Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A high-dimensional two-sample test for the mean using random subspaces[☆]

Måns Thulin[*]

Department of Mathematics, Uppsala University, Box 480, 751 06 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 2 May 2013
Received in revised form 22 December 2013
Accepted 23 December 2013
Available online 3 January 2014

Keywords:
Gene expression data
Gene-set testing
High-dimensional data
Random subspace
Test about the mean
Two-sample problem

ABSTRACT

A common problem in genetics is that of testing whether a set of highly dependent gene expressions differ between two populations, typically in a high-dimensional setting where the data dimension is larger than the sample size. Most high-dimensional tests for the equality of two mean vectors rely on naive diagonal or trace estimators of the covariance matrix, ignoring dependences between variables. A test using random subspaces is proposed, which offers higher power when the variables are dependent and is invariant under linear transformations of the marginal distributions. The p-values for the test are obtained using permutations. The test does not rely on assumptions about normality or the structure of the covariance matrix. It is shown by simulation that the new test has higher power than competing tests in realistic settings motivated by microarray gene expression data. Computational aspects of high-dimensional permutation tests are also discussed and an efficient R implementation of the proposed test is provided.

1. Introduction

A commonly encountered problem in modern genetic research, geological imaging, signal processing, astrometry and finance is that of comparing the mean vectors of two populations. In many of today's applications, the data averts analysis by classic statistical methods as the data dimension p typically is larger than the sample size n, which causes most standard procedures to break down. When $p < n$, comparisons of this type are usually done using Hotelling's T^2 test. In the high-dimensional setting where $p \geq n$, the sample covariance matrix is not invertible, meaning that Hotelling's test no longer can be used. In this paper we discuss two-sample tests in a high-dimensional setting where the variables have a non-negligible dependence structure. While the tests are discussed in the context of gene-set testing, we stress that they are equally applicable to other fields in which high-dimensional data occur.

In genetic research, one is often interested in identifying differentially expressed genes between two groups of patients based on data from a microarray experiment. Genes, however, do not function in isolation. Rather, they work together in complex networks. It is therefore often of greater interest to search for sets of genes, rather than individual genes, that are differentially expressed (Barry et al., 2005; Efron and Tibshirani, 2007; Goeman and Bühlmann, 2007; Newton et al., 2007; Nettleton et al., 2008; Barry et al., 2008; Chen and Qin, 2010). These gene-sets are determined a priori, typically by utilizing databases such as Gene Ontology (http://www.geneontology.org/) or Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/) or by grouping genes with similar chromosomal locations together.

A common approach to finding differentially expressed gene-sets is to use a two-step procedure, starting by performing individual tests for each gene. These gene-level tests are then aggregated into a single test for the entire gene set. Much, if

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

[*] Tel.: +46 184713389.
E-mail address: thulin@math.uu.se.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.12.003

not all, of the multivariate structure of the data set is lost when gene-level tests are used. Goeman and Bühlmann (2007), Efron (2007) and Gatti et al. (2010) demonstrated several problems with common tests based on this approach, including very high rates of false positives. The empirical Bayes methods applied e.g. by Efron et al. (2001) suffer from similar issues, caused by correlations between gene expressions (Qui et al., 2005).

By using a truly multivariate test for the gene set, it is possible to not only take the multivariate dependence structure of the gene expressions into account, but to gain more power from these dependences, as illustrated in Fig. 1. There are three approaches to modifying the covariance estimator in Hotelling's test statistic to allow for high-dimensional inference. The first approach is to use prior information about the covariance structure to estimate the covariance matrix. Jacob et al. (2012) presented such a test in the setting where location shift between the two populations is related to a known graph structure describing the dependence between the genes.

The second approach is to assume that the covariance matrix has a simple diagonal structure, either explicitly when deriving the null distribution or implicitly by using only the diagonal of the covariance matrix in the test statistic. The tests proposed by Bai and Saranadasa (1996), Srivastava (2007), Srivastava and Du (2008) and Srivastava et al. (2013) follow this approach, effectively assuming the expressions of different genes to be independent. This is an unrealistic assumption for gene expressions, where genetic regulatory networks tend to cause the expression to be highly correlated. The assumption is equally unrealistic in many other biomedical problems. As a further example, the prevalence of an allele is typically highly correlated with the prevalence of other alleles on neighboring loci.

