行政院國家科學委員會專題研究計畫 成果報告

由離群值建構的基因分析
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2118-M-009-001-
執 行 期 間 ：98 年 08 月 01 日至 99 年 07 月 31 日
執 行 單 位 ：國立交通大學統計學研究所

計 畫 主 持 人：陳鄰安

計畫參與人員：碩士班研究生-兼任助理人員：刁瀅潔
碩士班研究生-兼任助理人員：林書維
碩士班研究生-兼任助理人員：鄭秋煒
碩士班研究生-兼任助理人員：陳怡頻
碩士班研究生-兼任助理人員：林洋德
博士班研究生-兼任助理人員：魏裕中

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 99 年 10 月 22 日

Report of NSC Project on "Nonparametric Test based on Outlier Mean for Gene Expression Analysis

by

Lin-An Chen

Institute of Statistics, National Chiao Tung University

Contents

Nonparametric Test based on Outlier Mean for Gene Expression Analysis

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 a parametric study with specifying the normal distribution. Although the framework of a test for gene expression analysis based on outlier mean is then established, the understanding applying this outlier mean or outlier sum in nonparametric situation is very limited while gene expression data is generally non-normal. Hence, in this project, we study nonparametric gene expression analysis.

2. Research Purpose

Our purpose in this research is to establish an outlier mean based nonparametric test that is appropriate to be applied for gene expression analysis. First, we show that the outlier mean of Chen, Chen and Chan (2009) is an efficient technique in theoretical power performance. Second, a nonparametric statistical inference procedure may be theoretically very efficient but it is inefficient in practical application when it involves inefficient parame-

ters estimation. We see that the outlier mean based test involves unknown densities at tail quantiles so that its power may be remarkably reduced with inefficient extreme density estimation. Third, we propose an alternative design of outlier mean test that can avoid the difficulty of estimating unknown density poits.

3. Literature Review

There are some manuscripts dealt with approaches closed related to the outlier observations. 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. Chen, Chen and Chan (2010) developed parametric inferences based on outlier mean in gene expression that allows us to formulate the p value based on its asymptotic distribution.

A nonparametric approach allowing to formulate the p value is still not available.

4. Research Methods

The outlier mean proposed by Chen, Chen and Chan (2010) is

$$L_Y = \frac{\sum_{i=1}^{n_2} Y_i I\{Y_i \geq \hat{\eta}\}}{\sum_{i=1}^{n_2} I\{Y_i \geq \hat{\eta}\}}$$

that is to estimate the following population outlier mean

$$\mu_{\ell_Y} = E(Y|Y \geq \eta)$$

where Y_i's are sample from disease group and the cutoff point $\hat{\eta}$ is computed based on sample from normal group data.

In this research, we prove that $\sqrt{n_2}(L_Y - \mu_{\ell_Y})$ converges in distribution to a normal random variable having distribution $N(0, \sigma^2_{\ell_Y})$ for an unknown constants $\sigma^2_{\ell_Y}$. Then under $H_0 : F_x = F_y$, we have the following,

$$P_{H_0}\{\sqrt{n_2}(\frac{L_Y - \mu_{\ell_x}}{\sigma_{\ell_Y}}) \leq z\} \to \int_{-\infty}^{z} \phi(z)dz$$

for $z \in R$ where ϕ represents the probability density function of $N(0, 1)$ where we have μ_{ℓ_x} in the function since the sample outlier mean L_Y is to estimate μ_{ℓ_y} that is supposed to compare with μ_{ℓ_x}. If we have $\hat{\sigma}_{\ell_Y}$ and

$\hat{\mu}_{\ell_X}$, respectively, nonparametric estimates of σ_{ℓ_Y} and μ_{ℓ_X}, we may define an outlier mean based test as

$$\text{rejecting } H_0 \text{ if } n_2^{1/2}(\frac{L_Y - \hat{\mu}_{\ell_X}}{\hat{\sigma}_{\ell_Y}}) \geq z_{\alpha^*}.$$

Having this outlier mean based nonparametric 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 le vel α^* may be derived as bellows

$$\begin{aligned} & P_{F_Y}\{\sqrt{n_2}(\frac{L_Y - \hat{\mu}_{\ell_X}}{\hat{\sigma}_{\ell_Y}}) \geq z_{\alpha^*}\} \\ & = P_{F_Y}\{\sqrt{n_2}(\frac{L_Y - \mu_{\ell_Y}}{\sigma_{\ell_Y}}) \geq \frac{z_{\alpha^*}\hat{\sigma}_{\ell_Y} + \sqrt{n_2}(\hat{\mu}_{\ell_X} - \mu_{\ell_Y})}{\sigma_{\ell_Y}}\} \\ & \approx P\{Z \geq z_{\alpha^*} + \frac{\sqrt{n_2}(\mu_{\ell_X} - \mu_{\ell_Y})}{\sigma_{\ell_Y}}\} \end{aligned}$$

In this research, we consider two cutoff points, $\eta_1 = 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)$ and $\eta_2 = F_X^{-1}(\gamma)$ for studying outlier mean's power performance.

