University *of*
New Haven

University of New Haven
Digital Commons @ New Haven

Marketing and Quantitative Analysis Faculty
Publications

Marketing and Quantitative Analysis

Generalized F-tests for the Multivariate Normal Mean

Jiajuan (John) Liang
University of New Haven, Jliang@newhaven.edu

Man-Lai Tang
Hang Seng Management College

Follow this and additional works at: http://digitalcommons.newhaven.edu/
marketingandquantitativeanalysis-facpubs

Part of the Marketing Commons

Publisher Citation

Comments

Generalized *F*-tests for the Multivariate Normal Mean[1]

Jiajuan Liang[a,2], Man-Lai Tang[b]

[a]*Department of Management, University of New Haven*

300 Boston Post Road, West Haven, CT 06516

[b]*Department of Mathematics, Hong Kong Baptist University*

Kowloon Tong, Hong Kong, China

Abstract

Based on Läuter's (Biometrics, 1996) exact *t* test for biometrical studies related to the multivariate normal mean, we develop a generalized *F*-test for the multivariate normal mean and extend it to multiple comparison. The proposed generalized *F*-tests have simple approximate null distributions. A Monte Carlo study and two real examples show that the generalized *F*-test is at least as good as the optional individual Läuter's test and can improve its performance in some situations where the projection directions for the Läuter's test may not be suitably chosen. It is discussed that the generalized *F*-test could be superior to individual Läuter's tests and the classical Hotelling T^2-test for the general purpose of testing the multivariate normal mean. It is shown by Monte Carlo studies that the extended generalized *F*-test outperforms the commonly-used classical test for multiple comparison of normal means in the case of high dimension with small sample sizes.

AMS Classification: 62F03; 62F05

Keywords: F-test; Multiple comparison; Normal distribution; Principal components; Robustness; Spherical distribution.

---

1. Introduction

It is well known that the Hotelling T^2-test is a commonly-used method for testing the multivariate normal mean (Anderson, 1984). Theoretically, the null hypothesis for such a test can be reduced to

$$H_0: \quad \boldsymbol{\mu} = \mathbf{0} \qquad (1.1)$$

versus H_1: $\boldsymbol{\mu} \neq \mathbf{0}$ based on an i.i.d. sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is unknown and assumed to be positively definite ($\boldsymbol{\Sigma} > \mathbf{0}$). The classical Hotelling T^2-test is equivalent to an exact F-test (Anderson, 1984) and is based on the condition that the sample size n must be greater than the dimension p (i.e., $n > p$) so that the sample covariance matrix is nonsingular. This condition, however, may not be satisfied in some situations where experimental data are costly obtained. For example, in a longitudinal study that may last for months or even years, the available number of experimental subjects is often limited or costly to obtain but the number of time points for observing the experimental subjects is usually large. In this case, the number of time points is the dimension of the observed data and the number of experimental subjects is the sample size. We often face analysis of high dimensional data with small sample sizes in longitudinal studies or other medical research.

In their biometrical studies on treatment effects with multiple endpoints, Läuter (1996) and Läuter et al. (1996) developed some exact t- and F-tests. We will call these tests the Läuter's tests in the subsequent discussion. The remarkable advantage of Läuter's tests is that they are applicable to both cases of large and small sample sizes and even the case of $n \leq p$. This is a noteworthy improvement over the classical Hotelling T^2-test. Läuter's tests maintain the exact α-level for any pre-assigned significance level α (e.g., 1%, 5% and 10%) and possess some dimension stability. This implies that the tests are applicable for both $n > p$ and $n \leq p$. This kind of dimension stability can

also be found from the goodness-of-fit tests developed by Fang et al. (1998), Liang and Fang (2000), and Liang et al. (2000).

In this paper, we will propose a generalized F-test for a general purpose of testing the multivariate normal mean as in hypothesis (1.1) and extend it to multiple comparison of normal means. The proposed test can improve the power performance of Läuter's tests in some situations and possesses the same dimension stability property. This paper is organized as follows. Section 2 presents the theoretical outline for deriving the generalized F-test. Section 3 extends the results in Section 2 to multiple comparison of population means. A Monte Carlo study is given in Section 4. Applications of the proposed generalized F-test are illustrated by two real examples in Section 5. Some remarks are given in the last section.

2. The Generalized F-Test

Läuter's tests are based on the following Lemma (Läuter, 1996).

Lemma 1. *Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ $(n \times p,\ n > p)$ be the observation matrix of an i.i.d. sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from the p-dimensional normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $(\boldsymbol{\Sigma} > \boldsymbol{0})$ and $\boldsymbol{d} = f(\boldsymbol{X}'\boldsymbol{X})$ be a $p \times 1$ random vector that is a function of $\boldsymbol{X}'\boldsymbol{X}$ and is uniquely determined by $\boldsymbol{X}'\boldsymbol{X}$. Define the random vector*

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{d}. \tag{2.1}$$

Then $\boldsymbol{z}$ has a spherical distribution (Fang et al., 1990) with $P(\boldsymbol{z} = \boldsymbol{0}) = 0$.

Läuter (1996) proposed the following exact t-test for hypothesis (1.1):

$$T = \sqrt{n}\bar{z} \Big/ \left[\frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \right]^{\frac{1}{2}}, \tag{2.2}$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)'$ is obtained from (2.1) and $\bar{z} = \sum_{i=1}^{n} z_i/n$. Under the assumption in Lemma 1, T has an exact t-distribution $t(n-1)$ with $(n-1)$ degrees of freedom. Note that T in (2.2) does not depend on the dimension p. That explains why T has the property of dimension stability. In other words,

3

T is not sensitive to the increase of dimension and it is applicable to the case of $n \leq p$.

The choice of $\boldsymbol{d} = f(\boldsymbol{X}'\boldsymbol{X})$ in Lemma 1 is an interesting issue that deserves further study. Briefly, it acts as a projection direction for projecting the multivariate data in $\boldsymbol{X}$ onto the direction $\boldsymbol{d}$. Läuter's t-test in (2.2) is constructed from the one-dimensional "observations" $z_1, \ldots, z_n$ in $\boldsymbol{z} = (z_1, \ldots, z_n)'$ determined by (2.1). $z_1, \ldots, z_n$ may not be independent but have the same distribution (Fang et al., 1990). The spherical property of $\boldsymbol{z}$ in (2.1) guarantees T in (2.2) has an exact t-distribution $t(n-1)$. Theoretically, any direction $\boldsymbol{d} = f(\boldsymbol{X}'\boldsymbol{X})$ in Lemma 1 can lead to an exact t-test given by (2.2). Läuter et al. (1996) gave a discussion on some meaningful choices of $\boldsymbol{d}$ in Lemma 1. There is no optimal choice available in the literature. However, the principal component (PC) direction was found to have good performance for general purposes. This was also empirically verified by Liang et al. (2000).

For a general purpose without considering the directional interpretation of rejecting the null hypothesis (1.1), the exact t-test (2.2) is equivalent to the F-test given by

$$LF = T^2 = n\bar{z}^2 \Big/ \left[\frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2 \right]. \tag{2.3}$$

LF has an F-distribution $F(1, n-1)$ under Lemma 1. We will call a test for (1.1) based on (2.3) the Läuter's F-test in the subsequent context. Following the discussion in Läuter et al. (1996) on the PC directions, we define the p sample PC directions from the observation matrix $\boldsymbol{X}$ in Lemma 1 by

$$\left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} \right) \boldsymbol{D} = \boldsymbol{D}\boldsymbol{\Lambda}, \tag{2.4}$$

where $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p)'$ $(p \times p)$ consists of the p PC directions $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ (a diagonal matrix) with $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$. We impose the condition on the matrix $\boldsymbol{D}$ so that $\boldsymbol{D}$ has positive diagonal elements for the eigenvectors associated with the positive eigenvalues to ensure the unique

solution to the eigenvalue-eigenvector problem (2.4). The population PC directions $\{\widetilde{\boldsymbol{d}}_1, \ldots, \widetilde{\boldsymbol{d}}_p\}$ are defined by the solutions to the eigenvalue-eigenvector problem

$$\boldsymbol{\Sigma}\, \widetilde{\boldsymbol{D}} = \widetilde{\boldsymbol{D}}\, \widetilde{\boldsymbol{\Lambda}} \tag{2.5}$$

where $\widetilde{\boldsymbol{D}} = (\widetilde{\boldsymbol{d}}_1, \ldots, \widetilde{\boldsymbol{d}}_p)'$ $(p \times p)$, and $\widetilde{\boldsymbol{\Lambda}} = \operatorname{diag}(\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_p)$ with $\widetilde{\lambda}_1 \geq \ldots \geq \widetilde{\lambda}_p > 0$. $\widetilde{\boldsymbol{D}}$ is assumed to have positive diagonal elements to ensure the unique solution. The following theorem is the basis for constructing the generalized F-test.

Theorem 1. *Let $\{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p\}$ be the sample PC directions given by the solutions to eigenvalue-eigenvector problem (2.4) and let the random vectors*

$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{d}_1, \ \ldots, \ \boldsymbol{z}_p = \boldsymbol{X}\boldsymbol{d}_p, \tag{2.6}$$

where $\boldsymbol{X}$ $(n \times p)$ is observation matrix as in Lemma 1 and (2.4). Assuming the null hypothesis (1.1) is true and $n > p$, we have the following two assertions.

