Computational Statistics and Data Analysis 97 (2016) 87–97

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A high-dimension two-sample test for the mean using cluster subspaces[☆]

Jie Zhang[a], Meng Pan[b,*]

[a] Department of Physics, Jinan University, Guangzhou 510632, China

[b] Department of Optoelectronic Engineering, Jinan University, Guangzhou 510632, China

HIGHLIGHTS

- A two-sample test using hierarchical clustering was proposed.
- Hotelling's statistics are computed in cluster-subspaces and summed as the statistic.
- Highly correlated variables take priority for being processed.
- A cutoff distance is used to restrain the effect of statistical fluctuations.
- High performance was demonstrated in simulations and real data analysis.

ARTICLE INFO

Article history:
Received 12 June 2015
Received in revised form 1 December 2015
Accepted 3 December 2015
Available online 14 December 2015

Keywords:
High-dimension data
Two-sample problem
Hierarchical clustering
Hotelling's test

ABSTRACT

A common problem in modern genetic research is that of comparing the mean vectors of two populations – typically in settings in which the data dimension is larger than the sample size – where Hotelling's test cannot be applied.

Recently, a test using random subspaces was proposed, in which the data are randomly projected into several lower-dimensional subspaces, and Hotelling's test is well defined. Superior performance with competing tests was demonstrated when the variables were correlated.

Following the research of random subspaces, a modified test was proposed that might make more efficient use of covariance structure at high dimension. Hierarchical clustering is performed first such that highly correlated variables are clustered together. Next, Hotelling's statistics are computed for every cluster-subspace and summed as the new test statistic. High performance was demonstrated via simulations and real data analysis.

1. Introduction

A common problem in genetics is that of testing whether a set of dependent gene expressions differs between two populations, typically in a setting where the data dimension is larger than the sample size.

For correlated variables, a classic test is Hotelling's test. For instance, two samples $X = (X_1, \ldots, X_{n_1})$ and $Y = (Y_1, \ldots, Y_{n_2})$ of size n_1 and n_2 are generated in an independent and identically distributed (i.i.d.) manner from p-dimensional

[☆] Supplementary materials related to the implementation of the proposed method are available online (see Appendix A).

[*] Corresponding author. Tel.: +86 13620400960; fax: +86 20 85224401.
E-mail address: tpanm@jnu.edu.cn (M. Pan).

http://dx.doi.org/10.1016/j.csda.2015.12.004

multivariate normal distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, respectively, where the mean vectors μ_1 and μ_2 and positive-definite covariance matrix Σ are all fixed and unknown; the hypothesis testing problem of interest is

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2,$$

the Hotelling's test statistic is defined by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\overline{X} - \overline{Y})^T \widehat{\Sigma}^{-1}(\overline{X} - \overline{Y}),$$

where $\overline{X} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_j$ and $\overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$ are the sample means, $\widehat{\Sigma}$ is the pooled sample covariance matrix, given by $\widehat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n_1}(X_j - \overline{X})(X_j - \overline{X})^T + \frac{1}{n}\sum_{j=1}^{n_2}(Y_j - \overline{Y})(Y_j - \overline{Y})^T$, and $n = n_1 + n_2 - 2$ (Anderson, 1984). Under the null hypothesis H_0, $\frac{n-p+1}{np}T^2$ has an F-distribution with degrees of freedom p and $n - p + 1$ (Muirhead, 2005). Let the significance be chosen as α and the threshold be denoted as $F_\alpha(p, n - p + 1) = \frac{n-p+1}{np}T_\alpha^2$. Under the null hypothesis H_0, the probability of $T^2 \geq T_\alpha^2$ is called the false positive (type I error) rate. Under the alternative hypothesis H_1, the probability of $T^2 \geq T_\alpha^2$ is called the true positive rate (power).

By using a test suitable for correlated variables, it is possible not only to take the multivariate dependence structure into account but to gain more power from these dependences (Thulin, 2014). However, when $p > n$, the matrix $\widehat{\Sigma}$ is singular, and Hotelling's test cannot be applied.

