國 立 交 通 大 學

統計學研究所

碩 士 論 文

由混合變異建構基因表現分析之無母數檢定

Nonparametric Test based on Combined Variation for Gene Expression Analysis

研究生：連紫汝

指導教授：陳鄰安 博士

中華民國 一百 年 六 月

由混合變異建構基因表現分析之無母數檢定

Nonparametric Test based on Combined Variation for Gene Expression Analysis

研 究 生：連紫汝 Student：Tzu-Ju Lian

指導教授：陳鄰安 Advisor：Dr. Lin-An Chen

國 立 交 通 大 學

統計學研究所

碩 士 論 文

A Thesis

Submitted to Institute of Statistics College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of Master

in

Statistics

June 2011

Hsinchu, Taiwan, Republic of China

中華民國 一百 年 六 月

由混合變異建構基因表現分析之無母數檢定

研究生：連紫汝　　　　　　　　　　指導教授：陳鄰安博士

國立交通大學統計學研究所

摘要

在致病基因的檢測問題上，因為 Tomlins et al.(2005)的發現使得探討離群分配變成一個重要主題。不同於離群平均只能檢測中心位置之改變，我們提出一個統計量它同時可以檢測中心與離心兩種變異。這個統計量還有一個好處。由它所建立的檢定統計量不用估計未知分配的密度函數值。我們利用模擬分析比較了幾種檢定方法的檢力並且做了比較。我們也進一步做了一個簡單的實際資料分析。

Nonparametric Test based on Combined Variation for Gene Expression Analysis

Student：Tzu-Ju Lian   Advisor：Dr. Lin-An Chen

Institute of Statistics
National Chiao Tung University

SUMMARY

Observed by Tomlins et al. (2005), detection of the shift for outlier-distribution is a new topic useful in gene expression analysis. Alternative to the outlier mean test, we introduce a nonparametric statistic that can simultaneously detect the location shift and variation shift in the outlier distribution. There is an advantage, comparing with the outlier mean, that the test based on this statistic requires no prediction of distributional densities. Comparisons of this test statistic with some other methods in terms of mean square errors for estimation of their population parameters and powers for their abilities in detection of disease genes are simulated and displayed. Finally, a simple real data analysis is also performed and presented.

誌　謝

畢業將至，回想起兩年的研究所生活，除了在專業上更精進之外，更獲得許多學業以外的知識。

首先，最要感謝陳鄰安老師，謝謝您這些日子來的指導與教誨，從您身上我學到做學問的態度，啟發我對研究的熱忱，都是支持我繼續努力的動力。除此之外，老師也時常教導我許多人生哲理，使我獲益良多。一句謝謝您，代表我最深的敬意。感謝口試委員許文郁老師、謝文萍老師以及黃冠華老師對論文的指導與建議，使論文能更趨完善。

感謝所有的同學，彼此間的互相打氣，與論文奮戰的每一天，你們的陪伴及鼓勵，都支持著我繼續努力下去，也因為有你們，讓我的碩士生涯更多采多姿。謝謝親愛的室友們，不論在研究上、或在生活上妳們都是我最佳的傾聽者，謝謝妳們無私的關心。感謝所辦的郭姐，協助我們處理大小事務，讓我們可以專心做研究。感謝家人的支持，讓我能無後顧之憂的繼續唸書，當我最強力的後盾。

謝謝每個曾經幫助過我的人，在此致上最深的謝意。

連紫汝　謹誌于

國立交通大學統計學研究所

中華民國 一百 年 六 月

目　錄

Nonparametric Test based on Combined Variation for Gene Expression Analysis

SUMMARY

Observed by Tomlins et al. (2005), detection of the shift for outlier-distribution is a new topic useful in gene expression analysis. Alternative to the outlier mean test, we introduce a nonparametric statistic that can simultaneously detect the location shift and variation shift in the outlier distribution. There is an advantage, comparing with the outlier mean, that the test based on this statistic requires no prediction of distributional densities. Comparisons of this test statistic with some other methods in terms of mean square errors for estimation of their population parameters and powers for their abilities in detection of disease genes are simulated and displayed. Finally, a simple real data analysis is also performed and presented.

Key words: Gene expression analysis; Outlier mean; Outlier sum; t-test.

1. Introduction

DNA microarray technology, which simultaneously probes thousands of gene expression profiles, has been successfully used in medical research for disease classification (Agrawal et al. (2002); Alizadeh et al. (2000); Ohki et al. (2005)); Sorlie et al. (2003)). Among the existed techniques in differential genes detection, common statistical methods for two-group comparisons such as t-test, are not appropriate due to a large number of genes expressions and a limited number of subjects available. Several statistical approaches have been proposed to identify those genes where only a subset of the sample genes has high expression. Among them, Tomlins et al. (2005) observed that there is small number of outliers in samples of differential genes and then introduced a method called cancer outlier profile analysis that identifies outlier profiles by a statistic based on the median and the median absolute deviation of a gene expression profile. With this observation, a sequence of approaches then concentrated on detecting differential genes based on outlier samples while Tibshirani and Hastie (2007) and Wu (2007) suggested to use an outlier sum, the sum of all the gene expression values in the disease group that are greater than a specified cutoff point. The common disadvantage of these techniques is that the distribution theory of the proposed methods has not been discovered so that the distribution based p value can not been applied. Recently Chen, Chen and Chan (2010) considered the outlier mean (average of outlier sum) and developed its large sample theory that allows us to formulate the p value based on its asymptotic distribution. For evaluation, they performed simulation studies in a parametric study by specifying the normal distribution. Although the outlier sum or outlier mean is shown interesting in detection of influential genes through statistical analysis and some real data analysis, however, these techniques can detect only the location shift in the outlier distribution, not the distributional variation.