(1) The random vector $\boldsymbol{z}_i$ $(i = 1, \ldots, p)$ has a spherical distribution with
   $P(\boldsymbol{z}_i = \boldsymbol{0}) = 0$;

(2) $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ are asymptotically independent when $n \to \infty$.

Proof. *For assertion (1), it is noted from (2.4) that each eigenvector $\boldsymbol{d}_i$ is a function of $\boldsymbol{X}'\boldsymbol{X}$. That is, we can write $\boldsymbol{d}_i = f_i(\boldsymbol{X}'\boldsymbol{X})$. Assertion 1 follows from Lemma 1.*

For assertion (2), under the null hypothesis (1.1), the random matrix $\boldsymbol{X}'\boldsymbol{X}$ has a Wishart distribution $W(\boldsymbol{\Sigma}, n)$ (see Mardia et al. (1979, p. 66)). According to Theorem 8.3.3 of Mardia et al. (1979, p. 230), the sample PC direction $\boldsymbol{d}_i \sim N_p(\widetilde{\boldsymbol{d}}_i, \boldsymbol{V}_i/n)$ (approximately, for some covariance matrix $\boldsymbol{V}_i$ determined by $\boldsymbol{\Sigma}$), where $\widetilde{\boldsymbol{d}}_i$ is the population PC direction determined by (2.5), we can write

$$n^{1/2}(\boldsymbol{d}_i - \widetilde{\boldsymbol{d}}_i) = O_p(1), \quad or \quad \boldsymbol{d}_i = \widetilde{\boldsymbol{d}}_i + \boldsymbol{g}_n, \quad \boldsymbol{g}_n = O_p(n^{-1/2}), \tag{2.7}$$

where "$O_p(n^{-1/2})$" and "$O_p(1)$" are the common notations used in the limit theory of probability (see, e.g., Mardia et al. (1979, p. 52)). $\boldsymbol{g}_n = O_p(n^{-1/2})$

and $\boldsymbol{b}_n = O_p(1)$ are equivalent to

$$
\begin{aligned}
&\lim_{k\to\infty} \sup_{n\geq k} P\left(\|n^{1/2}\boldsymbol{g}_n\| > c\right) \to 0 \qquad as\ c \to \infty, \\
and \quad &\lim_{k\to\infty} \sup_{n\geq k} P\left(\|\boldsymbol{b}_n\| > c\right) \to 0 \qquad as\ c \to \infty,
\end{aligned}
\tag{2.8}
$$

respectively, where "$\|\cdot\|$" stands for the Euclidean norm of a vector. From the definition of $\boldsymbol{z}_i$ in (2.6), we have

$$
\boldsymbol{z}_i = \boldsymbol{X}\boldsymbol{d}_i = \boldsymbol{X}\,\widetilde{\boldsymbol{d}}_i + \boldsymbol{X}\boldsymbol{g}_n.
\tag{2.9}
$$

Since $\boldsymbol{X} \sim N_{n\times p}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$ under hypothesis (1.1), it is easy to verify that

$$
\boldsymbol{X}\,\widetilde{\boldsymbol{d}}_i \sim N_n(\boldsymbol{0}, \widetilde{\lambda}_i\,\boldsymbol{I}_n),
\tag{2.10}
$$

and $\{\boldsymbol{X}\,\widetilde{\boldsymbol{d}}_i:\ i = 1,\ldots,p\}$ are independent due to

$$
\begin{aligned}
cov(\boldsymbol{X}\,\widetilde{\boldsymbol{d}}_i, \boldsymbol{X}\,\widetilde{\boldsymbol{d}}_j) &= E\left\{vec(\boldsymbol{X}\,\widetilde{\boldsymbol{d}}_i)'[vec(\boldsymbol{X}\,\widetilde{\boldsymbol{d}}_j)']'\right\} \\
&= E\left\{(\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{d}}\,'_i)vec\boldsymbol{X}'[vec\boldsymbol{X}']'(\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{d}}_j)\right\} \\
&= (\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{d}}\,'_i)E\left\{vec\boldsymbol{X}'[vec\boldsymbol{X}']'\right\}(\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{d}}_j) \\
&= (\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{d}}\,'_i)(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma})(\boldsymbol{I}_n \otimes \widetilde{\boldsymbol{d}}_j) \\
&= \boldsymbol{I}_n \otimes (\widetilde{\boldsymbol{d}}\,'_i\boldsymbol{\Sigma}\,\widetilde{\boldsymbol{d}}_j) = \delta_{ij}\,\widetilde{\lambda}_i\,\boldsymbol{I}_m,
\end{aligned}
\tag{2.11}
$$

where $\widetilde{\lambda}_i$'s $(i = 1\ldots,p)$ in (2.10) and (2.11) are the eigenvalues of $\boldsymbol{\Sigma}$ in (2.5) with $\widetilde{\lambda}_1 \geq \ldots \geq \widetilde{\lambda}_p > 0$, $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$. Since $\boldsymbol{g}_n = O_p(n^{-1/2})$, it can be verified that

$$
\begin{aligned}
\|\boldsymbol{X}\boldsymbol{g}_n\|^2 &= \boldsymbol{g}'_n\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}_n = n\boldsymbol{g}'_n(\boldsymbol{X}'\boldsymbol{X}/n)\boldsymbol{g}_n \\
&\leq n\lambda_1\boldsymbol{g}'_n\boldsymbol{g}_n = n\lambda_1\|\boldsymbol{g}_n\|^2 = \lambda_1\|O_p(1)\|^2
\end{aligned}
\tag{2.12}
$$

according to Corollary A.9.2.1 of Mardia et al. (1979, p. 480), where λ_1 is the maximum eigenvalue of $\boldsymbol{X}'\boldsymbol{X}/n$ determined by (2.4). According to the conclusion $\lambda_1 \sim N(\widetilde{\lambda}_1, 2\,\widetilde{\lambda}_1^2/n)$ (approximately) in Theorem 8.3.3 of Mardia et al. (1979, p. 230), we can write

$$
n^{1/2}(\lambda_1 - \widetilde{\lambda}_1) = O_p(1) \quad or \quad \lambda_1 = \widetilde{\lambda}_1 + O_p(n^{-1/2}).
\tag{2.13}
$$

Substituting (2.13) into (2.12), we obtain

$$\|\boldsymbol{X}\boldsymbol{g}_n\|^2 \leq \widetilde{\lambda}_1 \,\|O_p(1)\|^2 + O_p(n^{-1/2})\|O_p(1)\|^2. \tag{2.14}$$

According to the definition for $O_p(1)$ and $O_p(n^{-1/2})$ in (2.8),

$$\widetilde{\lambda}_1 \,\|O_p(1)\|^2 + O_p(n^{-1/2})\|O_p(1)\|^2 \overset{P}{\to} 0, \tag{2.15}$$

where "$\overset{P}{\to}$" means convergence in probability. Combining (2.9)-(2.15), we can reach the conclusion that $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p\}$ given by (2.6) are asymptotically independent for a large sample size n as a result of the independence of $\{\boldsymbol{X}\,\widetilde{\boldsymbol{d}}_i: i = 1, \ldots, p\}$ and $\boldsymbol{X}\boldsymbol{g}_n \overset{P}{\to} 0$ ($n \to \infty$). This completes the proof.

From Theorem 1, each $\boldsymbol{z}_i$ $(i = 1, \ldots, p)$ given by (2.6) determines an exact Läuter's F-test

$$LF_i = n\bar{z}_i^2 \Big/ \left[\frac{1}{n-1}\sum_{j=1}^{n}(z_{ij} - \bar{z}_i)^2\right], \quad (i = 1, \ldots, p) \tag{2.16}$$

where $\bar{z}_i = \sum_{j=1}^{n} z_{ij}/n$ and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{in})'$ is given by (2.6). Under hypothesis (1.1), LF_i in (2.16) has an exact F-distribution $F(1, n-1)$ and $LF_1, \ldots, LF_p$ are asymptotically independent as a result of the asymptotic independence of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ in Theorem 1. Theoretically, each LF_i $(i = 1, \ldots, p)$ in (2.16) based on a single PC direction can be used as an individual exact F-test for hypothesis (1.1). A statistic that can collect the sample information from all PC directions in (2.4) should dominate any individual test based on a single PC direction. Since a large value of any of the Läuter's F-test in (2.16) indicates that the underlying normal distribution for the sample $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ may have a nonzero mean, we consider the following generalized F-statistic

$$GF = \max_{1 \leq i \leq r}\{LF_i\} \tag{2.17}$$

for testing hypothesis (1.1), where r represents the number of nonzero eigenvalues in (2.4), that is, $r = p$ if $n > p$, $r = n$ if $n < p$ and $r = p-1$ if $n = p$. A large value of GF implies possible rejection of hypothesis (1.1). We will call a test for (1.1) by GF in (2.17) the generalized F-test.

