

Exact F-tests for the Mean of A Class of Elliptically Contoured Distributions

Jiajuan Liang

College of Business, University of New Haven, West Haven, Connecticut, U.S.A.
Email: jliang@newhaven.edu

Abstract A class of F -tests are developed based on an i.i.d. sample from a subfamily of elliptically contoured distributions. The theory of spherical matrix distributions is employed for constructing the tests. The results from a Monte Carlo study on the empirical performance of the tests show that the tests can control type I error rates very well and have fairly good power performance even for very high dimension with a small sample size. An analysis on a real dataset in financial models illustrates possible applications of the proposed tests.

Keywords: Elliptically contoured distribution, F-test, high dimension, principal component, spherically symmetric distribution

1 Introduction

Testing the mean of a population plays an important role in parametric statistical hypotheses. Traditional tests for the population mean are basically based on the normal assumption. For example, the classical Student's t -test is the popular test for the univariate normal population mean. The Hotelling T^2 -test is the commonly-used statistic for testing the multivariate normal population mean. Both of these two tests assume the sample data from the normal population. Since the last few decades, researchers have found different ways to generalize the multivariate normal distribution to a much wider family of multivariate distributions that contain the normal distribution as a special case. Among these generalized distributions, the family of elliptically contoured distributions (simply called ECD) is one of the most thorough-studied distributional families that possess many similar properties to those of the multivariate normal distribution, see, for example, [1], [2], [3], [4], [5] and [7]. With more and more theoretical properties discovered, ECD has been applied to statistical modeling and high-dimensional data analysis in real-life problems, see [8] for some recent applications.

Following the notations in [5], we use $ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the family of elliptically contoured distributions with the mean vector $\boldsymbol{\mu}$ ($p \times 1$) and covariance matrix $\boldsymbol{\Sigma}$ ($p \times p$). Testing the mean $\boldsymbol{\mu}$ in $ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be reduced to testing the hypothesis

$$H_0 : \boldsymbol{\mu} = \mathbf{0}, \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}. \quad (1)$$

The challenge for constructing statistics for testing this hypothesis is that a distribution in $ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ may have a probability density function. If it has a density function, it is still unknown the exact mathematical expression of the function. Therefore, the traditional likelihood ratio approach is no longer applicable. Fang and Zhang [5] developed an approach to testing hypothesis (1) by assuming that the sample observation matrix has a left spherical matrix distribution, which is not based on an i.i.d. (independently identically distributed) sample. It remains an open problem to test hypothesis (1) with an i.i.d. sample.

In this paper, we try to tackle the problem of testing hypothesis (1) with an i.i.d. sample from a subfamily $ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_p$, where $\sigma > 0$ is unknown and \mathbf{I}_p stands for the $p \times p$ identity matrix. That is, we consider an i.i.d. sample is available from $ECD_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$. The theoretical development for the tests is presented in Section 2 as follows. Section 3 provides a Monte Carlo study on the empirical performance of the tests. Section 4 gives an example for illustrating possible applications of the tests in financial portfolio analysis. Some concluding remarks are summarized in the last section.

2 Theoretical Construction of the Tests

The construction of the tests is related to some simple knowledge in spherical matrix distributions [5]. A $p \times n$ random matrix \mathbf{X} is said to have a left spherical matrix distribution (denote by LSMD for simplicity) for any $p \times p$ constant orthogonal matrix $\mathbf{\Gamma}$, $\mathbf{\Gamma X}$ and \mathbf{X} have the same probability distribution. As illustrated in [5], if \mathbf{X} has an LSMD and $t(\mathbf{X})$ is a statistic defined by \mathbf{X} and $t(\mathbf{X})$ is affine invariant satisfying $t(\mathbf{AX}) = t(\mathbf{X})$ for any $p \times p$ nonsingular constant matrix \mathbf{A} , then $t(\mathbf{X})$ and $t(\mathbf{Y})$ have the same distribution with \mathbf{Y} having a matrix normal distribution $N_{p \times n}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_n)$, where “ \otimes ” stands for the Kronecker product. This property provides the fact that the probability distribution of an affine invariant statistic $t(\mathbf{X})$ remains unchanged in the family consisting of all LSMDs. Note that $N_{p \times n}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_n)$ is an LSMD. This implies that the distribution of $t(\mathbf{X})$ can be obtained by considering \mathbf{X} has a matrix normal distribution $N_{p \times n}(\mathbf{0}, \mathbf{I}_p \otimes \mathbf{I}_n)$. Then we have the following theorem.