We propose a statistic that can detect simultaneously the location shift and variation shift of the outlier distribution that is generalized from the combined control chart applied in quality control (see Cheng and Thaga (2006) for a review). In Section 2, we present the reasons for the need for the combined outlier quantity. In Section 3, we introduce an asymptotic distribution for the combined outlier quantity and use this theory to introduce a new test for gene expression analysis where a discussion of power based on this new test is given. In Section 4, a comparison between this test and a test combined from the outlier mean and outlier variance is given. Finally, the proofs of theorems are provided in Section 5.

2. Combined Outlier Quantity

In a general study that consists of n_1 subjects in the normal control group and n_2 subjects in the disease group, suppose that there are m genes to be investigated. Their gene expression can be represented as $X_{ij}, i = 1, 2, ..., n_1, j = 1, ..., m$ for normal control group and $Y_{ij}, i = 1, 2, ..., n_2, j = 1, 2, ..., m$ for the disease group. However, in our study, we restrict on one gene with expression variable X for group of normal subject and expression variable Y for group of disease subject where the distribution functions for them are F_X and F_Y respectively. We assume that we have observations $X_i, i = 1, ..., n_1$ and $Y_i, i = 1, ..., n_2$ for our study.

An important observation by Tomlins et al. (2005) from a study of prostate cancer, outlier genes are over-expressed only in a small number of disease samples. With defining a cutoff point $\hat{\eta}$ determined from the data of the variable X, Tibshirani and Hastie (2007) and Wu (2007) considered the sum of variables $Y_i' s$ that are over higher cutoff point $\hat{\eta}$ given by $\sum_{i=1}^{n_2} Y_i I(Y_i \geq \hat{\eta})$ as a test statistic for detection if the disease group distribution is different from the normal group distribution. Latter Chen, Chen and Chan (2010) developed the asymptotic distribution for its average, called the outlier mean, $\bar{Y}_{out} = (\sum_{i=1}^{n_2} I(Y_i \geq \hat{\eta}))^{-1} \sum_{i=1}^{n_2} Y_i I(Y_i \geq \hat{\eta})$ for constructing a distribution based p value. In this paper, we choose $\eta = F_X^{-1}(\gamma)$, the population γth quantile, and $\hat{\eta} = \hat{F}_X^{-1}(\gamma)$, the γth empirical quantile from the sample $X_1, ..., X_{n_1}$. Then, the population type outlier means for distributions of X and Y are

$$\mu_{X,out} = E(X|X \geq F_X^{-1}(\gamma)) \text{ and } \mu_{Y,out} = E(Y|Y \geq F_X^{-1}(\gamma)) \qquad (2.1)$$

and the population type outlier variances are

$$\sigma_{X,out}^2 = Var(X|X \geq F_X^{-1}(\gamma)) \text{ and } \sigma_{Y,out}^2 = Var(Y|Y \geq F_X^{-1}(\gamma)). \quad (2.2)$$

The outlier mean based analysis is to test if $\mu_{Y,out}$ is statistically different from $\mu_{X,out}$ and the outlier variance based analysis is to test if $\sigma_{Y,out}^2$ is statistically different from $\sigma_{X,out}^2$.

For the following two distribution settings,

$$\begin{aligned} &\text{Normal } : X \sim N(0,1), Y \sim N(\theta, \sigma^2), \sigma = 0.5 \\ &\text{Mixed normal } : X \sim N(0,1), Y \sim 0.9N(0,1) + 0.1N(\theta, \sigma^2), \sigma = 0.5 \end{aligned}$$

we choose parameter values of θ such that either outlier means are equal, i.e., $\mu_{X,out} = \mu_{Y,out}$, or outlier variances are equal, i.e., $\sigma_{X,out}^2 = \sigma_{Y,out}^2$. In Table 1, we display, for each distribution setting, two outlier means, two outlier variances.

Table 1. Equal outlier means and equal outlier variances

	θ	$\mu_{X,out}$	$\mu_{Y,out}$	$\sigma^2_{X,out}$	$\sigma^2_{Y,out}$	$\sigma^2_{Y,X}$
Normal						
(I) $\gamma = 0.85$	1.313	1.554	1.554	0.194	0.125	0.125
$\gamma = 0.9$	1.465	1.754	1.754	0.169	0.112	0.112
$\gamma = 0.95$	1.695	2.062	2.062	0.138	0.096	0.096
(II) $\gamma = 0.85$	1.799	1.554	1.866	0.194	0.194	0.292
$\gamma = 0.9$	1.861	1.754	1.977	0.169	0.169	0.218
$\gamma = 0.95$	2.012	2.062	2.210	0.138	0.138	0.159
Mixed Normal						
(I) $\gamma = 0.85$	1.313	1.554	1.554	0.194	0.170	0.170
$\gamma = 0.9$	1.465	1.754	1.754	0.169	0.145	0.145
$\gamma = 0.95$	1.695	2.062	2.062	0.138	0.115	0.115
(II) $\gamma = 0.85$	1.638	1.554	1.630	0.194	0.194	0.200
$\gamma = 0.9$	1.768	1.754	1.833	0.169	0.169	0.175
$\gamma = 0.95$	1.971	2.062	2.140	0.138	0.138	0.144

We have several comments for the results in Table 1:
We see that the outlier means $\mu_{X,out}$ and $\mu_{Y,out}$ for two three γ's in (I) and the outlier variances $\sigma^2_{X,out}$ and $\sigma^2_{Y,out}$ for two three γ's in (II) are all identical. This indicates that for any underlying distribution, there is chance that using outlier mean or outlier variance to test equality of two distributions may not be appropriate.