7

Since the Läuter's F-tests LF_i $(i = 1, \ldots, p)$ given by (2.16) have an exact F-distribution $F(1, n-1)$ and they are asymptotically i.i.d. under hypothesis (1.1), we can easily obtain the asymptotic null distribution of GF. For any $x > 0$,

$$P(GF < x) = P\left(\max_{1 \le i \le p}\{LF_i\} < x\right) \approx [F(x;\, 1, n-1)]^p, \qquad (2.18)$$

where $F(x;\, 1, n-1)$ represents the c.d.f. (cumulative distribution function) of the F-distribution $F(1, n-1)$. The p-value of the generalized F-test by GF in (2.17) can be approximately calculated by

$$P(GF > GF_0) \approx 1 - [F(GF_0;\, 1, n-1)]^p, \qquad (2.19)$$

where GF_0 stands for an observed value of GF calculated from a sample.

Theorem 2. *The generalized F-test for hypothesis (1.1) based on the GF-statistic (2.17) is robust in the sense that $GF(\boldsymbol{X}) \overset{\mathrm{d}}{=} GF(\boldsymbol{X}_0)$ for any random matrix $\boldsymbol{X}$ ($n \times p$) with a stochastic decomposition $\boldsymbol{X} \overset{\mathrm{d}}{=} s\boldsymbol{X}_0$, where "$\overset{\mathrm{d}}{=}$" means that the two sides of the equality have the same probability distribution, $s > 0$ is a positive random variable with probability 1 and $\boldsymbol{X}_0 \sim N_{n\times p}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$, which implies that the rows of $\boldsymbol{X}_0$ are i.i.d. and have a normal distribution $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$.*

Proof. *When the stochastic decomposition $\boldsymbol{X} \overset{\mathrm{d}}{=} s\boldsymbol{X}_0$ is true, the eigenvalue-eigenvector problem (2.4) can be written as*

$$\left(\frac{1}{n}s^2 \boldsymbol{X}_0' \boldsymbol{X}_0\right)\boldsymbol{D} = \boldsymbol{D}\boldsymbol{\Lambda}, \;\; or \;\; \left(\frac{1}{n}\boldsymbol{X}_0' \boldsymbol{X}_0\right)\boldsymbol{D} = \boldsymbol{D}\boldsymbol{\Lambda}/s^2. \qquad (2.20)$$

This implies that the eigenvectors $\{\boldsymbol{d}_i : i = 1, \ldots, p\}$ determined by $\boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_p)$ in (2.20) are the sample PC directions from the observation matrix $\boldsymbol{X}_0$ consisted of an i.i.d. sample from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then the random vectors

$$\hat{\boldsymbol{z}}_i = \boldsymbol{X}\boldsymbol{d}_i = s\boldsymbol{X}_0\boldsymbol{d}_i = s\boldsymbol{z}_i, \;\; with \;\; \boldsymbol{z}_i = \boldsymbol{X}_0\boldsymbol{d}_i, \qquad (2.21)$$

for $i = 1, \ldots, p$. By the definition of the GF-statistic (2.17), we have

$$GF(\boldsymbol{X}) = GF(\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_r) = \max_{1 \le i \le r}\{LF_i(\hat{\boldsymbol{z}}_i)\}, \qquad (2.22)$$

8

where $r =$ the number of positive eigenvalues in (2.20), and

$$GF(\boldsymbol{X}_0) = GF(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_r) = \max_{1 \le i \le r}\{LF_i(\boldsymbol{z}_i)\}. \tag{2.23}$$

Note that the statistics $LF_i(\boldsymbol{z}_i)$'s $(i = 1, \ldots, r)$ defined by (2.16) are asymptotically $(n \to \infty)$ i.i.d. and have the same F-distribution $F(1, n-1)$ when $\boldsymbol{z}_i$ is obtained from (2.21) according to Theorem 1. It is easy to verify that the location-scale invariance for the statistics $LF_i(\boldsymbol{z}_i)$'s:

$$LF_i(\hat{\boldsymbol{z}}_i) = LF_i(s\boldsymbol{z}_i) \equiv LF_i(\boldsymbol{z}_i), \quad i = 1, \ldots, r. \tag{2.24}$$

Summarizing (2.21) through (2.24), we can conclude

$$GF(\boldsymbol{X}) \stackrel{\mathrm{d}}{=} GF(\boldsymbol{X}_0), \ \text{if } \boldsymbol{X} \stackrel{\mathrm{d}}{=} s\boldsymbol{X}_0 \ \text{with } \boldsymbol{X}_0 \sim N_{n \times p}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}), \tag{2.25}$$

This completes the proof.

The conclusion in Theorem 2 implies that the GF-test (2.17) possesses robustness in the family of "generalized normal mixtures". That is, $GF(\boldsymbol{X})$ maintains the same distribution for $\boldsymbol{X} \stackrel{\mathrm{d}}{=} s\boldsymbol{X}_0$ with $s > 0$ being a positive random variable and $\boldsymbol{X}_0 \sim N_{n \times p}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$. The GF-test (2.17) also maintains almost all of the good properties (such as the dimension stability) that an individual Läuter's F-test possesses, except that GF does not have an exact F-distribution. For an arbitrarily chosen individual Läuter's F-test, when the direction $\boldsymbol{d} = f(\boldsymbol{X}'\boldsymbol{X})$ in Lemma 1 is not chosen as one of the sample PC directions as in (2.4), all individual Läuter's F-tests may have strong dependence among themselves. This may cause the α-level not to be maintained unless only one Läuter's F-test is used. One of the obvious weaknesses of using a single Läuter's F-test is the possible loss of sample information from other projection directions. The GF-test (2.17) can be considered to combine all sample information from individual PC directions into one statistic. This is a common practice in PC analysis (Jolliffe, 1986). A Monte Carlo study on the performance of the GF-test (2.17) will be given in Section 4.

3. Extension of the Generalized F-Test

The generalized F-test (2.17) can be extended to multiple comparison of normal population means as specified by the following hypothesis

$$H_0 : \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k, \qquad (k \geq 2) \tag{3.1}$$

versus the alternative hypothesis H_1: at least two means differ. This is exactly the problem of classical multivariate analysis of variance (MANOVA) when assuming normal populations with an identical covariance matrix. Let $\{\boldsymbol{x}_{ij} : i = 1, \ldots, n_i\}$ be an i.i.d. sample from a normal population $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ ($j = 1, \ldots, k$) and assume that the k samples are independent with one another. We want to test hypothesis (3.1). It is well-known that hypothesis (3.1) is commonly tested by the classical Wilks lambda statistic, which is defined by

$$\Lambda = |\boldsymbol{W}|/|\boldsymbol{W} + \boldsymbol{B}| \sim \Lambda(p, n - k, k - 1) \tag{3.2}$$

under hypothesis (3.1), where

$$\boldsymbol{W} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_j)' \text{ and } \boldsymbol{B} = \sum_{j=1}^{k} n_j (\bar{\boldsymbol{x}}_j - \bar{\bar{\boldsymbol{x}}})(\bar{\boldsymbol{x}}_j - \bar{\bar{\boldsymbol{x}}})' \tag{3.3}$$

are respectively the "within-samples" and "between-samples", and

$$\bar{\boldsymbol{x}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \boldsymbol{x}_{ij}, \quad \bar{\bar{\boldsymbol{x}}} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} \boldsymbol{x}_{ij}, \quad n = \sum_{j=1}^{k} n_j, \tag{3.4}$$

where $\bar{\boldsymbol{x}}_j$ stands for the j-th sample mean for the sample from the j-th population and $\bar{\bar{\boldsymbol{x}}}$ for the overall sample mean from all samples. The distribution $\Lambda(p, n - k, k - 1)$ in (3.2) is called the Wilks lambda distribution (see Mardia et al. (1979, p. 81)). Hypothesis (3.1) is rejected for small values of Λ in (3.2). The exact distribution of the Wilks Λ-statistic (3.2) is only available for the special cases of $k = 2$ and $k = 3$ (see Mardia et al. (1979, p. 83)):

$$\begin{aligned} k = 2 : \quad & \frac{n - p - 1}{p} \cdot \frac{1 - \Lambda(p, n - 2, 1)}{\Lambda(p, n - 2, 1)} \sim F(p, n - p - 1), \\ k = 3 : \quad & \frac{n - p - 2}{p} \cdot \frac{1 - \Lambda^{1/2}(p, n - 3, 2)}{\Lambda^{1/2}(p, n - 3, 2)} \sim F(2p, 2(n - p - 2)). \end{aligned} \tag{3.5}$$

For the general case of k, the asymptotic χ^2-distribution is employed when using the Wilks Λ-statistic (3.2):

$$-\left[n - k - \frac{1}{2}(p - k + 2)\right] \log \Lambda(p, n-k, k-1) \sim \chi^2((k-1)p), \quad n \to \infty. \quad (3.6)$$

Now we extend the GF-test (2.17) to testing hypothesis (3.1). Let

$$\boldsymbol{X} = (\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{1n_1}, \boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{2n_2}, \ldots, \boldsymbol{x}_{k1}, \ldots, \boldsymbol{x}_{kn_k})' : n \times p, \ n = \sum_{j=1}^{k} n_j \quad (3.7)$$

be the total observation matrix. The extended GF-test is based on the following theorem.

