Statistics and Probability Letters 97 (2015) 206–211

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Statistical properties of gene–gene correlations in omics experiments

Huaizhen Qin *,[1], Weiwei Ouyang [1]

Department of Biostatistics and Bioinformatics, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, New Orleans, LA 70112, United States

ARTICLE INFO

Article history:
Received 8 July 2014
Received in revised form 20 October 2014
Accepted 28 November 2014
Available online 8 December 2014

MSC:
primary 62J99
secondary 62G15

Keywords:
Transcriptomics
Proteomics
Differential magnitudes
Stochastic representations
Tight clustering

ABSTRACT

In this article, we obtain generic stochastic representations and asymptotic distributions of gene–gene correlations with respect to differential magnitudes, residual correlations, and sample size of experiment. Our results establish theoretical foundation for tight clustering of co-expressed genes.

1. Introduction

Advanced high-throughput technologies can assay expressions of thousands of genes simultaneously. Massive multi-omics data provide the potential to identify changes at RNA and protein levels between different biological (e.g., different developmental stages) or experimental conditions (e.g., treatment and control). In transcriptomics and proteomics experiments, two fundamental challenges are ranking and identifying differentially expressed genes given moderate sample sizes (typically called a "large p small n" problem). Various statistical procedures have been developed for gene ranking in two-sample transcriptomics and proteomics experiments, e.g., the Significance Analysis of Microarrays (SAM) t (Tusher et al., 2001), the moderated t (Smyth, 2004) and the Welch type t (Hu and Wright, 2007). These statistics are constructed to pool information across genes, and thus may improve ranking efficiency over the standard two-sample t. However, these popular methods do not explicitly exploit the correlations between co-expressed genes. As a result, these approaches could claim too many false positives (Morley et al., 2004; Wacholder et al., 2004).

To prevent the flood of false positives, multiple-testing corrections should be employed, e.g., Bonferroni correction (Hochberg, 1988; Holm, 1979) and false discovery rate (FDR) approaches (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). For this, one may appreciate the explicit null distribution of the adopted statistic to claim which genes are significantly differentially expressed. If the null distribution is intractable, e.g., that of the SAM t, one has to rely on permutation to claim significantly differentially expressed genes and estimate their corresponding FDRs. It is well-known that

* Corresponding author.
E-mail addresses: hqin2@tulane.edu (H. Qin), wouyang@tulane.edu (W. Ouyang).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.spl.2014.11.026
0167-7152/

permutation is not an appropriate method when the sample size is small (e.g., $n < 10$). There are other procedures in the literature that can handle correlations (Efron, 2007; Pawitan et al., 2006; Qiu et al., 2005), which exhibit some improved performances over standard procedures, but such improvements are mostly expressed in terms of reduced variance. The FCPC approach (Qin et al., 2008) explicitly utilizes gene–gene correlation to boost statistical power while preventing false positives in expression experiments. Again, this approach was developed according to empirical observations. The WGCNA R software package (Langfelder and Horvath, 2008) provided comprehensive functions for performing various aspects of weighted correlation network analysis.

As to be theoretically shown, gene–gene correlation in expression data has particular properties. In multivariate statistics, there exist theoretical results on sample correlation (Fang and Zhang, 1990; Ruben, 1966). Such results assume that all the bivariate data points are generated from an identical bivariate normal distribution. In a two-condition experiment (e.g., treatment and control), however, the distribution of bivariate data points in treatments differs from that in controls. Thus, the Bartlett decomposition and extensions (Li and Geng, 2003) in conventional correlation theory do not apply. The remaining of this article is organized as follows. In Section 2, we provide problem formulation to model the gene–gene correlations in transcriptomics and proteomics and connect it with the tight clustering algorithm (Qin et al., 2008). In Sections 3–5, we theoretically and numerically investigate the exact and asymptotic statistical behaviors of gene–gene correlation under different scenarios. In Section 6, we summarize the results of our theoretical and numerical findings. All technical proofs and numerical illustrations are presented in the Supplementary (see Appendix A).

