Journal of Multivariate Analysis 115 (2013) 204–216

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Tests for multivariate analysis of variance in high dimension under non-normality

Muni S. Srivastava [a], Tatsuya Kubokawa [b,*]

[a] Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario, Canada M5S 3G3
[b] Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

ARTICLE INFO

Article history:
Received 27 December 2011
Available online 23 October 2012

AMS 2010 subject classifications:
primary 62H15
secondary 62F05

Keywords:
Asymptotic distributions
High dimension
MANOVA
Multivariate linear model
Non-normal model
Sample size smaller than dimension

ABSTRACT

In this article, we consider the problem of testing the equality of mean vectors of dimension p of several groups with a common unknown non-singular covariance matrix $\boldsymbol{\Sigma}$, based on N independent observation vectors where N may be less than the dimension p. This problem, known in the literature as the multivariate analysis of variance (MANOVA) in high-dimension has recently been considered in the statistical literature by Srivastava and Fujikoshi (2006) [8], Srivastava (2007) [5] and Schott (2007) [3]. All these tests are not invariant under the change of units of measurements. On the lines of Srivastava and Du (2008) [7] and Srivastava (2009) [6], we propose a test that has the above invariance property. The null and the non-null distributions are derived under the assumption that $(N, p) \to \infty$ and N may be less than p and the observation vectors follow a general non-normal model.

1. Introduction

The problem of testing the equality of mean vectors of several groups with common unknown nonsingular covariance matrix, the so called MANOVA or multivariate analysis of variance has been considered many times in the statistical literature. For normally distributed observation vectors when the total sample size N is considerably larger than the dimension p of the vector, Wilks [9] likelihood ratio test is commonly used with Box's [2] approximation for the distribution of the test statistic. For dimension p larger than the sample size N, this testing problem has also been recently considered in the literature by Srivastava and Fujikoshi [8], Srivastava [5], Schott [3] and Yamada and Srivastava [10] for normally distributed observation vectors.

In this article, we consider a general model which includes normal distributions and propose a test that is invariant under the change of units of measurements. That is, the test statistic is invariant under the transformation by non singular diagonal matrices. Thus, without any loss of generality, we assume that the covariance matrix is a correlation matrix $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$, where $\boldsymbol{\Lambda}^{1/2}$ is the unique positive definite matrix. Since the MANOVA problem is a special case of the multivariate regression model, we assume that the $N \times p$ matrix of observations follow the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta} + \boldsymbol{U}\boldsymbol{\Lambda}^{1/2} \tag{1.1}$$

where $\boldsymbol{X}$ is an $N \times k$ matrix of known constants of rank k, $\boldsymbol{\Theta}$ is a $k \times p$ matrix of unknown parameters, $k \leq p$,

$$\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)',$$

[*] Corresponding author.
E-mail addresses: srivasta@utstat.toronto.edu (M.S. Srivastava), tatsuya@e.u-tokyo.ac.jp (T. Kubokawa).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.10.011

and $\boldsymbol{u}_i = (u_{i1}, \dots, u_{ip})'$ are independent and identically distributed with

$$E(\boldsymbol{u}_i) = \boldsymbol{0}, \qquad \mathbf{Cov}\,(\boldsymbol{u}_i) = \boldsymbol{I}_p, \qquad E(u_{ik}^4) = K_4 + 3, \tag{1.2}$$

and for $\nu_k \geq 0$, $\sum_{k=1}^{p} \nu_k \leq 4,\ i = 1, \dots, N$,

$$E\left[\prod_{k=1}^{p} u_{ik}^{\nu_k}\right] = \prod_{k=1}^{p} E(u_{ik}^{\nu_k}). \tag{1.3}$$

Here $\boldsymbol{\Lambda} = (\lambda_{ij}) = \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$ is the non-singular correlation matrix. For normally distributed $\boldsymbol{u}_i$ with zero mean vector and identity covariance matrix, the conditions (1.2)–(1.3) are satisfied with $K_4 = 0$.