Many studies have addressed the "large p, small n" problem. Chung and Fraser (1958) proposed a nonparametric test that treats each variable independently. Dempster (1958, 1960) proposed a so-called "non-exact" significance test based on the quantity $(\overline{X} - \overline{Y})^T(\overline{X} - \overline{Y})$, which can be viewed as replacing $\widehat{\Sigma}$ with $I_{p\times p}$. It was later refined by Bai and Saranadasa (1996) and Chen and Qin (2010). However, the statistics based on $(\overline{X} - \overline{Y})^T(\overline{X} - \overline{Y})$ lack desirable invariance properties under rescaling transformation. Srivastava and Du (2008) proposed a test based on $(\overline{X} - \overline{Y})^T D_{\widehat{\Sigma}}^{-1}(\overline{X} - \overline{Y})$, where $D_{\widehat{\Sigma}}$ is the diagonal matrix associated with $\widehat{\Sigma}$, i.e., $(\widehat{\Sigma})_{ii} = (D_{\widehat{\Sigma}})_{ii}$. To make use of the multivariate dependence structure, Srivastava (2007) proposed using the Moore–Penrose pseudo-inverse of $\widehat{\Sigma}$ when computing Hotelling's test statistic. Cai et al. (2014) applied a regularization technique to obtain a sparse estimator of the matrix and proposed a test statistic. Shen and Lin (2015) proposed a test that selects important variables against the null hypothesis.

To make more use of the multivariate dependence structure, Lopes et al. (2011, 2012) proposed a test in which the data are randomly pseudo-projected into several lower-dimensional spaces, where Hotelling's test is well defined. Hotelling's T^2 statistic is computed for each pseudo-projection, and the result is then averaged over all pseudo-projections. Superior performance with competing tests was demonstrated when the variables were correlated. Thulin (2014) proposed a modified test using random subspaces to improve the invariance properties. Random permutation resampling was utilized to improve the null distribution of the statistic.

This study followed the research of Thulin (2014) and proposed a test that can make more efficient use of the covariance structure at high dimension. Hierarchical clustering is performed first so that highly correlated variables are clustered together. Then, Hotelling's T^2 statistics are computed for every cluster-subspace and summed as the new test statistic.

The rest of the paper is organized as follows. In Section 2, we propose the new test based on cluster subspaces. In Section 3, we compare the test with other two-sample tests with Monte Carlo simulations. The new test and the other tests are applied to a breast cancer dataset in Section 4. In Section 5, we discuss how the clustering method can group the variables successfully. Conclusions are presented in Section 6.

2. Cluster subspaces test

In Thulin's random subspaces test, variables are randomly selected to construct subspaces in which Hotelling's statistics are computed; therefore, the correlations between the variables are utilized and higher power is obtained. In the test in this study, hierarchical clustering is performed so that highly correlated variables are grouped together; therefore, the covariance structure might be more efficiently utilized. By clustering, the high dimension data are also projected to cluster subspaces of lower dimension in which Hotelling's statistics can be computed and summed as the new test statistic.

2.1. Hierarchical clustering

Hierarchical clustering was widely used in modern genetic research to find related genes or individuals. By clustering, variables with high similarity metrics (or low distances) are grouped together (Eisen et al., 1998). In this study, 1—Pearson correlation coefficient was used as the distance. Highly correlated and thus small distance variables would be clustered together. The correlation coefficients are also related to the covariance matrix $\widehat{\Sigma}$ (if the distributions of all the variables are normalized so that the variances are equal to 1, the covariances are equal to the correlation coefficients).

There may be statistical fluctuations for the correlation coefficients. For large p, some of the coefficients may be large by chance. To restrain the effect of statistical fluctuations, clusters were first calculated based on a cutoff distance d_c. The Fisher z'-transformation of correlation coefficient r is $z' = \frac{1}{2}\ln\frac{1+r}{1-r}$. The standard error of z' is $\sigma_{z'} = 1/\sqrt{n-1}$