We then consider a test that can simultaneously interpret the combined change in both outlier mean $\mu_{Y,out}$ and outlier variance $\sigma^2_{Y,out}$. The combined outlier quantity is defined as

$$\sigma^2_{Y,X} = E\{(Y - \mu_{X,out})^2 | Y \geq F_X^{-1}(\gamma)\}.$$

This combined outlier quantity when Y and X have the same distribution is

$$\sigma^2_{X,out} = E\{(X - \mu_{X,out})^2 | X \geq F_X^{-1}(\gamma)\}.$$

The aim of combined outlier quantity is to verify if $\sigma^2_{Y,X}$ and $\sigma^2_{X,out}$ are identical. In Table 1, the values of combined outlier quantity $\sigma^2_{Y,X}$ in all two distributions and different γ's are displayed. With a comparison of $\sigma^2_{Y,X}$ and $\sigma^2_{X,out}$ in all situations, these two quantities are basically not identical.

This allows us to propose a combined outlier quantity based test for gene expression analysis.

We further consider the following three types of distribution setting,

Type 1: $X \sim N(0,1), Y \sim (\chi^2(10)+\theta)$,
Type 2: $X \sim t(10), Y \sim 0.9t(10)+0.1N(\theta,\sigma^2), \sigma=1$,
Type 3: $X \sim t(10), Y \sim 0.9t(10)+0.1(\chi^2(10)+\theta)$.

and present the differnces of outlier means, outlier variances and combined outlier quantity as

$$Df_m = \mu_{Y,out} - \mu_{X,out},\ Df_v = \sigma^2_{Y,out} - \sigma^2_{X,out},\ Df_{comb} = \sigma^2_{Y,X} - \sigma^2_{X,out}$$

in Table 2.

Table 2. Comparison of outlier means and outlier variances

	Df_m	Df_v	Df_{com}
Type 1			
$\theta = 0, \gamma = 0.85$	3.594	25.86	38.78
$\gamma = 0.9$	4.340	27.38	46.22
$\gamma = 0.95$	5.480	27.16	57.20
$\theta = 2$			
$\gamma = 0.85$	4.444	35.10	54.85
$\gamma = 0.9$	5.392	36.60	65.67
$\gamma = 0.95$	6.853	34.83	81.80
$\theta = 4$			
$\gamma = 0.85$	5.296	46.29	74.33
$\gamma = 0.9$	6.444	47.81	89.35
$\gamma = 0.95$	8.232	44.19	111.9
Type 2			
$\theta = 2, \gamma = 0.85$	0.222	0.167	0.216
$\gamma = 0.9$	0.205	0.122	0.164
$\gamma = 0.95$	0.153	0.044	0.067
$\theta = 4$			
$\gamma = 0.85$	0.965	1.519	2.451
$\gamma = 0.9$	1.062	1.337	2.467
$\gamma = 0.95$	1.118	0.953	2.203
Type 3			
$\theta = 0, \gamma = 0.85$	3.517	25.05	37.42
$\gamma = 0.9$	4.218	26.33	44.12
$\gamma = 0.95$	5.245	25.85	53.37
$\theta = 2$			
$\gamma = 0.85$	4.368	34.11	53.19
$\gamma = 0.9$	5.268	35.31	63.08
$\gamma = 0.95$	6.614	33.23	76.99
$\theta = 4$			
$\gamma = 0.85$	5.219	45.12	72.36
$\gamma = 0.9$	6.321	46.29	86.26
$\gamma = 0.95$	7.994	42.30	106.2

It is seen that the differences of combined outlier quantities are much more larger than the other two differences. This probably indicates that the combined outlier quantity may be more efficient in detecting the influential genes.

The sample estimator of combined outlier quantity is defined as

$$S^2_{Y,X} = [\sum_{i=1}^{n_2} I(Y_i \geq \hat{F}_X^{-1}(\gamma))]^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{\mu}_{X,out})^2 I(Y_i \geq \hat{F}_X^{-1}(\gamma)),$$

where the sample outlier mean is $\hat{\mu}_{X,out} = [\sum_{i=1}^{n_1} I(X_i \geq \hat{F}_X^{-1}(\gamma))]^{-1} \sum_{i=1}^{n_1} X_i I(X_i \geq \hat{F}_X^{-1}(\gamma))$. It is also interesting to evaluate the efficiencies in estimating the parameters of outlier mean, outlier variance and combined outlier quantity. We denote the mean square errors for $\mu_{X,out}, \mu_{Y,out}, \sigma^2_{X,out}, \sigma^2_{Y,out}$ and $\sigma^2_{Y,X}$ are, respectively, as $MSE_{\mu_{X,out}}, MSE_{\mu_{Y,out}}, MSE_{\sigma^2_{X,out}}, MSE_{\sigma^2_{Y,out}}$ and $MSE_{\sigma^2_{Y,X}}$. Under the following distribution setting, with $n = 30$,

$$X_1, ..., X_n \text{ iid } N(0,1), Y_1, ..., Y_n \text{ iid } 0.9N(0,1) + 0.1N(\mu, 1)$$

we display these results in Table 3.

Table 3. MSE's comparison for parameters' estimations ($n_1 = n_2 = n = 30$)

	$MSE_{\mu_{X,out}}$	$MSE_{\mu_{Y,out}}$	$MSE_{\sigma^2_{X,out}}$	$MSE_{\sigma^2_{Y,out}}$	$MSE_{\sigma^2_{Y,X}}$
$\mu = 1$					
$\gamma = 0.85$	0.0977	0.0996	0.0288	0.0426	0.1007
$\gamma = 0.9$	0.1235	0.1283	0.0283	0.0382	0.1269
$\gamma = 0.95$	0.2191	0.1788	0.0263	0.0335	0.1580
$\mu = 3$					
$\gamma = 0.85$	0.0981	0.2419	0.0276	0.3354	1.2345
$\gamma = 0.9$	0.1240	0.2895	0.0306	0.3415	1.3698
$\gamma = 0.95$	0.2137	0.3587	0.0265	0.3270	1.8171

It is seen that the MSE's for combined outlier quantity are relatively larger than the other outlier mean and outlier variance quantity. This is due to that a quantity that can simultaneously predict the difference in outlier mean and outlier variance should be more difficult. The appropriateness of the test based on combined outlier quantity needs to be justified through the power comparisons.

