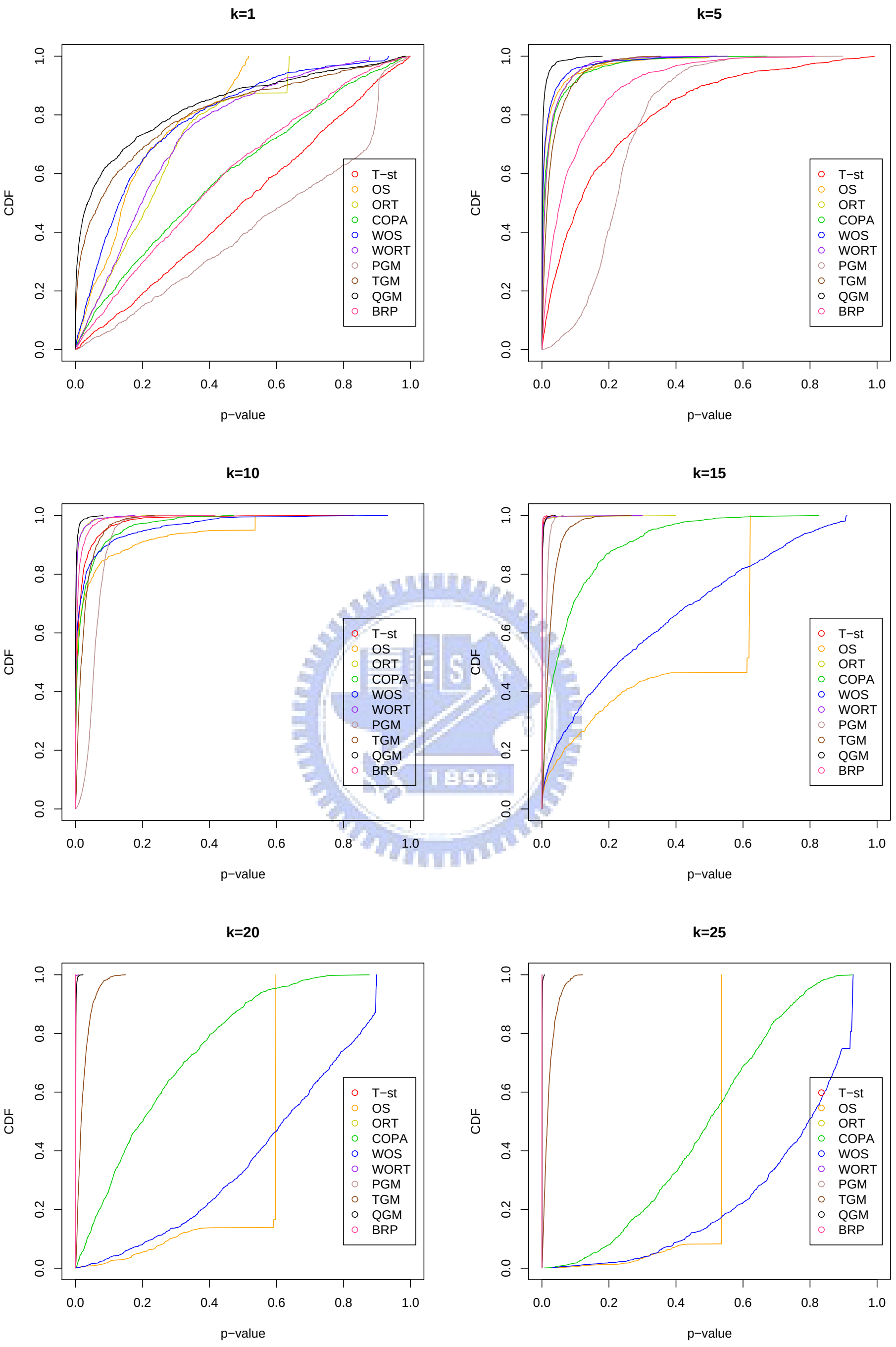

Figure 3: The true/false-positive rates plot as $\mu = 3$.

