Computational Statistics and Data Analysis 56 (2012) 510–521

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Permutation test for incomplete paired data with application to cDNA microarray data

Donghyeon Yu [a], Johan Lim [a], Feng Liang [b], Kyunga Kim [c], Byung Soo Kim [d], Woncheol Jang [e,*]

[a] Department of Statistics, Seoul National University, Seoul, Republic of Korea
[b] Department of Statistics, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, USA
[c] Department of Statistics, Sookmyung Women's University, Seoul, Republic of Korea
[d] Department of Applied Statistics, Yonsei University, Seoul, Republic of Korea
[e] Department of Epidemiology and Biostatistics, University of Georgia, 30602 Athens, GA, USA

ARTICLE INFO

Article history:
Received 15 December 2010
Received in revised form 19 August 2011
Accepted 21 August 2011
Available online 7 September 2011

Keywords:
Colorectal cancer
Incomplete paired data
Microarray data
Permutation test

ABSTRACT

A paired data set is common in microarray experiments, where the data are often incompletely observed for some pairs due to various technical reasons. In microarray paired data sets, it is of main interest to detect differentially expressed genes, which are usually identified by testing the equality of means of expressions within a pair. While much attention has been paid to testing mean equality with incomplete paired data in previous literature, the existing methods commonly assume the normality of data or rely on the large sample theory. In this paper, we propose a new test based on permutations, which is free from the normality assumption and large sample theory. We consider permutation statistics with linear mixtures of paired and unpaired samples as test statistics, and propose a procedure to find the optimal mixture that minimizes the conditional variances of the test statistics, given the observations. Simulations are conducted for numerical power comparisons between the proposed permutation tests and other existing methods. We apply the proposed method to find differentially expressed genes for a colorectal cancer study.

1. Introduction

Testing the equality of means in a paired data set is a problem which we often encounter in practice. In particular, when one wants to identify differentially expressed genes between normal and tumor tissues in microarray experiments, it is common to extract RNA samples from both normal and tumor tissues of the same individual. This indirect design is usually adopted for the experiment that uses a common reference RNA hybridized both to normal and tumor tissues. Ideally, this experiment produces n pairs of microarray data. However, for some individuals, either the normal tissue or the tumor tissue is not big enough for the experimenter to extract enough RNA for conducting the experiment. For this reason, one of the paired expression levels often fails to be assessed for these individuals.

A simple approach for the unpaired missing data is to remove those unpaired incomplete observations from the analysis, but this would lead to a power loss in testing because tests based on all the available data are supposed to have better performance. Another approach to deal with missing data is imputation, but this may work only when incomplete pairs

* Corresponding author.
E-mail addresses: bunguji@snu.ac.kr (D. Yu), johanlim@snu.ac.kr (J. Lim), liangf@uiuc.edu (F. Liang), kyunga@sookmyung.ac.kr (K. Kim), bskim@yonsei.ac.kr (B.S. Kim), jang@uga.edu (W. Jang).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.08.012

occur for a small number of genes. We may have, however, a sizeable number of missing arrays in either the tumor or the normal tissues.

Much attention has been paid to tests for incomplete paired data in previous literature. Many of the proposed tests are based on estimates of the mean difference using all the available data (Lin and Stivers, 1974; Ekbohm, 1974; Bhoj, 1974). These existing methods heavily rely on the assumption that the data are generated from a bivariate normal, but this assumption might not be true for gene expression data.

In this study, we develop a statistical method for incomplete paired data, which does not require any parametric assumption and utilizes the information from both paired and unpaired data. We consider the class of convex combinations of two linear statistics $\mathbf{T}_{\mathrm{p}}$ and $\mathbf{T}_{\mathrm{up}}$:

$$\mathbf{T}_{\lambda} = \lambda \mathbf{T}_{\mathrm{up}} + \big(1-\lambda\big)\mathbf{T}_{\mathrm{p}}, \quad 0 \le \lambda \le 1,$$

where $\mathbf{T}_{\mathrm{p}}$ and $\mathbf{T}_{\mathrm{up}}$ are statistics for paired and unpaired samples, respectively. The optimal combination is further chosen to minimize the conditional variance of permuted $\mathbf{T}_{\lambda}$, given the observations.

The rest of this paper is organized as follows. In Section 2, we introduce a statistical model for paired gene expression data and review the existing methods for incomplete paired data. In Section 3, we propose a permutation method for testing the difference of means. We numerically investigate the performance of our method based on simulations in Section 4. In Section 5, we apply the proposed method to identify differentially expressed genes from the colorectal cancer microarray data which previously studied by Kim et al. (2005). Finally, we discuss possible extensions of our method in Section 6.