Theorem 3. *Let the total observation matrix $\boldsymbol{X}$ be defined by (3.7) and $\boldsymbol{A}$ a constant matrix defined by*

$$\boldsymbol{A} = \left(a_{ij}\right) : \quad (n-1) \times n, \qquad a_{ij} = \begin{cases} \dfrac{1}{\sqrt{i(i+1)}}, & j = 1, \ldots, i, \\ \dfrac{-i}{\sqrt{i(i+1)}}, & j = i+1, \\ 0, & \text{otherwise.} \end{cases} \quad (3.8)$$

Define the random matrix

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} : \qquad (n-1) \times n, \quad\quad\quad\quad\quad (3.9)$$

and the eigenvalue-eigenvector problem

$$\frac{1}{n-1} \boldsymbol{Y}'\boldsymbol{Y}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{\Gamma}, \quad\quad\quad\quad\quad (3.10)$$

where $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p)$ $(p \times p)$ consists of the eigenvectors $\{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p\}$ and $\boldsymbol{\Gamma} = diag(\gamma_1, \ldots, \gamma_p)$ consists of the eigenvalues $\gamma_1 \geq \ldots \geq \gamma_p > 0$ (assuming $n - 1 > p$). Let

$$\boldsymbol{u}_i = \boldsymbol{Y}\boldsymbol{h}_i, \quad i = 1, \ldots, p. \quad\quad\quad\quad\quad (3.11)$$

Under hypothesis (3.1), we have the following assertions:

(1) Each $\boldsymbol{u}_i$ $(i = 1, \ldots, p)$ has a spherical distribution with $P(\boldsymbol{u}_i = \boldsymbol{0}) = 0$;

11

(2) $\{\boldsymbol{u}_i: \ i = 1,\ldots,p\}$ *are asymptotically independent* $(n \to \infty)$.

Proof. *Assuming hypothesis (3.1) is true, we have $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k = \boldsymbol{\mu}$, say. From the assumption on the normal samples in (3.7), the random matrix $\boldsymbol{X}$ in (3.7) has a matrix normal distribution $N_{n \times p}(\mathbf{1}_n \boldsymbol{\mu}', \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$. So the random matrix $\boldsymbol{Y}$ in (3.9) also has a matrix normal distribution. The mean and the covariance matrix of $\boldsymbol{Y}$ can be computed as follows. By the definition of the constant matrix $\boldsymbol{A}$ in (3.8), it can be easily verified that $\boldsymbol{A}$ satisfies $\boldsymbol{A}\mathbf{1}_n = \mathbf{0}$ and $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}_{n-1}$. Then*

$$
\begin{aligned}
E(\boldsymbol{Y}) &= \boldsymbol{A}E(\boldsymbol{X}) = \boldsymbol{A}\mathbf{1}_n \boldsymbol{\mu}' = \mathbf{0}, \\
cov(vec\boldsymbol{Y}') &= (\boldsymbol{A} \otimes \boldsymbol{I}_p)cov(vec\boldsymbol{X}')(\boldsymbol{A}' \otimes \boldsymbol{I}_p) \\
&= (\boldsymbol{A} \otimes \boldsymbol{I}_p)(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma})(\boldsymbol{A}' \otimes \boldsymbol{I}_p) \\
&= (\boldsymbol{A}\boldsymbol{A}') \otimes \boldsymbol{\Sigma} = \boldsymbol{I}_{n-1} \otimes \boldsymbol{\Sigma}.
\end{aligned} \tag{3.12}
$$

That is, $\boldsymbol{Y} \sim N_{(n-1) \times p}(\mathbf{0}, \boldsymbol{I}_{n-1} \otimes \boldsymbol{\Sigma})$. This implies that $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-1})'$ is an $(n-1) \times p$ observation matrix with i.i.d. observations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n-1}\}$ from the normal population $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Comparing the eigenvalue-eigenvector problems (3.10) and (2.4), and (3.11) with (2.6), we can conclude that the two assertions in Theorem 3 are true as a result of the two assertions in Theorem 1. This completes the proof.

Following the same approach as to defining the GF-statistic (2.17), we can define the multiple GF-statistic (denote by MGF) for testing hypothesis (3.1) as follows. Let $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{i,n-1})'$ $(i = 1, \ldots, p)$ be given by (3.11) and $\bar{u}_i = \sum_{j=1}^{n-1} u_{ij}/(n-1)$. Define

$$
MLF_i(\boldsymbol{u}_i) = (n-1)\bar{u}_i^2 \Big/ \left[\frac{1}{n-2}\sum_{j=1}^{n-1}(u_{ij} - \bar{u}_i)^2\right]. \quad (i = 1, \ldots, p) \tag{3.13}
$$

Under hypothesis (3.1), $MLF_i(\boldsymbol{u}_i)$ in (3.13) has an exact F-distribution $F(1, n-2)$ and $MLF_1, \ldots, MLF_p$ are asymptotically independent as a result of the asymptotic independence of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ in Theorem 3. The MGF-test for (3.1)

is defined by

$$MGF = \max_{1 \le i \le p} \{MLF_i\}. \tag{3.14}$$

A large value of MGF implies rejection of hypothesis (3.1). Referring to (2.18) and (2.19), we can obtain the approximate null distribution of MGF in (3.14):

$$P(MGF < x) \approx [F(x; 1, n-2)]^p, \quad x \ge 0, \tag{3.15}$$

where $F(x; 1, n-2)$ represents the c.d.f. of the F-distribution $F(1, n-2)$ as in (2.18). The approximate p-value of the MGF-test for (3.1) by

$$P(MGF > MGF_0) \approx 1 - [F(MGF_0; 1, n-2)]^p, \tag{3.16}$$

where MGF_0 stands for an observed value of MGF calculated from the observations $\{\boldsymbol{x}_{ij} : i = 1, \dots, n_i; j = 1, \dots, k\}$ and n is the total sample size given by (3.4). By referring to Theorems 2 and 3, we have the following direct corollary.

Corollary 1. *The MGF-test (3.14) for hypothesis (3.1) is robust in the sense that $MGF(\boldsymbol{Y}) \stackrel{d}{=} MGF(\boldsymbol{Y}_0)$ for any random matrix $\boldsymbol{Y}$ $((n-1) \times p)$ with a stochastic decomposition $\boldsymbol{Y} \stackrel{d}{=} R\boldsymbol{Y}_0$ with $R > 0$ being a positive random variable and $\boldsymbol{Y}_0 \sim N_{(n-1) \times p}(\boldsymbol{0}, \boldsymbol{I}_{n-1} \otimes \boldsymbol{\Sigma})$.*

The robustness of the MGF-test (3.14) ensures the power stability for observed data with some kind of departure from normal distribution. Based on Corollary 1, we can expect the same power performance for non-normal data $\boldsymbol{Y} \stackrel{d}{=} R\boldsymbol{Y}_0$ as for the normal data $\boldsymbol{Y}_0$ after the transformation (3.9) for the raw data. A simple empirical comparison between the MGF-test (3.14) and the classical Wilks Λ-test (3.2) is given in next section.

4. A Monte Carlo Study

4.1. The finite-sample property

The asymptotic distribution (2.18) of the GF-statistic (2.17) is obtained under large sample sizes. Its performance under finite-sample sizes can be

studied by the Monte Carlo method. A commonly-used method is to compare the finite-sample quantiles of the GF-statistic (2.17) with the quantiles of the asymptotic distribution (2.18) under the null hypothesis (1.1). This can be assessed by the Q-Q (quantile-quantile) plot method. The finite-sample quantiles of the GF-statistic (2.17) can be approximated by Monte Carlo simulation. Since the null distribution (2.18) does not depend on the unknown covariance matrix $\boldsymbol{\Sigma} > \mathbf{0}$ in hypothesis (1.1), we can generate empirical samples from the standard normal $N_p(\mathbf{0}, \boldsymbol{I}_p)$ in the Monte Carlo study. Let GF_k denote the empirical value of GF in (2.17) obtained from the k-th set of generated sample ($k = 1, \ldots, 10,000$). Repeating the simulation for 10,000 times, we obtain 10,000 values of GF and arrange them in ascending order

$$GF_{(1)} \leq \ldots \leq GF_{(10000)}. \tag{4.1}$$

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{99}) = (0.01, 0.02, \ldots, 0.99)$. The empirical finite-sample quantiles of GF at the percentages in $\boldsymbol{\alpha}$ can be approximated by

$$GF_{(10000\alpha_1)} \leq GF_{(10000\alpha_2)} \leq \ldots \leq GF_{(10000\alpha_{99})}. \tag{4.2}$$

The corresponding asymptotic quantiles of GF by the distribution in (2.18) are computed by

$$F^{-1}[(1 - \alpha_i)^{1/p}; 1, n - 1], \qquad i = 1, \ldots, 99, \tag{4.3}$$

where $F^{-1}(\cdot; 1, n-1)$ represents the inverse c.d.f. of the F-distribution $F(1, n-1)$. The Q-Q plots for a number of selected sets of (n, p) under 10,000 simulation replications are presented in Fig. 1, where all the Q-Q plots show satisfactory approximation: the empirical quantiles given by (4.2) (at the horizontal axes) are roughly equal to the asymptotic quantiles computed by (4.3) (at the vertical axes) because all plots are close to the equiangular line $y = x$.