4 Real Data

The data is for breast cancer in microarray data, which gotten by Department of Interdisciplinary Oncology Moffitt Cancer Center and Research Institute, University of South Florida. There are 54675 genes in the data, 143 healthy persons and 42 patients are included in the normal group and disease group dividedly. They found 1554 genes among all genes. We also choose 1554 significant genes by every methods and check how many of them are included in their choices. The number of the same choices of every methods could be seen in Table 1. By the way, before finding the index, data for each gene would be checked if the median for disease group is larger than the median for normal group. If not, we would change the sign for each data.

Table 7: The number of the same choices for the ten methods

t	OS	ORT	COPA	WOS	WORT	PGM	TGM	QGM	BRP
152	528	724	382	529	715	382	77	152	714

5 Appendix

5.1 T-statistics

First, we consider the gene that is not over-expressed,

$$X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_n \sim N(\mu_1, \sigma^2)$$

So, we get

$$\overline{X}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$$

$$\overline{X}_2 \sim N(\mu_1, \frac{\sigma^2}{n_2})$$

The t-statistic is

$$T = \frac{\overline{X}_2 - \overline{X}_1}{S}$$

Now, we try to find the distribution of $\overline{X}_2 - \overline{X}_1$ and S for the gene.

$$\overline{X}_2 - \overline{X}_1 \sim N(0, (\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}))$$

$$S^2 = \frac{1}{(n-2)}[\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2 + \sum_{i=n_1+1}^{n}(X_i - \overline{X}_2)^2]$$

We know that

$$\frac{\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2}{\sigma^2} \sim \chi^2_{n_1-1} \quad \text{and} \quad \frac{\sum_{i=n_1+1}^{n}(X_i - \overline{X}_2)^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

So,

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

Therefore,

$$T = \frac{\overline{X}_2 - \overline{X}_1}{S} = \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}[\frac{\frac{\overline{X}_2 - \overline{X}_1}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}}}{\sqrt{\frac{(n-2)S^2/\sigma^2}{n-2}}}] \sim \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}T_{n-2}$$

and the c.d.f. of T is

$$G(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})} \frac{1}{\sqrt{(n-2)\pi}} \frac{1}{(1 + (\frac{\frac{x^2}{(\frac{1}{n_1}+\frac{1}{n_2})}}{n-2}))^{\frac{n-1}{2}}} dx$$

Second, in disease group, there are k patients' genes over-expressed. So, we get

$$X_1, \dots, X_{n_1} \sim N(\mu_1, \sigma^2)$$

$$X_{n_1+1}, \dots, X_{n_1+k} \sim N(\mu_2, \sigma^2)$$

$$X_{n_1+k+1}, \dots, X_n \sim N(\mu_1, \sigma^2)$$

Now, we try to find the distribution of $\overline{X}_2 - \overline{X}_1$ and S for the gene.

$$\overline{X}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$$

$$\overline{X}_2 \sim N(\frac{k\mu_2 + (n_2 - k)\mu_1}{n_2}, \frac{\sigma^2}{n_2})$$

So, we get

$$\overline{X}_2 - \overline{X}_1 \sim N(\frac{k(\mu_2 - \mu_1)}{n_2}, (\frac{1}{n_1} + \frac{1}{n_2})\sigma^2)$$

$$S^2 = \frac{1}{(n-2)}[\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2 + \sum_{i=n_1+1}^{n}(X_i - \overline{X}_2)^2]$$

where

$$\frac{\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2}{\sigma} \sim \chi^2_{n_1-1}$$

