Full PDF
DMPS Page

Discussiones Mathematicae
Probability and Statistics 37 (2017) 147–148
doi:10.7151/dmps.1195

A TWO-SAMPLE TEST BASED ON CLUSTER SUBSPACES FOR EQUALITY OF MEAN VECTORS IN HIGH DIMENSION

Łukasz Smaga

Adam Mickiewicz University
Faculty of Mathematics and Computer Science

e-mail: ls@amu.edu.pl

Abstract

In this paper, a two-sample problem in a high-dimensional setting, where the data dimension is larger than the sample size, is considered. In such setting, the Hotelling's test is not applicable due to singularity of the pooled sample covariance matrix. Recently, Zhang and Pan (2016) proposed a permutation test based on several cluster subspaces of lower dimension, where the Hotelling's statistic can be applied. This paper considers a modification of this test using other dissimilarity measure. To calculate clusters, a cutoff measure is established. The new testing procedure is shown to be invariant under linear transformations of the marginal distributions. Simulation studies indicate that the new test performs comparable to or even better in certain situations than the test of Zhang and Pan (2016) in terms of power.

Keywords: cluster analysis, coefficient of determination, high-dimensional data, Pearson correlation coefficient, two-sample problem.

2010 Mathematics Subject Classification: 62H15, 62H30.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis (3rd ed., Wiley, London, 2003).

[2] Z. Bai and H. Saranadasa, *Effect of high dimension: by an example of a two sample problem*, Statist. Sinica **6** (1996) 311–329.

[3] S.X. Chen and Y.L. Qin, *A two-sample test for high-dimensional data with applications to gene-set testing*, Ann. Stat. **38** (2010) 808–835.

[4] L. Feng, C. Zou, Z. Wang and L. Zhu, *Two sample Behrens-Fisher problem for high-dimensional data*, Statist. Sinica **25** (2015) 1297–1312.

[5] L. Feng, C. Zou, Z. Wang and L. Zhu, *Composite T^2 test for high-dimensional data*, Statist. Sinica (2016).
doi:10.5705/ss.202015.0199

[6] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria (2017).
https://www.R-project.org/

[7] Ł. Smaga, *Diagonal and unscaled Wald-type tests in general factorial designs*, Electr. J. Stat. **11** (2017) 2613–2646.

[8] M.S. Srivastava and M. Du, *A test for the mean vector with fewer observations than the dimension*, J. Multivariate Anal. **99** (2008) 386–402.

[9] M. Thulin, *A high-dimensional two-sample test for the mean using random subspaces*, Comput. Stat. & Data Anal. **74** (2014) 26–38.

[10] J. Zhang and M. Pan, *A high-dimension two-sample test for the mean using cluster subspaces*, Comput. Stat. & Data Anal. **97** (2016) 87–97.

Received 5 October 2017
Accepted 16 November 2017