2. Existing methods

We represent incomplete paired data as a combination of paired and unpaired data sets as follows. Let (X_i, Y_i) $(i = 1, 2, \ldots, n_3)$ be paired observations, for the ith individuals under two different conditions. Denote X_i^{U} $(i = 1, 2, \ldots, n_1)$ and Y_j^{U} $(j = 1, 2, \ldots, n_2)$ as unpaired observations on X and Y, respectively. Then, the entire data set can be arranged as follows:

$$\begin{array}{lll} X_1^{\mathrm{U}}, \ldots, X_{n_1}^{\mathrm{U}} & ; & ; X_1, \ldots, X_{n_3} \\ & ; Y_1^{\mathrm{U}}, \ldots, Y_{n_2}^{\mathrm{U}} & ; Y_1, \ldots, Y_{n_3}. \end{array}$$

Suppose that $(X_i, Y_i)^{\mathrm{T}}$ is a bivariate normal random vector with mean $\boldsymbol{\mu} = (\mu_X, \mu_Y)^{\mathrm{T}}$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{XX} & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_{YY} \end{pmatrix}.$$

Then the marginal distributions of unpaired data are normal with mean μ_X and variance σ_{XX} for X^{U}, and normal with mean μ_Y and variance σ_{YY} for Y^{U}, respectively. We define $\overline{X}$, $\overline{Y}$, S_{XX}, S_{YY}, S_{XY}, and $r = S_{XY}/\sqrt{S_{XX}S_{YY}}$ as the sample means, the sample variances, the sample covariance, and the sample correlation for the paired data, respectively. We further define $\overline{X}^{\mathrm{U}}$, $\overline{Y}^{\mathrm{U}}$, S_{XX}^{U}, and S_{YY}^{U} as the sample means and the sample variances for the unpaired data. The primary interest here is testing the mean difference $\delta = \mu_X - \mu_Y$.

When missing data occur only on one side (i.e., $n_1 = 0$ or $n_2 = 0$), one can obtain a closed form of the maximum likelihood estimate (MLE) of $\boldsymbol{\mu}$ and Σ (Anderson, 1957). However, there is no closed form of the MLE for the incomplete data, when missing data occur on both sides (i.e., $n_1 \neq 0$ and $n_2 \neq 0$), though one can find the MLE of $\boldsymbol{\mu}$ and Σ with an iterative procedure (Orchard and Woodbury, 1970). To get a closed form of an estimator, Lin and Stivers (1974) proposed a modified MLE with non-iterative procedures and provided several test statistics based on their estimator. Ekbohm (1974) and Bhoj (1974) also suggested similar, but simpler, test statistics. In particular, Bhoj (1974) proposed a test statistic that was a linear combination of the paired and the unpaired t-statistics:

$$\mathbf{T}_{\mathrm{Bhoj}} = \lambda \frac{\overline{X}^{\mathrm{U}} - \overline{Y}^{\mathrm{U}}}{S_p^{\mathrm{U}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} + \big(1-\lambda\big)\frac{\overline{X}-\overline{Y}}{S_p/\sqrt{n_3}},$$

where

$$\big(S_p^{\mathrm{U}}\big)^2 = \frac{(n_1-1)S_{XX}^{\mathrm{U}} + (n_2-1)S_{YY}^{\mathrm{U}}}{n_1+n_2-2} \quad \text{and} \quad S_p^2 = \frac{S_{XX}+S_{YY}-2S_{XY}}{n_3}.$$

The test statistic is a linear combination of two independent t-distributed random variables. Thus, it can be adequately approximated by a t-distribution with f_B degrees of freedom, where f_B is calculated by equating the second and fourth moments of the test statistic and the t-distribution (Patil, 1965).

Recently, Kim et al. (2005) considered a nonlinear combination of paired and unpaired statistics, which is referred to as $\mathbf{t}_3$:

$$\mathbf{t}_3 = \frac{n_3\big(\overline{X}-\overline{Y}\big) + n_{\mathrm{H}}\big(\overline{X}^{\mathrm{U}}-\overline{Y}^{\mathrm{U}}\big)}{\sqrt{n_3 S_p^2 + n_{\mathrm{H}}^2\left(\frac{S_{XX}^{\mathrm{U}}}{n_1}+\frac{S_{YY}^{\mathrm{U}}}{n_2}\right)}},$$