Now, we want to find the distribution of $\sum_{i=n_1+1}^{n}(X_i - \overline{X}_2)^2$.
We separate $\sum_{i=n_1+1}^{n}(X_i - \overline{X}_2)^2$ into two parts:
$\sum_{i=n_1+1}^{n_1+k}(X_i - \overline{X}_2)^2$ and $\sum_{i=n_1+k+1}^{n}(X_i - \overline{X}_2)^2$
and the distribution of $(X_i - \overline{X}_2)^2$ for $i = n_1 + 1, \dots, n_1 + k$ is found as:

$$\begin{aligned} X_i - \overline{X}_2 &= X_i - \frac{X_i + \sum_{j=n_1+1, j\neq i}^{n_1+k} X_j + \sum_{j=n_1+k+1}^{n_2} X_j}{n_2} \\ &= (1 - \frac{1}{n_2})X_i - \frac{1}{n_2}(\sum_{j=n_1+1, j\neq i}^{n_1+k} X_j + \sum_{j=n_1+k+1}^{n} X_j) \end{aligned}$$

where

$$(1 - \frac{1}{n_2})X_i \sim N((1 - \frac{1}{n_2})\mu_2, (1 - \frac{1}{n_2})^2\sigma^2)$$

$$\frac{1}{n_2}X_j \sim N(\frac{\mu_2}{n_2}, (\frac{1}{n_2})^2\sigma^2) \quad \text{for} \quad j > n_1, j \neq i$$

$$\frac{1}{n_2}X_j \sim N(\frac{\mu_1}{n_2}, (\frac{1}{n_2})^2\sigma^2) \quad \text{for} \quad j = n_1 + k + 1, \dots, n$$

$$\therefore X_i - \overline{X}_2 \sim N((1 - \frac{k}{n_2})(\mu_1 + \mu_2), \frac{(n_2^2 - n_2 + 1)}{n_2^2}\sigma^2) \quad \text{for} \quad i = n_1 + 1, \dots, n_1 + k$$

Take $(1-\frac{k}{n_2})(\mu_1+\mu_2)=\mu^*$ and $\frac{(n_2^2-n_2+1)}{n_2^2}\sigma^2=\sigma^{*2}$.
Then, the p.d.f of $\sum_{i=n_1+1}^{n_1+k}(X_i-\overline{X}_2)^2$ is

$$f_1(x)=\sum_{i=0}^{\infty}\frac{e^{-\frac{\delta^*}{2}}(\frac{\delta^*}{2})^i}{i!}\frac{e^{-\frac{x}{2\sigma^{*2}}}(\frac{x}{\sigma^{*2}})^{\frac{k+2i}{2}-1}}{2^{\frac{k}{2}+i}\sigma^{*2}\Gamma(\frac{k}{2}+i)}$$

where $\delta^*=k(\frac{\mu^*}{\sigma^*})^2$.
Then, we find the distribution of $(X_i-\overline{X}_2)^2$ for $i=n_1+k+1,\ldots,n$:

$$\begin{aligned}X_i-\overline{X}_2&=X_i-\frac{X_i+\sum_{n_1+1}^{n_1+k}X_j+\sum_{j=n_1+k+1,j\neq i}^{n}X_j}{n_2}\\&=(1-\frac{1}{n_2})X_i-\frac{1}{n_2}(\sum_{j=n_1+1}^{n_1+k}X_j+\sum_{j=n_1+k+1,j\neq i}^{n}X_j)\end{aligned}$$

$$(1-\frac{1}{n_2})X_i\sim N((1-\frac{1}{n_2})\mu_1,(1-\frac{1}{n_2})^2\sigma^2)$$
$$\frac{1}{n_2}X_j\sim N(\frac{\mu_2}{n_2},(\frac{1}{n_2})^2\sigma^2)\quad\text{for}\quad j=n_1+1,\ldots,n_1+k$$
$$\frac{1}{n_2}X_j\sim N(\frac{\mu_1}{n_2},(\frac{1}{n_2})^2\sigma^2)\quad\text{for}\quad j=n_1+k+1,\ldots,n,j\neq i$$
$$\therefore X_i-\overline{X}_2\sim N(\frac{k}{n_2}(\mu_1-\mu_2),\frac{(n_2-1)\sigma^2}{n_2})\quad\text{for}\quad i=n_1+k+1,\ldots,n$$

