

The Independence of Colon Genes

Ruimei Gao

Abstract – Recent advances in technology to obtain DNA microarrays have made it possible to measure quantitatively the expressions of thousands of genes. These expression levels within subjects may be expected to be correlated. Since the number of subjects is usually quite small compared to the number of genes. Therefore, more and more statisticians are pursuing new methods to solve these large-dimensional problems. In this paper we use the new statistic proposed by Jiang, Bai and Zheng (2013) to test the independence of a microarray dataset, which has expression levels for 6500 human genes, for 40 samples of colon tumor tissue and 22 samples of normal colon tissue.

Keywords – Independence Test, High Dimensional Data, Central Limit Theorem, Colon Genes.

I. INTRODUCTION

Test of independence for random variables is a very classical hypothesis testing problem, which dates back to the seminal work by Pearson [8], followed by a huge literature regarding this topic and its variants. Comprehensive overview and detailed references on this problem can be found in most of the textbooks on multivariate statistical analysis. For instance, here we recommend the masterpieces by Muirhead [7] and by Anderson [2] for more details. However, large dimensional data are occurring in many modern science fields, such as finance, bioinformatics and wireless communication, etc. Unfortunately, many classical statistical methods cannot be directly used for these large dimensional data. Therefore, more and more statisticians are pursuing new methods to solve these large-dimensional problems, such as [3, 6]. This paper will focus on introducing the method proposed by [6] for testing the independence of large sets of random variables and also use it to a microarray dataset, which has expression levels for 6500 human genes, for 40 samples of colon tumor tissue and 22 samples of normal colon tissue. Let us take a review more specifically on some representative existing results in the literature, after necessary notation is introduced. Assume that

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) := N(\mu, \Sigma),$$

in which x_i possesses dimension p_i for $i = 1, 2$, such that $p_1 + p_2 = p$. Denote by μ_i the mean vector of the i -th sub-vector x_i and by Σ_{ij} the cross covariance matrix of x_i and x_j for $i, j = 1, 2$. Then $\mu := (\mu_1', \mu_2')$ and $\Sigma := (\Sigma_{ij})$ are the mean vector and covariance matrix of x respectively.

This work was supported by Science and technology innovation fund project of CUST (XJJLG-2014-01) and The science and technology research program during the twelfth five-year plan period of the education department of Jilin Province (2015-52).

In this work, we consider the following hypothesis testing

$$(T1) \quad H_0 : \Sigma_{12} = 0, \quad \text{v.s.} \quad H_1 : \text{not } H_0.$$

To this end, we draw n observations of x , namely $x_1, \dots, x_n$. In addition, the i -th sub-vector of x_j will be

denoted by x_{ij} for all $i = 1, 2$ and $j \in \{1, \dots, n\}$.

Conventionally, the corresponding sample means will be written as $\bar{x} = n^{-1} \sum_{j=1}^n x_j$ and $\bar{x}_i = n^{-1} \sum_{j=1}^n x_{ij}$. Then, we can construct the sample covariance matrix. Set

$$(1.1) \quad X := [x_1 - \bar{x}, \dots, x_n - \bar{x}], \quad X_i := [x_{i1} - \bar{x}_i, \dots, x_{in} - \bar{x}_i], \quad i = 1, 2.$$

The sample covariance matrix of x and the cross sample covariance matrix of x_m and x_l will be denoted by $\hat{\Sigma}$ and $\hat{\Sigma}_{ml}$ respectively, to wit,

$$\hat{\Sigma} := \frac{1}{n-1} XX', \quad \hat{\Sigma}_{ml} := \frac{1}{n-1} X_m X_l',$$

In the classical large n and fixed p case, the likelihood ratio statistic $\Lambda_n := W_n^{n/2}$ with

$$W_n := \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_{11}| \cdot |\hat{\Sigma}_{22}|}$$

is a favorable one aiming at the testing problem (T1). A celebrated limiting law on Λ_n under H_0 is

$$(1.2) \quad -2\kappa \log \Lambda_n \Rightarrow \chi_\rho^2, \quad \text{as } n \rightarrow \infty,$$

where

$$\kappa = n - \frac{p_1 + p_2 + 3}{2}, \quad \rho = p_1 p_2.$$

One can refer to [11] or Theorem 11.2.5 of [7], for instance. The left hand side of (1.2) is known as the Wilks statistic. It is well known that when the dimension p is very small relative to the sample size n , the theoretical result (1.2) behaves very well. However, it performs badly when the dimension p is large relative to the sample size n . A new method was proposed to deal with the large-dimensional complete independence test by [9]. However, their construction deeply depends on the sum of squared elements of the sample correlation matrix, and such a test would suffer from lack of invariance. Fujikoshi and Sakurai [5] derived the asymptotic distributions of the large dimensional canonical correlations. Their work is closely related, but their method has the similar problem and cannot be directly used for large dimensional independence test. Then Jiang, Bai and Zheng [6] proposed a statistic to discuss the independence test of large dimensional sets of variables by employing RMT. The classical likelihood ratio statistic is defined in (1.4). After some simple calculations, we obtain

$$W_n := \frac{|\hat{\Sigma}_{11}| |\hat{\Sigma}_{22} - \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12}|}{|\hat{\Sigma}_{11}| \cdot |\hat{\Sigma}_{22}|} = |I_{p_2} - \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1}|$$