3. The Test based on Combined Outlier Quantity

We here introduce some asymptotic properties of the combined outlier quantity and then provide a test based on its asymptotic distribution.

Theorem 3.1. (a)

$$n_2^{1/2}(S_{Y,X}^2-\sigma_{Y,X}^2)=n_1^{-1/2}\sum_{i=1}^{n_1}[\lambda_1(\gamma-I(X_i\le F_X^{-1}(\gamma)))+\lambda_2(X_i-\mu_{X,out})$$
$$I(X_i\ge F_X^{-1}(\gamma))]+\beta_Y^{-1}n_2^{-1/2}\sum_{i=1}^{n_2}[(Y_i-\mu_{X,out})^2-\sigma_{Y,X}^2]I(Y_i\ge F_X^{-1}(\gamma))+o_p(1)$$

where we let

$$\lambda_1=-\,[\beta_Y^{-1}(F_X^{-1}(\gamma)-\mu_{X,out})^2 f_Y(F_X^{-1}(\gamma))f_X^{-1}(F_X^{-1}(\gamma))+2\beta_X^{-1}F_X^{-1}(\gamma)(\mu_{Y,out}-\mu_{X,out})]$$
$$\lambda_2=-\,2\beta_X^{-1}(\mu_{Y,out}-\mu_{X,out})$$

with $\beta_Y=P(Y\ge F_X^{-1}(\gamma)),\beta_X=1-\gamma$.

(b) We have $n_2^{1/2}(S_{Y,X}^2-\sigma_{Y,out}^2)$ converges in distribution to $N(0,v_y)$ where

$$v_y=\gamma(1-\gamma)\lambda_1^2+\lambda_2^2E[(X-\mu_{X,out})^2I(X\ge F_X^{-1}(\gamma))]-2\lambda_1\lambda_2(1-\gamma)$$
$$E[(X-\mu_{X,out})I(X\ge F_X^{-1}(\gamma))]+\beta_Y^{-2}E\{(Y-\mu_{X,out})^4I(Y\ge F_X^{-1}(\gamma))\}-\sigma_{Y,X}^4.$$

where $\beta_Y^{-2}E\{(Y-\mu_{X,out})^4I(Y\ge F_X^{-1}(\gamma))\}-\sigma_{Y,X}^4=Var[(Y-\mu_{X,out})^2|Y\ge F_X^{-1}(\gamma)]$.

From the above theorem, then under $H_0: F_x=F_y$, we have the following,

$$P_{H_0}\{\sqrt{n_2}(\frac{S_{Y,X}^2-\sigma_{X,X}^2}{\sqrt{v_Y}})\le z\}\to\int_{-\infty}^{z}\phi(z)dz$$

for $z\in R$ where ϕ represents the probability density function of $N(0,1)$. If we further have $\hat\sigma_{X,X}^2$ and $\hat v_Y$, respectively, estimates of $\sigma_{X,X}^2$ and v_Y, we may define an outlier combined test as

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}(\frac{S_{Y,X}^2-\hat\sigma_{X,X}^2}{\sqrt{\hat v_Y}})\ge z_\alpha. \tag{3.1}$$

Having this outlier combined test, it is desired to verify the power performance of this test when there exists distributional shift for the disease group distribution. An approximate power with significant level α may be derived as bellows

$$
\begin{aligned}
\pi_Y =& P_{F_Y}\{\sqrt{n_2}(\frac{S^2_{Y,X} - \hat{\sigma}^2_{X,X}}{\sqrt{\hat{v}_Y}}) \geq z_\alpha\} \\
=& P_{F_Y}\{\sqrt{n_2}(\frac{S^2_{Y,X} - \sigma^2_{Y,X}}{\sqrt{v_Y}}) \geq \frac{z_\alpha \sqrt{\hat{v}_Y} + \sqrt{n_2}(\hat{\sigma}^2_{X,X} - \sigma^2_{Y,X})}{\sqrt{v_Y}}\} \\
\approx& P\{Z \geq z_\alpha + \sqrt{n_2}(\frac{\sigma^2_{X,X} - \sigma^2_{Y,X}}{\sqrt{v_Y}})\} \qquad (3.2)
\end{aligned}
$$

The test defined in (3.1) requires that estimator $\hat{v}_Y$ is consistent for parameter v_Y. There is difficulty in providing efficient density estimates involved in λ_1. There is one way to get rid of this difficulty since a level α test is restricted on size α when two distributions F_Y and F_X are identical.

Corollary 3.2. When Y and X have the same distribution, we have, by the fact that $\sigma^2_{X,X} = \sigma^2_{X,out}$,

$$
\begin{aligned}
& n_2^{1/2}(S^2_{Y,X} - \sigma^2_{X,out}) \\
&= -\beta_X^{-1}(F_X^{-1}(\gamma) - \mu_{X,out})^2 n_1^{-1/2} \sum_{i=1}^{n_1}(\gamma - I(X_i \leq F_X^{-1}(\gamma))) \\
&\quad + \beta_X^{-1} n_2^{-1/2} \sum_{i=1}^{n_2}[(X_i - \mu_{X,out})^2 - \sigma^2_{X,out}] I(X_i \geq F_X^{-1}(\gamma)) + o_p(1).
\end{aligned}
$$

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

$$
\begin{aligned}
v_X =& \beta_X^{-2}\gamma(1-\gamma)(F_X^{-1}(\gamma) - \mu_{X,out})^4 \\
&+ \beta_X^{-2} E[(X - \mu_{X,out})^4 I(X \geq F_X^{-1}(\gamma))] - \sigma^4_{X,out}.
\end{aligned}
$$

