Journal of Multivariate Analysis 162 (2017) 82–92

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Tests for the parallelism and flatness hypotheses of multi-group profile analysis for high-dimensional elliptical populations

Masashi Hyodo

Department of Mathematical Sciences, Graduate School of Engineering, Osaka Prefecture University, 1-1, Gakuen-cho, Naka-ku, Sakai-shi, Osaka 599-8531, Japan

A R T I C L E I N F O

Article history:
Received 8 September 2017
Available online 18 September 2017

AMS 2000 subject classifications:
primary 62H15
secondary 62F05

Keywords:
Elliptical population
High dimension
Profile analysis
Statistical hypothesis testing

A B S T R A C T

This paper is concerned with tests for the parallelism and flatness hypotheses in multi-group profile analysis for high-dimensional data. We extend to elliptical distributions the procedures developed for normal populations by Harrar and Kong (2016). Specifically, we prove that their statistics continue to be asymptotically normal when the underlying population is elliptical, and we obtain new tests by improving their estimator of the asymptotic variance. Using asymptotic normality, we show that the asymptotic size of the proposed tests is equal to the nominal significance level, and we also derive the asymptotic power. Finally, we present simulation results and find that the power of the new tests is superior to that of the original Harrar–Kong procedure.

1. Introduction

We consider a multi-sample testing problem for profile analysis for populations with elliptically contoured distributions. For group $g \in \{1, \ldots, a\}$, let $\boldsymbol{\mu}_g = (\mu_{g1}, \ldots, \mu_{gp})^\top$ be a p-dimensional real vector, Λ_g be a $p \times p$ nonnegative definite matrix, and ξ_g be a nonnegative function. The $p \times 1$ random vector $\boldsymbol{X}_g$ is said to have an elliptically contoured distribution, denoted $\boldsymbol{X}_g \sim \mathcal{C}_p(\xi_g, \boldsymbol{\mu}_g, \Lambda_g)$, if the characteristic function of $\boldsymbol{X}_g$ can be written, for any $\boldsymbol{t} \in \mathbb{R}^p$, as

$$\phi_g(\boldsymbol{t}) = e^{i\boldsymbol{t}^\top \boldsymbol{\mu}_g} \xi_g(\boldsymbol{t}^\top \Lambda_g \boldsymbol{t}).$$

As a result, $\mathrm{E}(\boldsymbol{X}_g) = \boldsymbol{\mu}_g$ and $\mathrm{var}(\boldsymbol{X}_g) = -2\xi_g'(0)\Lambda_g \equiv \Sigma_g$, respectively. Well-known examples of elliptical distributions include the multivariate normal, multivariate Student t, and contaminated normal distributions; see, e.g., Muirhead [8].

Let $\boldsymbol{X}_{g1}, \ldots, \boldsymbol{X}_{gn_g}$ be mutually independent copies of $\boldsymbol{X}_g$. We consider a test of the parallelism hypothesis

$$\mathcal{H}_{01} : \forall_{g\in\{1,\ldots,a-1\}} \;\; \boldsymbol{\mu}_g - \boldsymbol{\mu}_a = \gamma_g \mathbf{1}_p \quad \text{vs.} \quad \mathcal{A}_{01} : \neg\mathcal{H}_{01}. \tag{1}$$

Here, γ_g is an unknown real constant and $\mathbf{1}_p$ is a $p \times 1$ vector of 1's, i.e., $\mathbf{1}_p = (1, \ldots, 1)^\top$. We also consider tests of the flatness hypothesis

$$\mathcal{H}_{02} : \forall_{g\in\{1,\ldots,a\}} \;\; \mu_{g1} = \cdots = \mu_{gp} \quad \text{vs.} \quad \mathcal{A}_{02} : \neg\mathcal{H}_{02}, \tag{2}$$

E-mail address: hyodo@ms.osakafu-u.ac.jp.

http://dx.doi.org/10.1016/j.jmva.2017.09.004

and the level hypothesis

$$\mathcal{H}_{03} : \gamma_1 = \cdots = \gamma_{a-1} = 0 \quad \text{vs.} \quad \mathcal{A}_{03} : \neg\mathcal{H}_{03}. \tag{3}$$