2. Problem formulation

In two-condition (e.g., treatment and control) transcriptomics/proteomics experiments, the bivariate normal model is commonly employed, either implicitly or explicitly, to describe the relationship between the expression levels (x, y) of two genes. For ease of understanding, we focus on the following bivariate model

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{cases} (u_i, v_i)' & \text{for } i = 1, \ldots, n_0, \\ (\mu + u_i, \nu + v_i)' & \text{for } i = n_0 + 1, \ldots, n_0 + n_1 \stackrel{\text{def}}{=} n, \end{cases} \tag{1}$$

where (x_i, y_i)'s denote observed expression levels of subject i, μ and ν denote unknown real treatment means and (u_i, v_i)'s denote a set of mutually independent bivariate normal vectors of mean 0, variance 1, and unknown correlation $\rho \in (-1, 1)$, namely,

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{2}$$

Under this model, μ and ν represent real treatment mean expression levels of two genes, respectively, and ρ reflects the residual dependency between the two expression levels except for the treatment effects. As in classical context, the Pearson coefficient of correlation between the expression levels of these two genes, i.e., $\boldsymbol{x} = (x_1, \ldots, x_n)'$ and $\boldsymbol{y} = (y_1, \ldots, y_n)'$, is defined as

$$r_{xy} = (\boldsymbol{x}'\boldsymbol{y} - n\bar{x}\bar{y})/\sqrt{\left(\boldsymbol{x}'\boldsymbol{x} - n\bar{x}^2\right)\left(\boldsymbol{y}'\boldsymbol{y} - n\bar{y}^2\right)}, \tag{3}$$

where $\bar{x}$ denotes the mean of all x_i's, and $\bar{y}$ denotes the mean of all y_i's, respectively. When $\mu = \nu = 0$, namely, both genes are stably expressed, (x_i, y_i)'s are independent duplicates from the bivariate normal distribution with zero mean and correlation ρ. For this case, an unbiased estimator of ρ can be constructed in terms of r_{xy} (Olkin and Pratt, 1958), and the properties of r_{xy} per se are well addressed in the literature, which are summarized in Section 3. However, when $\mu^2 + \nu^2 \neq 0$, the statistic r_{xy} has different properties from the ordinary sample correlation. Sections 4 and 5 address the properties of r_{xy} when one and both of μ and ν are zero, respectively.

In forward search clustering, we can assign a gene to a cluster if the minimum of the correlations between the gene and the genes in the cluster exceeds a tightness threshold $\rho_0 (> 0)$. For any pair of genes, define $\tilde{r}_{xy} = r_{xy}/\sqrt{1 - r_{xy}^2}$. Then, the tail probability

$$\psi(\rho_0) \stackrel{\text{def}}{=} \Pr\left(r_{xy} \geq \rho_0\right) = \Pr\left(\sqrt{n-2}\tilde{r}_{xy} \geq \sqrt{n-2}\tilde{\rho}_0\right) \tag{4}$$

measures the probability to cluster them to an identical cluster of tightness $\rho_0 \in (0, 1)$, where $\tilde{\rho}_0 = \rho_0/\sqrt{1 - \rho_0^2}$. It is well known that $\sqrt{n-2}\tilde{r}_{xy} \sim t_{n-2}$ if the two genes are stably expressed across controls and treatments with $\rho = 0$ (Section 3). In context, we investigate the distributional properties of r_{xy} and the performances of the tail probability ψ.

3. Statistical properties of correlations between stably expressed genes

In the scenario of $\mu = \nu = 0$, r_{xy} is a well-known consistent estimator for ρ as sample size n tends to infinity. For large n, Ruben (1966) obtained a simple approximate normalization for $\tilde{r}_{xy}$. This approximation outperforms the usual large-sample