Suppose that we have estimators $\hat{\sigma}^2_{X,X}$ and $\hat{v}_X$, respectively, for estimation of $\sigma^2_{X,X}$ and v_X. We then can define the following test

$$
\text{Combined test : rejecting } H_0 \text{ if } n_2^{1/2}\frac{S^2_{Y,X} - \hat{\sigma}^2_{X,X}}{\sqrt{\hat{v}_X}} > z_\alpha. \qquad (3.3)
$$

The interest by applying this test of (3.3) is that v_X itself involves no density point so that estimation of it is much easier. We can similarly derive the approximate power for the above test as

$$
\pi_X = P_X\{\sqrt{n_2}(\frac{S^2_{Y,X} - \hat{\sigma}^2_{X,X}}{\sqrt{\hat{v}_X}}) \geq z_\alpha\}. \qquad (3.4)
$$

Power representations (3.2) and (3.4) provide approximate powers based on tests in (3.1) and (3.3). We display the powers of this test (3.1) in Table 4 when the underlying distributions for control group and disease group as

$$X \sim N(0,1) \text{ and } Y \sim (1-\delta)N(0,1) + \delta N(\theta,1).$$

Table 4. Asymptotic power π_Y for mixed normal distribution ($n = 30$)

	$\theta = 1$	$\theta = 3$	$\theta = 5$	$\theta = 10$
$\delta = 0.1$				
$\gamma = 0.8$	0.078	0.281	0.281	0.541
$\gamma = 0.85$	0.074	0.247	0.414	0.534
$\delta = 0.2$				
$\gamma = 0.8$	0.098	0.422	0.682	0.825
$\gamma = 0.85$	0.090	0.345	0.618	0.810

Without simulation study, it is not known if (3.2) and (3.4) present appropriate powers for these two tests. If they are actually in-appropriate, the critical points z_α require an adjustment. We will answer this in next section.

4. Power Comparison by Simulation and a Simple Real Data Analysis

Two tasks will be done in this section. First, we will show by simulation that the setting of critical point z_α of (3.4) by approximation theory is too conservative and we will study present the appropriate level α critical point. Second, we will compare this outlier combined test with a combination of t-test and F-test in terms of power. The classical t-test is designed to detect a change in distributional mean and F-test is to detect a change in distributional variation. Hence, a combination of t-test and F-test is to detect the shift in mean and variation simultaneously. It is then desired to compare powers of these two combined tests.

A t and F combined test is

$$\text{rejecting } H_0 \text{ if } \frac{\bar{Y} - \bar{X}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha/2}(n_1 + n_2 - 2)$$

$$\text{or } \frac{S_X^2}{S_Y^2} > F_{\alpha/2}(n_1 - 1, n_2 - 1) \text{ or } < \frac{1}{F_{\alpha/2}(n_2 - 1, n_1 - 1)}$$

where $S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2+\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}{n_1+n_2-2}$, $S_X^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$.

We consider a simulation with sample size $n = n_1 = n_2$ and replications $m = 100,000$ to evaluate the power when X and Y are from the following setting of distribution:

$$X \sim N(0,1) \text{ and } Y \sim 0.9N(0,1) + 0.1N(\theta,1).$$

In Tables 5 and 6, we display the simulated results for $n = 50$ and $n = 100$ when level of significance is 0.05 and in Table 7, we display the simulated results for $n = 50$ when $\alpha = 0.1$.

We have comments for the results in Tables 5, 6 and 7:

(a) Although the contamination percentage of outlier in mixed normal distribution is small as 0.1 the combined outlier quantity of cutoff with small γ's are more powerful than it with larger γ's.

(b) The tests based on the combined outlier quantity of cutoff with small γ's are relatively more powerful than the t and F combined test. This indicates that simultaneously detect the shift in outlier mean and outlier variance is appropriate when we choose γ appropriately for the cutoff.

(c) The power for the test based on the combined outlier quantity is increasing when the contaminated location shift θ is increasing.

We next consider that alternative distribution has a constant shift as

$$\text{Setting I } : X \sim N(0,1) \text{ and } Y \sim (1-\delta)N(0,1) + \delta\{\theta\}$$
$$\text{Setting II } : X \sim N(0,1) \text{ and } Y \sim (1-\delta)t(10) + \delta\{\theta\}$$

We list the simulated results in Tables 8-11.

We consider a real data of control group and disease group that includes 22,283 genes. Considering the significance level $\alpha = 0.05$, the constants z^*'s in table are the critical points designed to ensure that the sizes of the tests and γ's are appropriately 0.05. Then, we evaluate the percentages of gene numbers to be rejected for all the respective tests in all γ's. The computed results are displayed in Table 12.

Table 12. Percentages of genes larger than critical values ($\alpha = 0.05$)

	Outlier mean	Outlier variance	Combined q	t-test
$\gamma = 0.6$	$0(z^* = 2.68)$	$0.1247(z^* = 2.85)$	$0.1195(z^* = 3.07)$	0.0325
$\gamma = 0.65$	$0(z^* = 2.46)$	$0.1263(z^* = 2.97)$	$0.1168(z^* = 3.28)$	
$\gamma = 0.7$	$0(z^* = 2.21)$	$0.1286(z^* = 3.11)$	$0.1165(z^* = 3.51)$	
$\gamma = 0.75$	$0(z^* = 1.86)$	$0.1256(z^* = 3.42)$	$0.1157(z^* = 3.85)$	
$\gamma = 0.8$	$0(z^* = 1.54)$	$0.1242(z^* = 3.74)$	$0.1092(z^* = 4.35)$	
$\gamma = 0.85$	$0(z^* = 1.23)$	$0.1230(z^* = 4.45)$	$0.1102(z^* = 5.28)$	
$\gamma = 0.9$	$0.00004(z^* = 1.01)$	$0.1247(z^* = 5.54)$	$0.1049(z^* = 7.25)$	
$\gamma = 0.95$	$0.0004(z^* = 0.77)$	$0.1267(z^* = 9.15)$	$0.1154(z^* = 15.5)$	

We have several comments on the results in this table:

(a) It is seen that the outlier mean test performed poorly with very low percentages of genes to be rejected. This shows that it can not detect any gene as influentials.