The third approach is to use an estimator that allows for dependence, but that can be used in the absence of prior information. The test proposed by Chen and Qin (2010) follows this approach. Recently, Lopes et al. (2011, 2012) proposed another such test in which the data is randomly pseudo-projected into several lower-dimensional spaces. Hotelling's T^2 statistic is computed for each pseudo-projection, and the result is then averaged over all pseudo-projections. Lopes et al. showed by asymptotic arguments and a simulation study that their test has substantially higher power than competing tests when the variables are correlated. There are however two downsides to their proposed method. First, it relies on an asymptotic null distribution derived under the assumption of normality. For finite sample sizes, this often leads to far too conservative p-values. Second, the test statistic is not invariant under linear transformations of the marginal distributions. This is a serious drawback, as it is common for genetic data to be rescaled by dividing the marginal distributions by their respective standard deviations.

In this paper, we show how accurate p-values for the Lopes et al. test can be obtained by using random permutations. We then propose a modified test statistic, which uses random subspaces instead of random pseudo-projections. The new test statistic is, conditioned on the random subspaces chosen, invariant under linear transformations of the marginal distributions.

A common problem in gene-set testing is that of identifying gene-sets, or pathways, that are related to cancer. For a pathway to induce cancer, a mutation must have occurred in at least one of its genes. Depending on where in this pathway the gene is located, the mutation can cause changes in the expressions of only a handful of genes or in all genes in the pathway. Motivated by this problem setting, we perform a Monte Carlo comparison of four multivariate two-sample tests and two tests based on gene-level t-tests. We compare the type I error rates and powers of the tests under different models for pathway dependences and mutation locations. Some tests that require normality are modified so that p-values are computed using permutations rather than asymptotic null distributions, resulting in better type I error rates as well as higher power. We also contrast the invariance properties of the test statistics. While invariance properties tend to be overlooked in the biomedical literature, they are of great importance in multivariate testing and need to be taken into account when choosing which test to use.

High-dimensional permutation tests are heavily computer-intensive. For that reason, we discuss some computational aspects of such tests, and show how to efficiently implement the proposed test in R.

Methods based on random projections and random subspaces have not been studied to a great extent in the statistical literature, but are common in machine learning, where these techniques mainly have been used for clustering and classification. See Durand and Atkison (2011) and Bingham and Mannila (2001) for reviews and some applications and Varmuza et al. (2010) for applications in chemometrics. Most authors have used only a single random projection or subspace, although there are a few exceptions, including the recent paper by Lopes et al. (2012). Cuesta-Albertos et al. (2006) used multiple random projections for goodness-of-fit testing but did not find the increase in power to be large enough to motivate the added computational complexity. Mielniczuka and Teisseyre (2014) used multiple random subsets for model selection, Fraiman and Svarc (2013) used random projections for robust estimation and Fern and Brodley (2003) used multiple random projections for clustering, in a manner that bears resemblance to the algorithms presented in this paper, and found that it improved the performance of their clustering algorithms. Recently, Henrion et al. (2011) proposed a subspace method for outlier detection in high-dimensional data sets that is somewhat similar to the random subspaces test presented in the present paper. Their method differs from ours in several ways, the most important difference being that their goal is to give an anomaly score *to each observation*, rather than to compute a statistic that can be used for inference about the underlying population. Also worth mentioning is a recent paper by Wei et al. (2013), who described a general hypothesis testing framework based on a non-random projection to the real line, determined by a linear classifier.

The rest of the paper is organized as follows. In Section 2 we review the Lopes et al. test and discuss its drawbacks. In Section 3 we propose a new test based on random subspaces. In Section 4 we compare several gene-set tests in terms of invariance, type I error rates and power. The new test is applied to a breast cancer data set in Section 5. In Section 6 we