Thus, consider the test statistic

$$(1.3) \ln(W_n) = \ln |I_{p_2} - \hat{\Sigma}_{21} \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1}|.$$

Jiang, Bai and Zheng[6] showed that if $r_{n1} = p_2 / p_1 \rightarrow r_1 \in (0, +\infty)$ and

$r_{n2} = p_2 / (n - 1 - p_1) \rightarrow r_2 \in (0, 1)$, then

$$\frac{\ln W_n - p_2 \cdot e(f) - m(f)}{\sqrt{v(f)}}$$

converges weakly to standard normal distribution. Here

$$e(f) = \ln(\alpha_n (1 - r_{n2})^2) + \frac{1 - r_{n2}}{r_{n2}} \ln(c_n)$$

$$+ \frac{1 - r_{n1}}{r_{n1}} \ln(c_n - d_n h_n) - \frac{r_{n1} + r_{n2}}{r_{n1} r_{n2}} \ln \frac{c_n h_n - d_n r_{n2}}{h_n},$$

$$m(f) = \frac{1}{2} \ln \frac{(ch - r_2 d)^2}{(c^2 - d^2) h^2}, v(f) = 2 \ln \frac{c^2}{c^2 - d^2},$$

$$h = \sqrt{r_1 + r_2 - r_1 r_2}, \alpha = (n - 1 - p_1) p_1^{-1},$$

and

$$c, d = \frac{1}{2} (\sqrt{(1+h)^2 + r_1 r_2^{-1} (1-r_2)^2} \pm \sqrt{(1-h)^2 + r_1 r_2^{-1} (1-r_2)^2})$$

$c > d > 0$.

Now we apply the proposed test statistic to a microarray dataset, which comes from a study of gene expression in colon adenocarcinoma. Gene expression levels have been measured using the Affymetrix technology. This microarray dataset was derived from 40 tumor and 22 normal colon tissues for more than 6500 human genes, i.e. $n = 62$. A section of 2000 genes with highest minimal intensity across the samples was made by Alon et al. [1]. The description of the dataset can be found in [4] and [10]. The preprocessed dataset is from <http://www.tatsuya.e.u-tokyo.ac.jp/>, where the expression levels have been transformed by the $\log_{10}$ transformation.

It was shown in [10] that the covariance matrices of tumor group and normal group should be equal. Hence we can use the proposed test statistic to investigate the independence of these genes. Before that we need the following preprocessing steps: (i) reorder: reorder the genes according to the magnitude of the t-statistic for comparing the two groups. (ii) partition: partition the reordered genes into 80 blocks in sequence with each one containing 25 genes. (iii) screen: get rid of the blocks which have singular sample covariance matrices. Then it remains 77 blocks and 1925 genes. For testing the independence of each two of these 77 block genes, based on the proposed test statistic, we get all the p -values are essentially one. Thus there is strong evidence that the 77 block genes are independent.

REFERENCES

- [1] Alon, U., Barkai, N., Motterman, D., Gish, K., Mack, S. and Levine, J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.*, 96(1999), 6745-6750.

- [2] Anderson, T., An introduction to multivariate statistical analysis. Third Edition. Wiley New York, 2003.
- [3] Bai, Z. D., Jiang, D. D., Yao, J. F. and Zheng, S. R., Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* 37(2009), 3822-3840.
- [4] Dettling, M. and Bühlmann, P., Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(2002), 1061-1069.
- [5] Fujikoshi, Y. and Sakurai, T., High-dimensional asymptotic expansions for the distributions of canonical correlations. *J. Multivariate Anal.*, 100(2009), 231-242.
- [6] Jiang D. D., Bai Z. D., Zheng S. R., Testing the independence of sets of large-dimensional variables. *Science China Mathematics*, 56 (2013), 135-147.
- [7] Muirhead, R. J., Aspects of Multivariate Statistical Theory. Wiley, New York, 1982.
- [8] Pearson, K., On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50 (1900), 157-175.
- [9] Schott, J. R., Testing for complete independence in high dimensions. *Biometrika*, 92(2005), 951-956.
- [10] Srivastava, M. S. and Reid, N., Testing the structure of the covariance matrix with fewer observations than the dimension. *Journal of Multivariate Analysis*, 112(2012), 156-171.
- [11] Wilks, S. S., On the independence of k sets of normally distributed statistical variables. *Econometrica*, 3(1935), 309-326.

AUTHOR'S PROFILE



Ruimei Gao (1983-)

Female, PhD, graduated in Northeast Normal University in Changchun, Jilin, P. R. China. The degree was earned in 2012. The author's major field of study is multivariate statistical analysis and singularities theories.

She works in Department of Science, Changchun University of Science and Technology after graduation. The published articles are as follows:

- [1] L.L. Kong, R.M. Gao, D.H. Pei, J.H. Zhang, Singularities of generic lightcone pedal surfaces of spacelike curves in Minkowski 4-space, *Houston Journal of Mathematics* [J], 2010, 36 (3), 697-710.
- [2] Z.G. Wang, R.M. Gao, X.M. Fan, Q.X. Han, Stability analysis of multi-group deterministic and stochastic epidemic models with vaccination rate [J]. *Chinese Physics B*, 2014, 23(9), 090201-01-090201-16.
- [3] R.M. Gao, D.H. Pei, The Supersolvable Order of Hyperplanes of an Arrangement, *Communication in Mathematical Research*, 2013, 29(3): 231-238.

The current and previous research interests are multivariate statistical analysis and singularities theories.