(b) The tests based on outlier variance and outlier combined quantity are with relatively moderate percentages of genes been claimed influential. Since the genes are measured simultaneously from the same subjects, there is need a simultaneous test that would remarkedly reduce the percentages of genes to be claimed influetial. We will not further pursuit this study. However, we see that only outlier variance and outlier combined quantity are with hope to be able to find genes been influential.

5. Appendix

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

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

ASSUMPTION 2: *Pobability density function* f_X *of distribution* F_X *is bounded away from zero in neighborhoods of* $F_X^{-1}(\alpha)$ *for* $\alpha \in (0,1)$ *and the population cutoff point* η.

ASSUMPTION 3: *Probability density function* f_Y *is bounded away from zero in a neighborhood of the population cutoff point* η.

Proof of Theorem 3.1: First, we consider the following expansion

$$\sum_{i=1}^{n_2}(Y_i-\hat{\mu}_{X,out})^2 I(Y_i\geq \hat{F}_X^{-1}(\gamma))=\sum_{i=1}^{n_2}(Y_i-\mu_{X,out})^2 I(Y_i\geq \hat{F}_X^{-1}(\gamma))$$
$$+(\hat{\mu}_{X,out}-\mu_{X,out})^2\sum_{i=1}^{n_2} I(Y_i\geq \hat{F}_X^{-1}(\gamma))-2(\hat{\mu}_{X,out}-\mu_{X,out})\sum_{i=1}^{n_2}[(Y_i$$
$$-\mu_{Y,out})+n_2(\mu_{Y,out}-\mu_{X,out})]I(Y_i\geq \hat{F}_X^{-1}(\gamma)). \tag{5.1}$$

From the theory for the outlier mean by Chen, Chen and Chan (2010), we may see that $n_2^{1/2}(\hat{\mu}_{Y,out}-\mu_{Y,out})=O_p(1)$, $n_1^{1/2}(\hat{\mu}_{X,out}-\mu_{X,out})=O_p(1)$ and $n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_{Y,out})I(Y_i\geq \hat{F}_X^{-1}(\gamma))=O_p(1)$. We then, from (5.1), may re-write the combined quantity as

$$n_2^{1/2}(S_{Y,X}^2-\sigma_{Y,X}^2)$$
$$=n_2^{1/2}(\sum_{i=1}^{n_2} I(Y_i\geq \hat{F}_X^{-1}(\gamma)))^{-1}\{\sum_{i=1}^{n_2}(Y_i-\mu_{X,out})^2[I(Y_i\geq F_X^{-1}(\gamma)+n_2^{-1/2}T)$$
$$-I(Y_i\geq F_X^{-1}(\gamma))]+\sum_{i=1}^{n_2}[(Y_i-\mu_{X,out})^2-\sigma_{Y,X}^2]I(Y_i\geq F_X^{-1}(\gamma))\}$$
$$-2(\mu_{Y,out}-\mu_{X,out})n_2^{1/2}(\hat{\mu}_{X,out}-\mu_{X,out})+o_p(1), \tag{5.2}$$

where we let $T=n_1^{1/2}(\hat{F}_X^{-1}(\gamma)-F_X^{-1}(\gamma))$.

With Assumptions 2 and 3, and techniques from Ruppert & Carroll (1980) and Chen & Chiang (1996), we may see that

$$n_2^{-1/2}\sum_{i=1}^{n_2}(Y_i-\mu_{X,out})^2[I(Y_i\geq F_X^{-1}(\gamma)+n_2^{-1/2}T^*)-I(Y_i\geq F_X^{-1}(\gamma))]$$
$$=-(F_X^{-1}(\gamma)-\mu_{X,out})^2 f_Y(F_X^{-1}(\gamma))T^* \tag{5.3}$$

for any sequence $T^*=O_p(1)$.

We may also see from Chen, Chen and Chang (2010) that the outlier mean $\hat{\mu}_{X,out}$ has the following representation

$$n_1^{1/2}(\hat{\mu}_{X,out}-\mu_{X,out})=\beta_X^{-1}F_X^{-1}(\gamma)n_1^{-1/2}\sum_{i=1}^{n_1}(\gamma-I(X_i\leq F_X^{-1}(\gamma)))$$
$$+\beta_X^{-1}n_1^{-1/2}\sum_{i=1}^{n_1}(X_i-\mu_{X,out})I(X_i\geq F_X^{-1}(\gamma))+o_p(1). \tag{5.4}$$

The result of this theorem is induced by plugging (5.3) and (5.4) into (5.2) and applying a representation for empirical quantile $\hat{F}_X^{-1}(\gamma)$ in Chen, Chen and Chang (2010). $\square$