Insert Fig. 1 here

4.2. Type I error rates

The goodness-of-fit of the finite-sample distribution of GF in (2.17) by the asymptotic distribution in (2.18) can be further studied by comparing the empirical type I error rates of GF with the approximate type I error rates computed from the asymptotic distribution in (2.18). Let $x_{1-\alpha}$ be the $(1-\alpha)$-quantile (e.g., $\alpha = 1\%$, 5% and 10%) computed from (2.18). That is,

$$P(GF < x_{1-\alpha}) \approx [F(x_{1-\alpha};\, 1, n-1)]^p = 1-\alpha, \text{ or } x_{1-\alpha} = F^{-1}[(1-\alpha)^{1/p};\, 1, n-1].$$
(4.4)

After generating 5,000 sets of empirical samples from $N_p(\mathbf{0}, \boldsymbol{I}_p)$ for each pair of (n, p), we summarize the empirical type I error rates of GF in Table 1 for $\alpha = 1\%$, 5% and 10%. The results in Table 1 show that the empirical type I error rates of GF in (2.17) are well approximated by the type I error rates computed from the distribution in (2.18) for most cases of (n, p) with $n > p$ or $n \leq p$. This also provides the justification of using the asymptotic distribution in (2.18) as the finite-sample null distribution of generalized F-statistic GF in various cases of (n, p).

Insert Table 1 here

4.3. Power study

Since the asymptotic distribution in (2.18) gives satisfactory type I error rates as presented in Table 1, we can use it as the finite-sample null distribution in computing the empirical power of the GF-statistic (2.17). Without loss of generality, we can generate empirical samples from an alternative normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by assuming $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ and $\boldsymbol{\mu} = c\mathbf{1}_p$, where $\mathbf{1}_p$ stands for the vector of ones $(p \times 1)$, c is a constant to be taken as 0 (0.05) 0.30 (i.e., the values c=0, 0.05, 0.10, ..., 0.30). We call $\mathbf{1}_p$ the power direction in Table 2. It indicates that the power is computed based on the increasing distance of the mean vector from the origin $(c = 0)$ along this direction. We have tried different choices of power directions in our simulation and obtained similar

15

power performance to that based on the direction $\mathbf{1}_p$ for each selected set of (n, p). The statistics in Table 2 are:

(a) GF: the generalized F-statistic (2.17);

(b) LF_1: Läuter's F-test in (2.16) by choosing the first PC direction $\boldsymbol{d}_1$ determined by (2.4);

(c) LF_m: Läuter's F-test in (2.16) by choosing the middle PC direction $\boldsymbol{d}_m$ determined by (2.4) with $m = [p/2]$ if $n > p$, $m = [n/2]$ if $n < p$, and $m = [(p-1)/2]$ if $n = p$, where $[\cdot]$ denotes the integer part of a real number;

(d) LF_p and LF_n: Läuter's F-test in (2.16) by choosing the PC direction $\boldsymbol{d}_p$ (for $n > p$) or $\boldsymbol{d}_n$ (for $n < p$) or d_{p-1} (for $n = p$) determined by (2.4).

The following conclusions can be summarized:

(1) The generalized F-test GF seems to be powerful for all pairs of (n, p) when the normal mean $\boldsymbol{\mu} = c\mathbf{1}_p$ increases along the direction $\mathbf{1}_p$ at $c \geq 0.20$ for $n > p$ and at $c = 0.15$ for $n \leq p$. It has approximately the same power performance as that for the Läuter's F-test LF_1 in the case of $n \leq p$;

(2) GF could improve the power performance of the Läuter's F-test LF_1 in the case of large sample sizes. For example, for the cases $n = 100$, $p = 5$ and $p = 10$, $c = 0.10$, GF could improve the power of the best Läuter's F-test LF_1 by more than 30%;

(3) Not all PC directions are suitable for constructing the Läuter's F-tests. When using the PC directions $\boldsymbol{d}_i$ for $i \geq m = [p/2]$ ($n > p$) or $i \geq m = [n/2]$ ($n \leq p$) in (2.4), Läuter's F-tests may completely lose power. This is indicated by the zero power values in Table 2;

(4) GF completely outperforms the individual Läuter's F-tests LF_m and LF_p ($n > p$) or LF_n ($n \leq p$) constructed from the PC directions associated with small eigenvalues in (2.4).

Insert Table 2 here

16

4.4. Empirical study on multiple comparison

In this subsection we present a simple Monte Carlo study on the multiple comparison of normal means specified by hypothesis (3.1). The MGF-test (3.14) is compared with the classical Wilks Λ-test (3.2) in the performance of controlling type I error rates and power. Because the asymptotic distribution (see (3.15)) of the MGF-test (3.14) does not depend on the common mean $\boldsymbol{\mu}$ and the common covariance matrix $\boldsymbol{\Sigma}$ under the null hypothesis (3.1), we can choose an arbitrary common mean vector $\boldsymbol{\mu} = \mathbf{1}_p$, say, and $\boldsymbol{\Sigma} = \left(\sigma_{ij}\right)$ with $\sigma_{ii} \equiv 1$ and $\sigma_{ij} \equiv 0.5$ for $i \neq j$ $(i, j = 1, \ldots, p)$ in the Monte Carlo study. The results for selected dimensions $p =$5, 10, 15 and 20 with the balanced sample size for all populations $n_i \equiv 5$ and $n_i \equiv 10$ are summarized in Table 3, where the critical values for the MGF-test are computed by (3.15) and those for Wilks' Λ-test by (3.6), the number of simulation repetition is 2,000. For comparison, the corresponding results from Wilks Λ-test are also presented in Table 3. It shows that the type I error rates of the MGF-test (3.14) are satisfactorily controlled for all cases as considered. But Wilks Λ-test has poor control of type I error rates for the cases of high dimension with small sample sizes. For example, for the cases of $p \geq 15$ and $n_i \equiv 5$ in Table 3, Wilks Λ-test has much higher type I error rates than those of the MGF-test. This implies that, in the case of high dimension with small sample sizes, Wilks Λ-test tends to reject the null hypothesis more frequently than does the MGF-test when the null hypothesis is actually true.

$$\boxed{\text{Insert Table 3 here}}$$

Power comparison between the MGF-test and Wilks Λ-test can be partially studied by assuming equal difference between any two population means in the alternative hypothesis of the null hypothesis (3.1). That is, let

$$\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i = c\mathbf{1}_p, \qquad i = 1, \ldots, k - 1, \tag{4.5}$$

in the alternative hypothesis of (3.1) and the same common covariance matrix $\boldsymbol{\Sigma}$ as that in Table 3, where $c > 0$ in (4.5) is a constant for measuring the increasing difference between the population means. The simulation results for the selected cases are presented in Fig. 2, where the number of simulation repetition=2,000 and the balanced sample size $n_i \equiv 10$, the critical values for the MGF-test and those for Wilks' Λ-test are obtained in the same way as in Table 3, the number of simulation repetition is 2,000. Fig. 2 shows that the MGF-test clearly outperforms Wilks Λ-test, especially when the dimension becomes higher and higher while the sample size remains unchanged.

$$\boxed{\text{Insert Fig. 2 here}}$$

5. Applications

Example 1. The real data were collected from 19 depressive patients acquired at the beginning and at the end of a six-week therapy. The nine variables represent the changes of the absolute theta power of electroencephalogram (EEG) during the therapy in nine selected channels ($n = 19$ and $p = 9$). The full data can be found in in Läuter et al. (1996). The purpose is to see whether the six-week therapy is effective (effectiveness means that significant nonzero changes of the absolute theta power are observed after the six-week therapy).

The sample data in this example can be considered to be coming from a 9-dimensional population associated with a 9-dimensional random vector $\boldsymbol{x} = (X_1, \ldots, X_9)'$. Without verifying the multivariate normal assumption on the data set, Läuter et al. (1996) employed Läuter's (1996) exact t-test (2.2) to test hypothesis (1.1) by choosing two projection directions: (a) the SS-test by choosing the projection direction $\boldsymbol{d} = \text{diag}(\boldsymbol{X}'\boldsymbol{X})\mathbf{1}_p$ ($\boldsymbol{X}$ is the $n \times p$ observation matrix, and $\text{diag}(\boldsymbol{X}'\boldsymbol{X})$ stands for the diagonal matrix with the same diagonal elements as in $\boldsymbol{X}'\boldsymbol{X}$); and (b) the PC-test by choosing the first PC direction associated with the largest eigenvalue λ determined by

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{d} = \lambda\text{diag}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{d}, \quad \boldsymbol{d}'\text{diag}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{d} = 1.$$

Läuter et al. (1996) got the p-value 0.0489 for the SS-test and the p-value 0.0487 for the PC tests. Hence, they claimed that the six-week therapy is effective at $\alpha = 5\%$ and therefore reject the null hypothesis in (1.1). There are two questions needed to be clarified in their analysis: (1) Is the multivariate normal assumption on the underlying distribution of the sample satisfied? and (2) What is the actual significance level when using the two individual Läuter's exact t-tests simultaneously? While question (1) may be easily answered by employing some popular tests for multinormality such as Mardia's (1970) skewness and kurtosis tests, it is difficult to answer question (2) since it is hard to show whether the two individual Läuter's exact t-tests are asymptotically independent.