The problem of testing in the model (1.1) is that of testing the hypothesis

$$H : \boldsymbol{C\Theta} = \boldsymbol{0} \quad \text{vs.} \quad A : \boldsymbol{C\Theta} \neq \boldsymbol{0},$$

where $\boldsymbol{C}$ is a $q \times k$ known matrix of rank $q \leq k$. For example, in testing the equality of $k = (q + 1)$ mean vectors, the observation matrix $\boldsymbol{Y}$ is of the form given by

$$\boldsymbol{Y} = (\boldsymbol{y}_{11}, \dots, \boldsymbol{y}_{1N_1}; \dots; \boldsymbol{y}_{k1}, \dots, \boldsymbol{y}_{kN_k})', \tag{1.4}$$

where N_i independent vectors are obtained from the ith group with mean vector $\boldsymbol{\mu}_i$, $i = 1, \dots, q + 1$, and $N = N_1 + \dots + N_{q+1}$. All the observation vectors have the same covariance matrix which we have assumed in this article as non singular correlation matrix $\boldsymbol{\Lambda}$. To write the problem of testing the equality of $k = (q + 1)$ mean vectors as a regression model, we define a vector $\boldsymbol{1}_r = (1, \dots, 1)'$ as an r-vector with all the elements equal to one,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{1}_{N_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1}_{N_2} & \boldsymbol{0} \\ \vdots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{1}_{N_k} \end{pmatrix} : N \times k \tag{1.5}$$

and

$$\boldsymbol{\Theta} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k)' : k \times p, \quad k = q + 1. \tag{1.6}$$

Thus, the regression model representing the mean vectors of $k = (q + 1)$ groups is given by (1.1) with $\boldsymbol{Y}, \boldsymbol{X}$ and $\boldsymbol{\Theta}$ defined respectively in (1.4)–(1.6). The problem of testing the equality of $k = (q + 1)$ mean vectors is given by $H : \boldsymbol{C\Theta} = \boldsymbol{0}$ against the alternative $A : \boldsymbol{C\Theta} \neq \boldsymbol{0}$ where $\boldsymbol{C}$ is now given by $q \times (q + 1)$ matrix.

$$\boldsymbol{C} = (\boldsymbol{I}_q, -\boldsymbol{1}_q) : q \times k, \quad k = q + 1. \tag{1.7}$$

In general, for testing the hypothesis $H : \boldsymbol{C\Theta} = \boldsymbol{0}$, we consider the variation due to the hypothesis given by

$$\boldsymbol{B} = \boldsymbol{Y}'\boldsymbol{G}\boldsymbol{Y}, \tag{1.8}$$

where

$$\boldsymbol{G} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1}\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \tag{1.9}$$

is an $N \times N$ matrix of rank $q < N$. The matrix $\boldsymbol{G}$ is an idempotent matrix of rank q, $\boldsymbol{G}^m = \boldsymbol{G}$ for a positive integer m. That is, there are q eigenvalues that are equal to 1 and the remaining $N - q$ eigenvalues are zero.

For asymptotic results for regression models under non-normal distributions, some assumptions on the so-called design matrix $\boldsymbol{X} = (x_{ij})$ are made. For example, it is common to assume that $N^{-1}\boldsymbol{X}'\boldsymbol{X}$ goes to a positive definite matrix and that x_{ij}'s are uniformly bounded. In our case, we assume that $\boldsymbol{G} = (g_{ij})$, $g_{ij} = O(N^{-1})$. This gives $\sum_{i=1}^{N}\sum_{j=1}^{N} g_{ij}^2 = O(1)$, which in our case is $q < \infty$. This also gives $\sum_{i=1}^{N} g_{ii}^2 = O(N^{-1})$. The above assumption along with assumptions on the correlation matrix are stated below.

Assumption (A). A(1) For $\boldsymbol{G} = (g_{ij})$, $g_{ij} = O(N^{-1})$,
A(2) $\lim_{p\to\infty}(\mathrm{tr}\,[\boldsymbol{\Lambda}^2]/p) < \infty$,
A(3) $\lim_{p\to\infty}(\mathrm{tr}\,[\boldsymbol{\Lambda}^4]/p^2) = 0$,
A(4) $N = O(p^\delta)$, $\delta > 1/2$, $q < \infty$,
A(5) $\lim_{(n,p)\to\infty}\{(pq)^{-1}\mathrm{tr}\,[\boldsymbol{\Lambda}\boldsymbol{M}\boldsymbol{M}']\} = 0$,

where

$$\boldsymbol{M} = \boldsymbol{\Theta}'\boldsymbol{C}'[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}']^{-1/2}, \qquad \boldsymbol{G}_+ = (g_{ij+}), \tag{1.10}$$

and $g_{ij+} = |g_{ij}|,\ i \neq j,\ i, j = 1, \dots, N,\ g_{ii} \geq 0$.