Table 5. Power comparison between t and F combined test and outlier combined test ($n = 50, \alpha = 0.05$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.049(\alpha^* = 0.034)$	$0.049(z_{\alpha^*} = 4.02)$
$\gamma = 0.65, \theta = 0$		$0.049(z_{\alpha^*} = 4.47)$
$\gamma = 0.7, \theta = 0$		$0.049(z_{\alpha^*} = 4.89)$
$\gamma = 0.75, \theta = 0$		$0.049(z_{\alpha^*} = 5.74)$
$\gamma = 0.8, \theta = 0$		$0.0473(z_{\alpha^*} = 6.83)$
$\gamma = 0.85, \theta = 0$		$0.0493(z_{\alpha^*} = 9.55)$
$\gamma = 0.9\theta = 0$		$0.0492(z_{\alpha^*} = 14.53)$
$\gamma = 0.6$		
$\theta = 1$	0.092	0.101
$\theta = 3$	0.539	0.755
$\theta = 5$	0.921	0.978
$\theta = 10$	0.994	0.995
$\gamma = 0.65$		
$\theta = 1$		0.097
$\theta = 3$		0.733
$\theta = 5$		0.978
$\theta = 10$		0.995
$\gamma = 0.7$		
$\theta = 1$		0.095
$\theta = 3$		0.717
$\theta = 5$		0.975
$\theta = 10$		0.995
$\gamma = 0.75$		
$\theta = 1$		0.090
$\theta = 3$		0.683
$\theta = 5$		0.972
$\theta = 10$		0.995
$\gamma = 0.8$		
$\theta = 1$		0.085
$\theta = 3$		0.639
$\theta = 5$		0.966
$\theta = 10$		0.994
$\gamma = 0.85$		
$\theta = 1$		0.078
$\theta = 3$		0.562
$\theta = 5$		0.949
$\theta = 10$		0.995
$\gamma = 0.9$		
$\theta = 1$		0.075
$\theta = 3$		0.478
$\theta = 5$		0.912
$\theta = 10$		0.995

Table 6. Power comparison between t and F combined test and outlier combined test ($n = 100, \alpha = 0.05$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.049(\alpha^* = 0.034)$	$0.049(z_{\alpha^*} = 3.07)$
$\gamma = 0.65, \theta = 0$		$0.049(z_{\alpha^*} = 3.28)$
$\gamma = 0.7, \theta = 0$		$0.049(z_{\alpha^*} = 3.51)$
$\gamma = 0.75, \theta = 0$		$0.049(z_{\alpha^*} = 3.85)$
$\gamma = 0.8, \theta = 0$		$0.0484(z_{\alpha^*} = 4.35)$
$\gamma = 0.85, \theta = 0$		$0.0494(z_{\alpha^*} = 5.28)$
$\gamma = 0.9\theta = 0$		$0.0497(z_{\alpha^*} = 7.25)$
$\gamma = 0.6$		
$\theta = 1$	0.123	0.136
$\theta = 3$	0.821	0.949
$\theta = 5$	0.994	0.999
$\theta = 10$	0.999	0.999
$\gamma = 0.65$		
$\theta = 1$		0.128
$\theta = 3$		0.943
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.7$		
$\theta = 1$		0.121
$\theta = 3$		0.933
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.75$		
$\theta = 1$		0.116
$\theta = 3$		0.920
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.8$		
$\theta = 1$		0.108
$\theta = 3$		0.898
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.85$		
$\theta = 1$		0.098
$\theta = 3$		0.854
$\theta = 5$		0.998
$\theta = 10$		0.999
$\gamma = 0.9$		
$\theta = 1$		0.089
$\theta = 3$		0.763
$\theta = 5$		0.997
$\theta = 10$		0.999

Table 7. Power comparison between t and F combined test and outlier combined test ($n = 50$, $\alpha = 0.1$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.099(\alpha^* = 0.067)$	$0.099(z_{\alpha^*} = 2.18)$
$\gamma = 0.65, \theta = 0$		$0.099(z_{\alpha^*} = 2.30)$
$\gamma = 0.7, \theta = 0$		$0.099(z_{\alpha^*} = 2.440)$
$\gamma = 0.75, \theta = 0$		$0.099(z_{\alpha^*} = 2.620)$
$\gamma = 0.8, \theta = 0$		$0.0983(z_{\alpha^*} = 2.93)$
$\gamma = 0.85, \theta = 0$		$0.0993(z_{\alpha^*} = 3.41)$
$\gamma = 0.9\theta = 0$		$0.1001(z_{\alpha^*} = 4.41)$
$\gamma = 0.6$		
$\theta = 1$	0.205	0.229
$\theta = 3$	0.887	0.975
$\theta = 5$	0.997	0.999
$\theta = 10$	0.999	0.999
$\gamma = 0.65$		
$\theta = 1$		0.217
$\theta = 3$		0.973
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.7$		
$\theta = 1$		0.212
$\theta = 3$		0.968
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.75$		
$\theta = 1$		0.202
$\theta = 3$		0.962
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.8$		
$\theta = 1$		0.189
$\theta = 3$		0.950
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.85$		
$\theta = 1$		0.180
$\theta = 3$		0.931
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.9$		
$\theta = 1$		0.166
$\theta = 3$		0.883
$\theta = 5$		0.999
$\theta = 10$		0.999

Table 8. Power comparison between t and F combined test and outlier combined test ($n = 50, \alpha = 0.05, \delta = 0.1$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.049(\alpha^* = 0.034)$	$0.049(z_{\alpha^*} = 4.02)$
$\gamma = 0.65, \theta = 0$		$0.049(z_{\alpha^*} = 4.47)$
$\gamma = 0.7, \theta = 0$		$0.049(z_{\alpha^*} = 4.89)$
$\gamma = 0.75, \theta = 0$		$0.049(z_{\alpha^*} = 5.74)$
$\gamma = 0.8, \theta = 0$		$0.0473(z_{\alpha^*} = 6.83)$
$\gamma = 0.85, \theta = 0$		$0.0493(z_{\alpha^*} = 9.55)$
$\gamma = 0.9\theta = 0$		$0.0492(z_{\alpha^*} = 14.53)$
$\gamma = 0.6$		
$\theta = 3$	0.482	0.717
$\theta = 5$	0.925	0.984
$\theta = 10$	0.995	0.996
$\gamma = 0.65$		
$\theta = 3$		0.685
$\theta = 5$		0.984
$\theta = 10$		0.995
$\gamma = 0.7$		
$\theta = 3$		0.658
$\theta = 5$		0.982
$\theta = 10$		0.995
$\gamma = 0.75$		
$\theta = 3$		0.608
$\theta = 5$		0.978
$\theta = 10$		0.995
$\gamma = 0.8$		
$\theta = 3$		0.554
$\theta = 5$		0.974
$\theta = 10$		0.995
$\gamma = 0.85$		
$\theta = 3$		0.451
$\theta = 5$		0.956
$\theta = 10$		0.995
$\gamma = 0.9$		
$\theta = 3$		0.361
$\theta = 5$		0.917
$\theta = 10$		0.995