Actually, the 9-dimensional sample data set used by Läuter et al. (1996) was found to show evidence of departure from multinormality by Mardia's (1970) multivariate skewness and kurtosis tests (e.g., p-values based on 10,000 replications of simulation are 0.0053 and 0.0000, respectively) and by the low-dimensional projection tests for multinormality in Liang et al. (2000) with almost all p-values below 0.001. It was pointed out by Liang et al. (2000) that the source of non-multinormality of the whole data set is very likely to come from the two variables X_2 and X_5 based on the single one-dimensional skewness and kurtosis tests. We apply the power transformation $y = \text{sign}(x)|x|^{\beta}$ ($\beta > 0$) to the observations from X_2 and X_5. Here, $\text{sign}(x)$ denotes the common sign function. It turns out that $\beta = 1/19$ gives a p-value=0.1202 for Mardia's 9-dimensional skewness test and a p-value= 0.8264 for Mardia's 9-dimensional kurtosis test by generating 10,000 sets of standard normal samples (Note: Mardia's skewness and kurtosis are location-scale invariant. We can generate standard normal samples in the simulation without loss of generality). Therefore, after the power transformation $y = \text{sign}(x)|x|^{1/19}$ on the observations from the two variables X_2 and X_5, we can consider the 9-dimensional

data set from the population associated with the 9-dimensional random vector

$$\boldsymbol{x} = (X_1, \, \text{sign}(X_2)|X_2|^{1/19}, \, X_3, \, X_4, \, \text{sign}(X_5)|X_5|^{1/19}, \, X_6, \ldots, X_9)' \qquad (5.1)$$

as multivariate normally distributed. Hence, the multivariate normal assumption is now satisfied and we can carry out the Läuter's F-tests (2.16) and the generalized F-test (2.17). The results are summarized in Table 4, where the observed values of the statistics are computed from the sample data after the above transformation (5.1), the exact P-V (P-V=p-value) of the corresponding statistics are computed from their exact distributions: for the Läuter's tests LF_i $(i = 1, \ldots, 9)$, their exact distributions are the same F-distribution $F(1, n-1) = F(1, 18)$ $(n = 19)$; for the generalized F-test GF, its exact distribution is taken as the asymptotic distribution given in (2.18). The empirical P-V for each statistic in Table 4 is obtained from 10,000 sets of empirical standard normal samples. The p-values for all tests in Table 4 are greater than the significance level $\alpha = 5\%$. This provides enough evidence on the insignificance of all tests in Table 4. That is, we should not reject hypothesis (1.1) and consider there is no evidence to show the effectiveness of the six-week therapy based on the experimental data at the significance level $\alpha = 5\%$. From Table 4, the p-value of Läuter's test LF_1 is $0.0559 < 10\%$ but all other Läuter's tests have p-values that are greater than 10%. This implies that individual Läuter's tests may give inconsistent results in practice.

Insert Table 4 here

Example 2. In longitudinal studies, sample data are usually observed at different time points. Suppose that we have observations at p time points: $t = 1, 2, \ldots, p$, here $t = 1$ means the initial time point (e.g., before the start of an experiment). The observations of a random index (variable) X from n experimental subjects at the p time points are denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, \boldsymbol{x}_{ip})'$ for $i = 1, \ldots, n$. For example, each observed value x_{ij} $(i = 1, \ldots, n; \; j = 1, \ldots, p)$

20

may denote the tumor volume of mouse i at week j after some treatment in a cancer research project. Assume that the data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are independently observed and from the same population and $E(\boldsymbol{x}_i) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ $(i = 1, \ldots, n)$. We want to test the null hypothesis

$$H_0: \ \mu_1 = \mu_2 = \ldots = \mu_p. \tag{5.2}$$

Note that this is not the usual one-way ANOVA (analysis of variance) problem since the observations at different time points are usually dependent and may not have the same variance in longitudinal studies. In order to test hypothesis (5.2) by our generalized F-test (2.17) and the Läuter's F-tests (2.16), let

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)', \qquad \boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)' = \boldsymbol{X}\boldsymbol{A},$$

$$\boldsymbol{A} = \left(a_{ij}\right): \quad p \times (p-1), \qquad a_{ij} = \begin{cases} 1, & \text{if } i = j, \\ -1, & \text{if } i = j+1, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Hence, testing hypothesis (5.2) from the observation matrix $\boldsymbol{X}$ in (5.3) is equivalent to testing hypothesis

$$H_0: \ \boldsymbol{\mu}_y = (\mu_1 - \mu_2, \ \mu_2 - \mu_3, \ \ldots, \ \mu_{p-1} - \mu_p)' = \boldsymbol{0}, \quad \boldsymbol{0}: \ (p-1) \times 1, \tag{5.4}$$

from the observation matrix $\boldsymbol{Y}$ in (5.3), where $\boldsymbol{\mu}_y$ is the mean vector of the i.i.d. observations $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ in (5.3). If we assume that $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in (5.3) is an i.i.d. p-dimensional normal sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ in (5.3) is a $(p-1)$-dimensional i.i.d. normal sample from $N_{p-1}(\boldsymbol{\mu}_y, \boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A})$ with $\boldsymbol{A}$ given in (5.3). If hypothesis (5.4) is rejected at level α (e.g., 1%, 5% and 10%), we can conclude that hypothesis (5.2) is also rejected at level α. It is obvious that hypothesis (5.4) is only one of the many equivalent alternatives to hypothesis (5.2).

Now we apply the above transformation method to a real data set. Tan et al. (2005) examined the multivariate normal assumption on a data set

used in a cancer research project. The data of tumor volumes of mice in experiment were observed weekly for a period of 12 weeks. It is the aim to study the activity of a new anticancer drug irinotecan (CPT-11) against neuroblastoma in xenograft models for two treatment regimens. The tumor-bearing mice in Group I received 0.4 mg/kg CPT-11 and those in Group II received CPT-11 0.26 mg/kg. Mice from the same strain were used and they are virtually genetically identical. Eleven mice have been successfully cultured with subcutaneous transplant of tumor and are divided into two groups. The tumor volumes (cm^3) were measured at the initial time and once every week for 12 weeks. Missing data arise because six mice died of toxicity or were sacrificed due to its tumor volume quadrupled early (it is marked "*" in Table 4, meaning missing values in the observation). The full data set in Table 5 is borrowed from Tan et al. (2005).

Insert Table 5 here

From Table 5, it can be seen that most of the mice (except mouse #10) can survive in the first 9 weeks. We have complete data from 10 mice (except mouse 10) in the first 9 weeks. It is of interest to see whether there is a significant difference among the cancer tumor volumes in the 10 mice during the first 9 weeks. This leads us to consider the statistical hypothesis (5.2) with $p = 10$ (i.e., weeks 0, 1, ..., 9). That is, we want to test hypothesis (5.4) based on the observed data in $\boldsymbol{Y}$ given by (5.3) with dimension $p - 1 = 9$. As in Example 1, we also employ Mardia's (1970) multivariate skewness and kurtosis tests to check the multinormality assumption of the data in $\boldsymbol{Y}$ given by (5.3). By generating 10,000 sets of standard normal samples, we obtain p-value=0.8090 for Mardia's skewness test and p-value=0.1855 for Mardia's kurtosis test. Therefore, we can consider that the data in $\boldsymbol{Y}$ given by (5.3) with dimension $p - 1 = 9$ are from a 9-dimensional normal population. Hence, we can apply the Läuter's tests (2.16) and the generalized F-test (2.17) to

22

hypothesis (5.4). The results are summarized in Table 6, where the observed values of the statistics are computed from the data in $\boldsymbol{Y}$ given by (5.3), the exact P-V (P-V=p-value) of the corresponding statistics are computed from their exact distributions: for the Läuter's tests LF_i ($i = 1, \ldots, 9$), their exact distributions are the same F-distribution $F(1, n-1) = F(1, 9)$ ($n = 10$); for the generalized F-test GF, its exact distribution is taken as the asymptotic distribution given in (2.18). The empirical P-V for each statistic in Table 6 is obtained from 10,000 sets of empirical standard normal samples. The p-values of almost all tests (except LF_2, whose exact p-value= 0.0164 < 5%) in Table 6 show that all tests are insignificant at level $\alpha = 5\%$, implying no significant difference among the cancer tumor volumes observed in the first 9 weeks from the 10 selected mice. That is, we do not reject hypothesis (5.4) at level 5% based on the observed data from the first 9 weeks. This can be interpreted as no significant increase for the cancer tumors in the mice during the first 9 weeks. The drug seems to take effect after week nine for most mice. It is noted that the individual Läuter's tests in Table 5 give inconsistent results: seven of them have p-values> 10%, implying insignificance at level 10%, while two of them (LF_1 and LF_2) have p-values< 10%, implying significance at level 10%. This results in inconsistent interpretation in practical application.

Insert Table 6 here

6. Remarks

From the limited Monte Carlo study and the applications in two real data sets, we conclude that the generalized F-test GF in this paper is generally at least as good as the individual Läuter's F-tests constructed from the PC directions in (2.4). GF could markedly improve the individual Läuter's F-tests when the Läuter's test chooses an unsuitable projection direction such as those PC directions associated with small eigenvalues. Application in the two examples shows that the generalized F-test GF is superior to the individual

Läuter's F-tests based on PC directions when they are used separately, which may lead to differences among their p-values for the same data set.