Take $(1-\frac{k}{n_2})(\mu_1+\mu_2)=\mu^{**}$ and $\frac{(n_2^2-n_2+1)}{n_2^2}\sigma^2=\sigma^{**2}$.
Then, the p.d.f of $\sum_{i=n_1+k+1}^{n}(X_i-\overline{X}_2)^2$ is

$$f_2(x)=\sum_{i=0}^{\infty}\frac{e^{-\frac{\delta^{**}}{2}}(\frac{\delta^{**}}{2})^i}{i!}\frac{e^{-\frac{x}{2\sigma^{**2}}}(\frac{x}{\sigma^{**2}})^{\frac{n_2-k+2i}{2}-1}}{2^{\frac{n_2-k}{2}+i}\sigma^{**2}\Gamma(\frac{n_2-k}{2}+i)}$$

where $\delta^{**}=(n_2-k)(\frac{\mu^{**}}{\sigma^{**}})^2$.
Since $\sum_{i=n_1+1}^{n_1+k}(X_i-\overline{X}_2)^2\sim f_1(x)$ and $\sum_{i=n_1+k+1}^{n}(X_i-\overline{X}_2)^2\sim f_2(x)$,

$$\sum_{i=n_1+1}^{n}(X_i-\overline{X}_2)^2\sim f_3(x)$$

where

$$f_3(x)=\int_{-x}^{x}\frac{1}{2}e^{-\frac{\delta^*+\delta^{**}}{2}-\frac{1}{4}(\frac{x+w}{\sigma^{*2}}+\frac{x-w}{\sigma^{**2}})}(\sum_{i=0}^{\infty}\frac{(\frac{\delta^*}{2})^i}{i!}\frac{(\frac{x+w}{2\sigma^{*2}})^{\frac{k+2i}{2}-1}}{2^{\frac{k}{2}+i}\sigma^{*2}\Gamma(\frac{k}{2}+i)})(\sum_{i=0}^{\infty}\frac{(\frac{\delta^{**}}{2})^i}{i!}\frac{(\frac{x-w}{2\sigma^{**2}})^{\frac{n_2-k+2i}{2}-1}}{2^{\frac{n_2-k}{2}+i}\sigma^{**2}\Gamma(\frac{n_2-k}{2}+i)})dw$$

Since $\sum_{i=1}^{n_1}(X_i-\overline{X}_1)^2\sim f_4(x)=\frac{1}{2^{\frac{n_1}{2}}\Gamma(\frac{n}{2})}e^{-\frac{\sigma^2x}{2}}\sigma^{n_1}x^{\frac{n_1}{2}-1}$,
so, the p.d.f of $S=\sqrt{\frac{1}{(n-2)}[\sum_{i=1}^{n_1}(X_i-\overline{X}_1)^2+\sum_{i=n_1+1}^{n}(X_i-\overline{X}_2)^2]}$ is

$$f_5(x)=\int_0^x 4xvf_3(x^2-v^2)f_4(v^2)dv$$

The p.d.f. of $\overline{X}_2 - \overline{X}_1$ is

$$f_6(x) = \frac{1}{\sqrt{2\pi(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}} e^{-\frac{(x-\frac{k(\mu_2-\mu_1)}{n_2})^2}{2(\frac{1}{n_1}+\frac{1}{n_2})\sigma^2}}$$

So, the p.d.f. of $T^* = \frac{\overline{X}_2-\overline{X}_1}{S}$ is

$$f_7(t^*) = \int_{-\infty}^{\infty} w f_6(t^* w) f_5(w) dw$$

The cumulative distribution function of the p-value of the disease gene is $1 - G(t^*)$ where $t^* \sim f_7$

