Statistica Sinica **34** (2024), 1-18
doi:https://doi.org/10.5705/ss.2202022.0048

HOMOGENEITY TESTS FOR HIGH-DIMENSIONAL MEAN VECTORS AND COVARIANCE MATRICES

Wenwen Guo[1], Xinyuan Song*[2] and Hengjian Cui[1]

[1] *Capital Normal University and* [2] *The Chinese University of Hong Kong*

Abstract: This study aims to develop homogeneity tests for high-dimensional mean vectors and covariance matrices, in which the number of features may be greater than the sample size. We introduce two categorically weighted statistics to test the equality of means and of covariance matrices. We establish the asymptotic distributions of the proposed test statistics under certain mild conditions, and develop simplified algorithms to facilitate the implementation and application. Simulation studies demonstrate the satisfactory performance of the proposed tests in terms of the empirical size and power. We also apply the proposed test procedures to two microarray data sets.

Key words and phrases: High-dimension, homogeneity, K-sample problem, location and scale, MANOVA.

1. Introduction

Despite numerous studies on homogeneity tests for distributions or distribution features (mean vectors or covariance matrices) in different populations, a crucial remaining problem is establishing whether gene expression levels differ among predefined patient populations in order to identify a disease's capital causal gene. However, in modern biological and financial studies, the data dimension is often much larger than the sample size. This "large p, small n" paradigm poses a considerable challenge to classical homogeneity tests, which were originally designed for fixed-dimensional problems.

This study focuses on homogeneity tests for high-dimensional mean vectors and covariance matrices. Assume that homogeneity tests for means, consider R groups. When $R = 2$, the traditional Hotelling T^2 test is optimal for normally distributed data when p is fixed. Several extensions of the Hotelling T^2 test have been proposed to accommodate high dimensionality; examples include those of Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010), Cai, Liu and Xia (2013), Feng, Zou and Wang (2016), and Chang et al. (2017). When $R > 2$, researchers often use a multivariate analysis of variance (MANOVA) to investigate whether the population mean vectors are the same under the "large n, small p" paradigm. Cai and Xia (2014) test the equality of multiple high-dimensional sparse mean vectors under dependency. Recently, Hu et al. (2017)

*Corresponding author.