Harrar and Kong [3] give expressions that are equivalent to hypotheses (1)–(3). Expression (1) is equivalent to

$$\widetilde{\mathcal{H}}_{01} : \boldsymbol{\mu}^\top K_{01}\boldsymbol{\mu} = 0 \quad \text{vs.} \quad \widetilde{\mathcal{A}}_{01} : \boldsymbol{\mu}^\top K_{01}\boldsymbol{\mu} > 0$$

with $K_{01} = P_a \otimes P_p$, where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_a^\top)^\top$ and $P_k = I_k - k^{-1}\mathbf{1}_k\mathbf{1}_k^\top$ for $k \in \{a, p\}$. Setting $K_{02} = (a^{-1}\mathbf{1}_a\mathbf{1}_a^\top) \otimes P_p$ and $K_{03} = D_a \otimes (p^{-1}\mathbf{1}_p\mathbf{1}_p^\top)$, we can also express hypotheses $\widetilde{\mathcal{H}}_{0x}$ with $x \in \{2, 3\}$ in the form

$$\widetilde{\mathcal{H}}_{0x} : \boldsymbol{\mu}^\top K_{0x}\boldsymbol{\mu} = 0 \quad \text{vs.} \quad \widetilde{\mathcal{A}}_{0x} : \boldsymbol{\mu}^\top K_{0x}\boldsymbol{\mu} > 0,$$

where $D_a = \text{diag}(n_1, \ldots, n_a) - n_{(a)}^{-1}\boldsymbol{n}\boldsymbol{n}^\top$. Here, $\boldsymbol{n} = (n_1, \ldots, n_a)^\top$ and $n_{(a)} = n_1 + \cdots + n_a$.

Srivastava [11] derived the likelihood ratio test for hypotheses (1)–(3) for two normal populations. However, the likelihood ratio tests for (1) and (2) cannot be applied to situations where $n_{(a)} \ll p$, e.g., microarray data, even for normal populations with covariance homogeneity.

In profile analysis, Takahashi and Shutoh [12] considered approximate tests for hypotheses (1) and (2) for two normal populations with equal covariance matrices. Harrar and Kong [3] extended these tests to multi-group normal populations without assuming equal covariance matrices. They also obtained the approximate test for hypothesis (3) based on matching moments.

In parallel, the effect of non-normality in profile analysis has been investigated. Okamoto et al. [9] used a perturbation method to obtain the asymptotic expansions of the distributions of test statistics for elliptical populations. Maruyama [6] extended the results under more general conditions using a different method introduced by Kano [5]. Note that these results are derived as $n_{(a)} \to \infty$.

In this paper, we propose new approximate tests for hypotheses (1) and (2) for high-dimensional elliptical populations without assuming equal covariance matrices. We note that the rank of K_{03} is at most $a - 1$, i.e., it does not grow with p; accordingly, $(n_{(a)}, p)$ asymptotic considerations are not relevant in pursuing our primary interest, which is to test (1) and (2). To this end, we show that the asymptotic normality of the test statistics proposed by Harrar and Kong [3] holds when the underlying distribution is elliptical. An improved estimator of the asymptotic variance of these test statistics also enables us to propose new approximate tests for (1) and (2) for high-dimensional elliptical populations.

The remainder of this paper is organized as follows. Preliminary asymptotic results for approximate tests are presented in Section 2. Using these results, we construct approximate tests for (1) and (2) and derive the asymptotic power and size of these tests for elliptical populations in Section 3. In Section 4, the numerical accuracy of the proposed tests is investigated, and the results are illustrated with a short numerical example. Section 5 concludes the paper. Technical proofs are given in Appendix.

2. Preliminary asymptotic results

We define an $a \times a$ non-random matrix

$$(R_a)_{ij} = \begin{cases} d_i & \text{if } i = j, \\ \psi\delta_i\delta_j & \text{if } i \neq j, \end{cases}$$

with $d_i, \delta_i, \psi \in \mathbb{R}$ for $i, j \in \{1, \ldots, a\}$. Then we consider the random variable

$$T = \overline{\boldsymbol{X}}^\top (R_a \otimes P_p)\overline{\boldsymbol{X}} - \sum_{g=1}^{a} \frac{d_g \text{tr}(P_p S_g)}{n_g},$$

where $\overline{\boldsymbol{X}} = (\overline{\boldsymbol{X}}_1^\top, \ldots, \overline{\boldsymbol{X}}_a^\top)^\top$ and

$$S_g = \frac{1}{n_g - 1}\sum_{i=1}^{n_g}(\boldsymbol{X}_{gi} - \overline{\boldsymbol{X}}_g)(\boldsymbol{X}_{gi} - \overline{\boldsymbol{X}}_g)^\top$$

with $\overline{\boldsymbol{X}}_g = (\boldsymbol{X}_{g1} + \cdots + \boldsymbol{X}_{gn_g})/n_g$.

Remark 1. If $\psi = -1$, $d_i = 1 - 1/a$, and $\delta_i = 1/\sqrt{a}$ for all $i \in \{1, \ldots, a\}$, then T is the test statistic for $\mathcal{H}_{01}$. If $\psi = d_i = a^{-1}$ and $\delta_i = 1$ for all $i \in \{1, \ldots, a\}$, then T is the test statistic for $\mathcal{H}_{02}$.

Here, T is an unbiased estimator of $\boldsymbol{\mu}^\top (R_a \otimes P_p)\boldsymbol{\mu}$, i.e.,

$$\text{E}(T) = \boldsymbol{\mu}^\top (R_a \otimes P_p)\boldsymbol{\mu}. \tag{4}$$