5. Results and Discussions

We have derived the asymptotic variance $\sigma^2_{\ell_Y}$ for cutoff point $\eta_1 = 2F_X^{-1}(1-\alpha) - F_X^{-1}(\alpha)$ as

$$\begin{aligned} \sigma^2_{\ell_Y} &= \alpha(1-\alpha)((1-\alpha)b_1 - \alpha b_2)^2 + 2(1-2\alpha)\alpha^3(b_1+b_2)^2 \\ &\quad + \alpha(1-\alpha)(\alpha b_1 - (1-\alpha)b_2)^2 + \frac{1}{\beta_Y^2}Var\{(Y-\mu_Y)I(Y \geq \eta)\}. \end{aligned}$$

where

$$b_1 = \frac{1}{\beta_Y}(\eta - \mu_Y) f_Y(\eta) \gamma^{1/2} f_X^{-1}(F_X^{-1}(\alpha)),$$
$$b_2 = \frac{-2}{\beta_Y}(\eta - \mu_Y) f_Y(\eta) \gamma^{1/2} f_X^{-1}(F_X^{-1}(1-\alpha)).$$

We have observed that the outlier mean may have satisfactory power performance when we have consistent estimators $\hat{\mu}_{\ell_Y}$ and $\hat{\sigma}_{\ell_Y}$ to construct a test. However, $\hat{\sigma}_{\ell_Y}$ involves estimations of $f_Y(2F_X^{-1}(1-\alpha)), f_X(F_X^{-1}(1-\alpha))$ and $f_X(F_X^{-1}(\alpha))$ while estimation of density function of tail quantile is extremely difficult in practice. Without an alternative proposal avoiding this density estimation, the outlier mean based test won't be powerful in detection of influential genes while the sample sizes in gene expression analysis are generally not allowed to be very large.

Hence, we propose an alternative cutoff point $\eta_2 = F_X^{-1}(\gamma)$. The asymptotic variance of the outlier mean with this cutoff point estimator is

$$\sigma^2_{\ell_Y} = \beta_Y^{-2}(F_X^{-1}(\gamma) - \mu_Y)^2 \gamma_{yx} (f_Y(F_X^{-1}(\gamma)) f_X^{-1}(F_X^{-1}(\gamma)))^2 \gamma(1-\gamma)$$
$$+ \beta_Y^{-2} Var\{(Y - F_X^{-1}(\gamma)) I(Y \geq F_X^{-1}(\gamma))\}.$$

With these two cutoff points, we have studied the power performance

under the following distributional settings:

Normal: $X \sim N(0,1), Y \sim N(\theta, \sigma^2)$,

Mixed normal I: $X \sim N(0,1), Y \sim 0.9N(0,1) + 0.1N(\theta, \sigma^2)$.

Mixed normal II: $X \sim N(0,1), Y \sim 0.8N(0,1) + 0.2N(\theta, \sigma^2)$

Laplace distribution: $X \sim Laplace(0,1)$ and $Y \sim Laplace(\theta, 1)$

t-distribution: $X \sim t(5)$ and $Y \sim t(5) + \theta$,

Case I: $X \sim N(0,1)$ and $Y \sim 0.9N(0,1) + 0.1(\chi^2(10) + \theta)$

Case II: $X \sim t(10)$ and $Y \sim 0.9t(10) + 0.1(\chi^2(10) + \theta)$

6. Judgements for Research Results

We have several comments for the computed results in the paper:

1. The power increases as location parameter θ increses indicating that when there are more wide outliers the outlier means are more efficient in detection of distributional shift.

2. For location shift models (Normal, Laplace and t distributions), the outlier means with cutoff point of larger percentage α is more powerful. Hence, choosing smaller cutoff point (larger α) is advisable for application when there is a locational shift.

3. For a distributional shift of only a small proportion (Mixed normal), the outlier mean with smaller percentage α is more powerful. Hence, choosing larger cutoff point (smaller α) is advisable.

7. 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.

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.

無研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：陳鄰安	計畫編號：98-2118-M-009-001-
計畫名稱：由離群值建構的基因分析	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	5	0	100%	人次	
		博士生	1	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	0	1	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

本計劃提出由離群平均來做基因選取之無母數分析。因基因資料已被證實大都為非常態且在本人之前發表文章於審查時副編輯表示非常態之分析非常重要，故本方法將具有相當之應用價值。