It is a remarkable advantage that both the proposed generalized F-test and the Läuter's F-tests are applicable to testing the multivariate normal mean with high dimension and possibly very small sample size compared to the dimension. Classical tests such as the Hotelling T^2-test or likelihood-ratio-type tests require that the sample size should be larger than the dimension to ensure good approximation by the limiting distribution of a test statistic. In terms of dimension stability, both the proposed generalized F-test and the Läuter's F-tests are superior to the Hotelling T^2-test or likelihood-ratio-type tests. Example 2 illustrates the application of the generalized F-test and the Läuter's F-tests in the case of very small sample size compared to the dimension.

Extension of the GF-test (2.17) to the MGF-test (3.14) for multiple comparison of normal population means greatly improves the classical Wilks Λ-test in controlling type I error rates and maintaining high power in the case of high dimension with small sample sizes. This implies that the MGF-test (3.14) is especially suitable for the case of high dimension with small sample sizes.

It is noted that the GF-test (2.17), the Läuter's F-tests and the MGF-test (3.14) do not possess location-scale invariance as compared to the Hotelling T^2-test and Wilks Λ-test. Glimm and Läuter (2003) pointed out that location-scale invariance is not always an advantage in some practical data analysis. Both the GF-test (2.17) and the MGF-test (3.14) possess some kind of robustness against a departure from multinormality for the raw data as summarized in Theorem 2 and Corollary 1. This implies that the GF-test (2.17) and the MGF-test (3.14) can still give reliable results for some kind of non-normal data as mentioned in Theorem 2 and Corollary 1. It is an open problem to find the noncentral distributions of the Läuter's F-tests (2.16), the GF-test (2.17), and the MGF-test (3.14). This exceeds the scope of this paper.

References

Anderson T.W., 1984. An Introduction to Multivariate Statistical Analysis (2nd Edition). John Wiley & Sons, New York.

Fang, K.T., Kotz, S., Ng, K.W., 1990. Symmetric Multivariate and Related Distributions. Chapman and Hall, London and New York.

Fang, K.T., Li, R., Liang, J., 1998. A multivariate version of Ghosh T_3-plot to detect non-multinormality. Comput. Statist. & Data Anal. 28, 371–386.

Glimm, E., Läuter, J., 2003. On the admissibility of stable spherical multivariate tests. J. Multivar. Anal. 86, 254–265.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer-Verlag, New York.

Läuter, J., 1996. Exact t and F tests for analyzing studies with multiple endpoints. Biometrics 52, 964–970.

Läuter, J., Glimm, E., Kropf, S., 1996. New multivariate tests for data with an inherent structure. Biometrical Journal 38, 5–23.

Liang, J., Fang, K.T., 2000. Some applications of Läuter's technique in tests for spherical symmetry. Biometrical Journal 42, 923–936.

Liang, J., Li, R., Fang, H., Fang, K.T., 2000. Testing multinormality based on low-dimensional projection. J. Statist. Plann. & Infer. 86, 129–141.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika 57, 519–530.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press, London and New York.

Tan, M., Fang, H.B., Tian, G.L., Wei, G., 2005. Testing multivariate normality in incomplete data of small sample size. J. Multivar. Anal. 93, 164–179.

Table 1

Empirical type I error rates by using the approximate distributions (No. of simulations=5,000)

(n,p)	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	(n,p)	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
(20,3)	0.0100	0.0484	0.0968	(40,3)	0.0100	0.0526	0.0964
(20,5)	0.0090	0.0506	0.0994	(40,5)	0.0086	0.0484	0.1012
(20,8)	0.0092	0.0492	0.1030	(40,8)	0.0104	0.0524	0.1006
(20,10)	0.0108	0.0514	0.1060	(40,10)	0.0120	0.0510	0.1004
(25,30)	0.0096	0.0428	0.0894	(30,30)	0.0092	0.0502	0.1002
(30,35)	0.0080	0.0424	0.0872	(35,35)	0.0096	0.0550	0.1146
(35,40)	0.0100	0.0500	0.1016	(40,40)	0.0090	0.0514	0.1058
(40,45)	0.0084	0.0424	0.0932	(45,45)	0.0114	0.0546	0.1058

Table 2

Power comparison between the generalized F-test and selected Läuter's tests
(No. of simulations=2,000)

Sample		Dimension $p = 5$, Power Direction=$c\mathbf{1}_5$, $\alpha = 5\%$						
Size	Stat.	$c = 0.00$	$c = 0.05$	$c = 0.10$	$c = 0.15$	$c = 0.20$	$c = 0.25$	$c = 0.30$
$n = 40$	GF	0.0590	0.0690	0.2345	0.8000	0.9995	1.0000	1.0000
	LF_1	0.0500	0.0650	0.2335	0.7440	0.9930	1.0000	1.0000
	LF_m	0.0550	0.0615	0.1530	0.2815	0.1035	0.0030	0.0000
	LF_p	0.0510	0.0565	0.0700	0.0560	0.0145	0.0005	0.0000
$n = 100$	GF	0.0480	0.1000	0.6255	0.9970	1.0000	1.0000	1.0000
	LF_1	0.0465	0.0805	0.4640	0.9655	0.9995	1.0000	1.0000
	LF_m	0.0555	0.0875	0.3135	0.3290	0.0575	0.0015	0.0000
	LF_p	0.0435	0.0595	0.1360	0.0855	0.0145	0.0000	0.0000
		Dimension $p = 10$, Power Direction=$c\mathbf{1}_{10}$, $\alpha = 5\%$						
		$c = 0.00$	$c = 0.05$	$c = 0.10$	$c = 0.15$	$c = 0.20$	$c = 0.25$	$c = 0.30$
$n = 40$	GF	0.0535	0.0700	0.2795	0.9365	1.0000	1.0000	1.0000
	LF_1	0.0465	0.0640	0.3165	0.9125	1.0000	1.0000	1.0000
	LF_m	0.0520	0.0620	0.1000	0.0525	0.0015	0.0000	0.0000
	LF_p	0.0555	0.0515	0.0490	0.0155	0.0010	0.0000	0.0000
$n = 100$	GF	0.0555	0.1120	0.7510	1.0000	1.0000	1.0000	1.0000
	LF_1	0.0515	0.0995	0.5775	0.9970	1.0000	1.0000	1.0000
	LF_m	0.0565	0.0790	0.1695	0.0615	0.0020	0.0000	0.0000
	LF_p	0.0540	0.0625	0.0785	0.0220	0.0015	0.0000	0.0000
Sample		Dimension $p = 30$, Power Direction=$c\mathbf{1}_{30}$, $\alpha = 5\%$						
Size	Stat.	$c = 0.00$	$c = 0.05$	$c = 0.10$	$c = 0.15$	$c = 0.20$	$c = 0.25$	$c = 0.30$
$n = 25$	GF	0.0515	0.0520	0.1710	0.9590	1.0000	1.0000	1.0000
	LF_1	0.0545	0.0780	0.3715	0.9815	1.0000	1.0000	1.0000
	LF_m	0.0420	0.0420	0.0230	0.0010	0.0000	0.0000	0.0000
	LF_n	0.0505	0.0430	0.0170	0.0095	0.0080	0.0045	0.0015
$n = 30$	GF	0.0505	0.0495	0.2660	0.9970	1.0000	1.0000	1.0000
	LF_1	0.0580	0.0830	0.4550	0.9955	1.0000	1.0000	1.0000
	LF_m	0.0435	0.0405	0.0220	0.0000	0.0000	0.0000	0.0000
	LF_n	0.0520	0.0450	0.0185	0.0045	0.0075	0.0015	0.0010

Table 3.

Comparison of the empirical type I error rates between Wilk's Λ-test and the MGF-test

		Multiple comparison of $k = 5$ population means					
		$p = 5$			$p = 10$		
	Test	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$n_i \equiv 5$	Λ	0.0085	0.0485	0.0985	0.0165	0.0755	0.1330
	MGF	0.0075	0.0485	0.1000	0.0090	0.0485	0.1055
$n_i \equiv 10$	Λ	0.0070	0.0430	0.0880	0.0085	0.0605	0.1090
	MGF	0.0135	0.0545	0.1035	0.0105	0.0505	0.1120
	Test	$p = 15$			$p = 20$		
$n_i \equiv 5$	Λ	0.0650	0.1805	0.2860	0.7385	0.8605	0.9145
	MGF	0.0120	0.0440	0.1000	0.0085	0.0500	0.1010
$n_i \equiv 10$	Λ	0.0110	0.0635	0.1130	0.0175	0.0775	0.1385
	MGF	0.0110	0.0635	0.1175	0.0135	0.0595	0.1120
		Multiple comparison of $k = 10$ population means					
		$p = 5$			$p = 10$		
	Test	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$n_i \equiv 5$	Λ	0.0105	0.0475	0.0945	0.0145	0.0605	0.1180
	MGF	0.0060	0.0465	0.0960	0.0045	0.0390	0.0820
$n_i \equiv 10$	Λ	0.0110	0.0505	0.1000	0.0060	0.0410	0.0875
	MGF	0.0085	0.0505	0.0985	0.0115	0.0535	0.0970
	Test	$p = 15$			$p = 20$		
$n_i \equiv 5$	Λ	0.0205	0.0690	0.1310	0.0265	0.1065	0.1900
	MGF	0.0100	0.0530	0.1000	0.0140	0.0545	0.1080
$n_i \equiv 10$	Λ	0.0105	0.0555	0.1095	0.0115	0.0515	0.1105
	MGF	0.0115	0.0535	0.1025	0.0105	0.0600	0.1065