Table 9. Power comparison between t and F combined test and outlier combined test ($n = 50, \alpha = 0.05, \delta = 0.2$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.049(\alpha^* = 0.034)$	$0.049(z_{\alpha^*} = 4.02)$
$\gamma = 0.65, \theta = 0$		$0.049(z_{\alpha^*} = 4.47)$
$\gamma = 0.7, \theta = 0$		$0.049(z_{\alpha^*} = 4.89)$
$\gamma = 0.75, \theta = 0$		$0.049(z_{\alpha^*} = 5.74)$
$\gamma = 0.8, \theta = 0$		$0.0473(z_{\alpha^*} = 6.83)$
$\gamma = 0.85, \theta = 0$		$0.0493(z_{\alpha^*} = 9.55)$
$\gamma = 0.9\theta = 0$		$0.0492(z_{\alpha^*} = 14.53)$
$\gamma = 0.6$		
$\theta = 3$	0.883	0.939
$\theta = 5$	0.999	0.999
$\theta = 10$	0.999	0.999
$\gamma = 0.65$		
$\theta = 3$		0.913
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.7$		
$\theta = 3$		0.887
$\theta = 5$		0.999
$\theta = 10$		1
$\gamma = 0.75$		
$\theta = 3$		0.824
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.8$		
$\theta = 3$		0.753
$\theta = 5$		0.999
$\theta = 10$		1
$\gamma = 0.85$		
$\theta = 3$		0.605
$\theta = 5$		0.995
$\theta = 10$		0.999
$\gamma = 0.9$		
$\theta = 3$		0.467
$\theta = 5$		0.975
$\theta = 10$		0.999

Table 10. Power comparison between t and F combined test and outlier combined test ($n = 50, \alpha = 0.05, \delta = 0.1$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.049(\alpha^* = 0.034)$	$0.049(z_{\alpha^*} = 4.02)$
$\gamma = 0.65, \theta = 0$		$0.049(z_{\alpha^*} = 4.47)$
$\gamma = 0.7, \theta = 0$		$0.049(z_{\alpha^*} = 4.89)$
$\gamma = 0.75, \theta = 0$		$0.049(z_{\alpha^*} = 5.74)$
$\gamma = 0.8, \theta = 0$		$0.0473(z_{\alpha^*} = 6.83)$
$\gamma = 0.85, \theta = 0$		$0.0493(z_{\alpha^*} = 9.55)$
$\gamma = 0.9\theta = 0$		$0.0492(z_{\alpha^*} = 14.53)$
$\gamma = 0.6$		
$\theta = 3$	0.620	0.772
$\theta = 5$	0.951	0.987
$\theta = 10$	0.995	0.996
$\gamma = 0.65$		
$\theta = 3$		0.739
$\theta = 5$		0.986
$\theta = 10$		0.996
$\gamma = 0.7$		
$\theta = 3$		0.709
$\theta = 5$		0.986
$\theta = 10$		0.996
$\gamma = 0.75$		
$\theta = 3$		0.658
$\theta = 5$		0.982
$\theta = 10$		0.996
$\gamma = 0.8$		
$\theta = 3$		0.596
$\theta = 5$		0.976
$\theta = 10$		0.996
$\gamma = 0.85$		
$\theta = 3$		0.484
$\theta = 5$		0.956
$\theta = 10$		0.996
$\gamma = 0.9$		
$\theta = 3$		0.384
$\theta = 5$		0.912
$\theta = 10$		0.995

Table 11. Power comparison between t and F combined test and outlier combined test ($n = 50, \alpha = 0.05, \delta = 0.2$)

	p_{tF}	p_{com}
$\gamma = 0.6, \theta = 0$	$0.049(\alpha^* = 0.034)$	$0.049(z_{\alpha^*} = 4.02)$
$\gamma = 0.65, \theta = 0$		$0.049(z_{\alpha^*} = 4.47)$
$\gamma = 0.7, \theta = 0$		$0.049(z_{\alpha^*} = 4.89)$
$\gamma = 0.75, \theta = 0$		$0.049(z_{\alpha^*} = 5.74)$
$\gamma = 0.8, \theta = 0$		$0.0473(z_{\alpha^*} = 6.83)$
$\gamma = 0.85, \theta = 0$		$0.0493(z_{\alpha^*} = 9.55)$
$\gamma = 0.9\theta = 0$		$0.0492(z_{\alpha^*} = 14.53)$
$\gamma = 0.6$		
$\theta = 3$	0.923	0.950
$\theta = 5$	0.999	0.999
$\theta = 10$	0.999	0.999
$\gamma = 0.65$		
$\theta = 3$		0.925
$\theta = 5$		0.999
$\theta = 10$		1
$\gamma = 0.7$		
$\theta = 3$		0.900
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.75$		
$\theta = 3$		0.842
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.8$		
$\theta = 3$		0.768
$\theta = 5$		0.999
$\theta = 10$		0.999
$\gamma = 0.85$		
$\theta = 3$		0.617
$\theta = 5$		0.995
$\theta = 10$		0.999
$\gamma = 0.9$		
$\theta = 3$		0.473
$\theta = 5$		0.973
$\theta = 10$		0.999

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, Dung-Tsa and Chan, Wenyaw. (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.

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. 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.

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