5.2 COPA

For non-over-expressed gene,

$$X_{n_1+1}, \ldots, X_n \sim N(\mu_1, \sigma^2)$$

So, the c.d.f. and p.d.f of X_i for $i = n_1, \ldots, n$ are

$$F\,(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu_1)^2}{2\sigma^2}} dt = \frac{1}{2}\left(1 + erf\frac{x-\mu_1}{\sqrt{2}\sigma}\right)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

The c.d.f. of 90% quantile of $X_{n_1+1}, \ldots, X_n$is

$$G(x) = \int_{0}^{x} \frac{n_2!}{\lfloor 0.9n_2 \rfloor!(n_2 - \lfloor 0.9n_2+1 \rfloor)!} F(t)^{\lfloor 0.9n_2 \rfloor}(1-F(t))^{n_2-\lfloor 0.9n_2+1 \rfloor} f(t) dt$$

For disease group which includes k patients with overexpressed genes.
$X_{n_1+1}, \ldots, X_{n_1+k} \sim N(\mu_2, \sigma^2)$
$X_{n_1+k+1}, \ldots, X_n \sim N(\mu_1, \sigma^2)$
So, the c.d.f. and p.d.f of X_i for $i = n_1+1, \ldots, n$ are

$$F_1\,(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu_2)^2}{2\sigma^2}} dt = \frac{1}{2}\left(1 + erf\frac{x-\mu_2}{\sqrt{2}\sigma}\right)$$

$$f_1(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$$

And the c.d.f. and p.d.f of X_i for $i = n_1+1, \ldots, n$ are

$$F_2\,(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu_1)^2}{2\sigma^2}} dt = \frac{1}{2}\left(1 + erf\frac{x-\mu_1}{\sqrt{2}\sigma}\right)$$

$$f_2(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

Then, the p.d.f. of 90% quantile of $X_{n_1+1}, \ldots, X_n$ is

$$
\begin{aligned}
& f_3(x) = \frac{n_2!}{\lfloor 0.9n_2 \rfloor !(n_2 - \lfloor 0.9n_2 + 1 \rfloor)!} \cdot \\
& [f_1(x) \sum_{i=0}^{min(k,\lfloor 0.9n_2 \rfloor)} F_2(x)^i (1 - F_2(x))^{k-i} F_1(x)^{\lfloor 0.9n_2 \rfloor} (1 - F_1(x))^{n_2 - \lfloor 0.9n_2+1 \rfloor - k + i} \\
& + f_2(x) \sum_{i=0}^{min(k-1,\lfloor 0.9n_2 \rfloor)} F_2(x)^i (1 - F_2(x))^{k-1-i} F_1(x)^{\lfloor 0.9n_2 \rfloor - 1} (1 - F_1(x))^{n_2 - k + i - \lfloor 0.9n_2 \rfloor}]
\end{aligned}
$$

The cumulative distribution function of the p-value of the disease gene is $1 - G(q^*)$ where $q^* \sim f_3$

References

[1] Bickel, P.J. and Levina, E. (2003) Some theory for Fisher's Linear Discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations.

[2] Fan, J. and Fan, Y. (2007). High Dimensional Class Using Features Annealed Independence Rules.

[3] Tibshirani, R. and Hastie, T. (2007). Outlier sums for differential gene expression analysis. Biostatistics, 8, 1,pp.2-8.

[4] Wu, B. (2007). Cancer outlier differential gene expression detection. Biostatistics, 8, 3, pp. 566-575.

[5] P. Baldi and A.D. Long. (2001). A Bayesian framewoek for the analysis of microarray expression date: regularized t-test and statistical inferences of gene change. Bioinformatics, 17(6):509-519, 2001.

[6] Erik, K., Anders, S, Mats, R., and Olle, N. (2007). Weighted analysis of general microarray experiments.