Table 4

Comparison between empirical p-values and exact p-values of the GF-test and Läuter's tests

Test	Observed value	Exact P-V	Empirical P-V
GF	4.1749	0.4044	0.4439
LF_1	4.1749	0.0559	0.0530
LF_2	0.0509	0.8240	0.8225
LF_3	0.3341	0.5704	0.5764
LF_4	0.2064	0.6550	0.6599
LF_5	0.0548	0.8176	0.8130
LF_6	2.4705	0.1334	0.1323
LF_7	0.0346	0.8546	0.8567
LF_8	0.1546	0.6988	0.6991
LF_9	0.9772	0.3360	0.3352

Table 5

Volumes (cm^3) of NB-SD tumor measured in 12 weeks for different combinations

		Weeks												
Group	Mouse	0	1	2	3	4	5	6	7	8	9	10	11	12
	1	2.34	2.48	2.04	1.06	1.26	0.91	0.84	1.08	0.86	0.59	*	*	*
	2	1.11	1.54	0.81	0.93	1.37	1.00	0.98	1.70	0.77	0.51	1.88	4.97	*
I	3	0.96	0.99	0.99	0.53	0.72	0.36	0.17	0.63	0.45	0.48	2.14	2.89	*
	4	0.66	0.60	0.49	0.78	1.40	1.33	1.11	1.31	0.69	0.98	3.44	3.08	*
	5	2.08	2.15	1.87	0.83	0.78	0.26	0.16	0.64	0.30	0.51	2.02	*	*
	6	1.09	1.04	0.76	0.77	0.72	0.29	0.20	0.29	0.68	0.55	0.55	*	*
	7	0.74	0.93	0.83	0.59	0.60	0.41	0.22	0.31	0.15	0.18	1.03	8.21	*
II	8	0.94	1.12	1.67	2.69	3.51	2.77	2.36	2.89	2.49	2.64	3.96	*	*
	9	1.84	1.99	2.75	4.29	6.41	4.04	3.20	3.89	4.10	6.38	*	*	*
	10	1.21	1.41	1.97	2.07	2.98	2.30	*	*	*	*	*	*	*
	11	1.24	1.32	1.63	2.43	3.00	2.04	1.08	1.07	0.39	0.88	1.86	5.90	*

Table 6

Comparison between empirical p-values and exact p-values of the GF-test and Läuter's tests

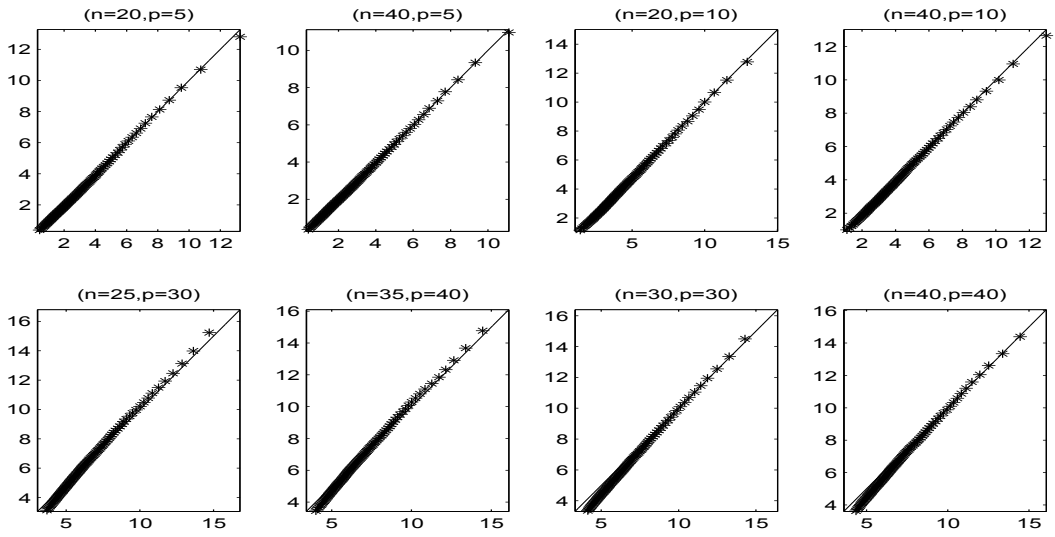| Test	Observed value	Exact P-V	Empirical P-V
GF	8.6533	0.1386	0.1486
LF_1	4.2508	0.0693	0.0710
LF_2	8.6533	0.0164	0.0155
LF_3	0.0486	0.8304	0.8316
LF_4	0.1406	0.7164	0.7177
LF_5	0.3415	0.5733	0.5852
LF_6	0.0012	0.9730	0.9736
LF_7	0.1608	0.6978	0.6984
LF_8	0.2633	0.6202	0.6094
LF_9	0.4154	0.5353	0.5324

Fig. 1. Q-Q plots for the empirical quantiles (4.2) (the horizontal axis) of the generalized F-statistic (2.17) versus the asymptotic quantiles (the vertical axis) computed by (4.3).
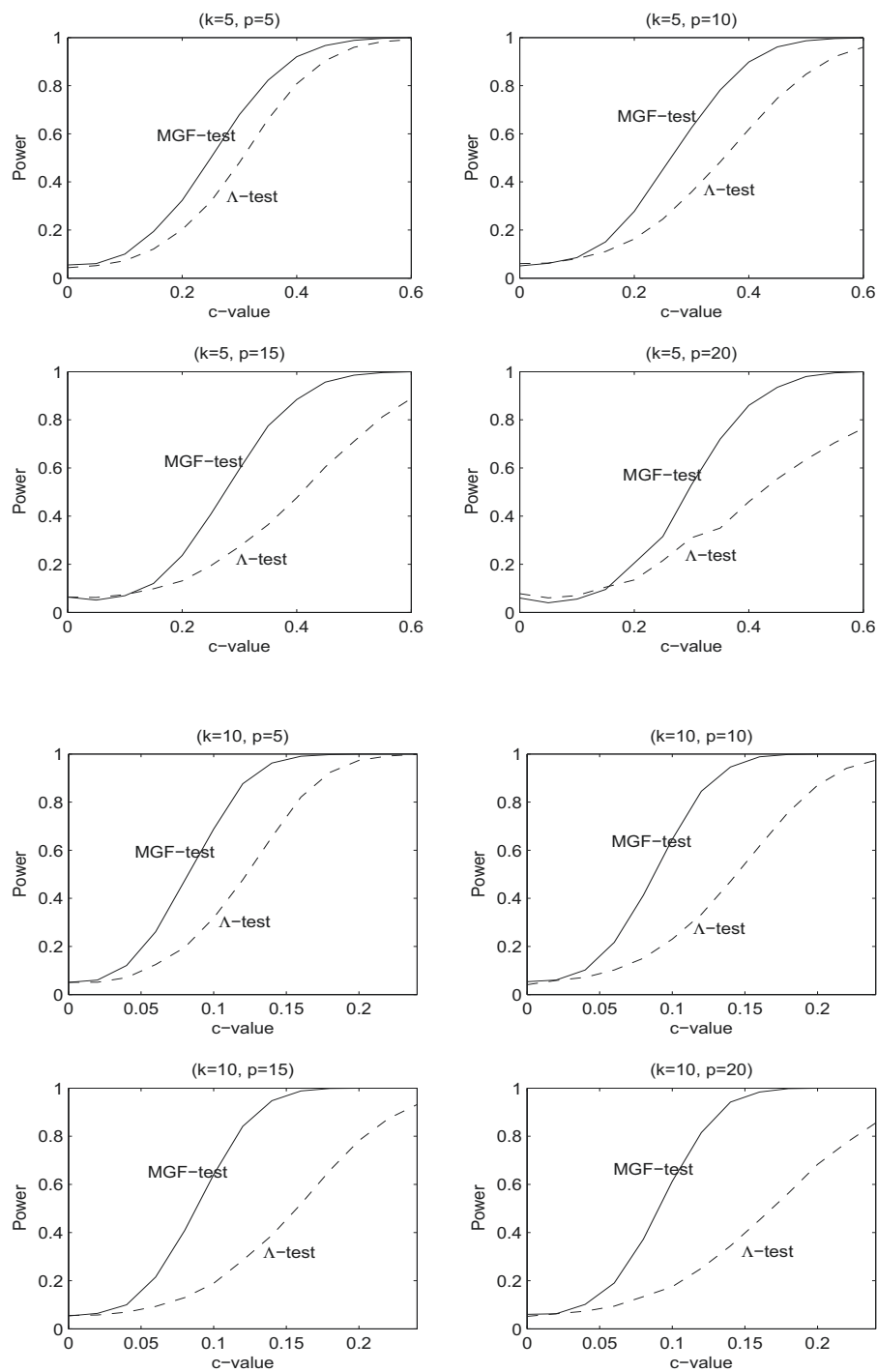
Fig. 2. Power comparison (level $\alpha = 5\%$) between the MGF-test and Wilks Λ-test. In each plot, the population mean difference $\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i = c\mathbf{1}_p$ $(i = 1, \ldots, k-1)$ with the increasing c-values as indicated. k =number of populations, p =population dimension.