Theorem 1. Let $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be an i.i.d. sample from $ECD_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ ($p \times n$). Assume that the probability $P(\mathbf{X} = \mathbf{0}) = 0$. Define the eigenvalue-eigenvectors of $\mathbf{X}'\mathbf{X}$ as

$$\frac{1}{p}(\mathbf{X}'\mathbf{X}) = \lambda_i \mathbf{d}_i, \quad (2)$$

associated with some positive eigenvalues $\lambda_1 > \dots > \lambda_r > 0$ for some $r \leq \min(n, p) - 1$ ($i = 1, \dots, r$). For $q = 1, \dots, r$, define the matrices

$$\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_q), \quad \mathbf{Z} = \mathbf{XD}, \quad \mathbf{H} = \mathbf{Z}' \left(\frac{1}{p} \mathbf{1}_p \mathbf{1}_p' \right) \mathbf{Z}, \quad \mathbf{G} = \mathbf{Z}' \left(\mathbf{I}_p - \frac{1}{p} \mathbf{1}_p \mathbf{1}_p' \right) \mathbf{Z}, \quad (3)$$

where $\mathbf{1}_p$ stands for the $p \times 1$ vector of ones. Construct the series of statistics

$$F_q = \frac{p-q}{q} \text{trace}(\mathbf{HG}^{-1}) = \frac{p-q}{pq} \mathbf{1}_p' \mathbf{Z}' \mathbf{G}^{-1} \mathbf{Z} \mathbf{1}_p, \quad (4)$$

for $q = 1, \dots, r$. Then, under the null hypothesis in (1), F_q has an exact F -distribution $F(q, p-q)$ with degrees of freedom $(q, p-q)$.

Proof. Note that under the null hypothesis in (1), each observation \mathbf{x}_i ($i = 1, \dots, n$) in the i.i.d. sample $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ has an $ECD(\mathbf{0}, \sigma^2 \mathbf{I}_p)$, which reduces to a spherically symmetric distribution [5]. This results in the fact that $\mathbf{\Gamma x}_i$ and \mathbf{x}_i have the same probability distribution for any $p \times p$ constant orthogonal matrix $\mathbf{\Gamma}$ for $i = 1, \dots, n$. Because $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ is an i.i.d. sample, this gives the fact that $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and $\mathbf{\Gamma X} = (\mathbf{\Gamma x}_1, \dots, \mathbf{\Gamma x}_n)$ have the same distribution. This implies that under the null hypothesis (1), the random matrix \mathbf{X} has an LSMD. The construction of the statistics F_q in (4) is following the same way to construct an F -statistic based on an i.i.d. multivariate normal sample given by [9]. Following the same thought as in Theorem 2 of [9], we can easily arrive at the conclusion that F_q has an exact F -distribution $F(q, p-q)$ with degrees of freedom $(q, p-q)$. This completes the proof.

The statistics F_q ($q = 1, \dots, r$) can be applied to test hypothesis (1). If the population distribution $ECD_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$ has a probability density function, it can be proved that the number of positive eigenvalues $r = \min(n, p) - 1$. For any given significance level $0 < \alpha < 1$, one should reject the null hypothesis in (1) if $F_q > F(q, p-q; 1-\alpha)$, where $F(q, p-q; 1-\alpha)$ stands for the $100(1-\alpha)$ -percentile of the F -distribution $F(q, p-q)$ for any choice of $q = 1, \dots, r$. We will carry out a Monte Carlo study on the empirical performance of different choices of q in the next section.

3 A Monte Carlo Study

3.1 Empirical Type I Error Rates

The following subfamilies of $ECD_p(\boldsymbol{\mu}, \mathbf{I}_p)$ ($\sigma = 1$) (Chapter 3 in [4]) are chosen for simulating the empirical type I error rates by using MATLAB code, where the C_i ($i = 1, 2, 3, 4$) are suitable normalizing constants.

(1) The multivariate t-distribution has a density function of the form

$$f_t(\|\mathbf{x} - \boldsymbol{\mu}\|) = C_1 \left(1 + \frac{\|\mathbf{x} - \boldsymbol{\mu}\|^2}{m}\right)^{-\frac{p+m}{2}}, \quad m > 0,$$

where “ $\|\cdot\|$ ” stands for the Euclidean norm of a vector, let $m = 5$;

(2) the Kotz type distribution $Kotz(N, m, s)$ with parameters $N = 2$, $m = 1$ and $s = 0.5$ has a density function of the form

$$f_k(\|\mathbf{x} - \boldsymbol{\mu}\|) = C_2 \|\mathbf{x} - \boldsymbol{\mu}\|^{2(N-1)} \exp\{-r\|\mathbf{x} - \boldsymbol{\mu}\|^{2s}\},$$

where N , r and s are parameters. Let $N = 2$, $r = 1$ and $s = 0.5$.

(3) The Pearson type II distribution with the density function of the form

$$f_{P2}(\|\mathbf{x} - \boldsymbol{\mu}\|) = C_3 (1 - \|\mathbf{x} - \boldsymbol{\mu}\|^2)^{-m},$$

where $m > -1$ is a parameter. Let $m = 3/2$.

(4) The multivariate Cauchy distribution has a density function of the form

$$f_{Cauchy}(\|\mathbf{x} - \boldsymbol{\mu}\|) = C_4 (1 + \|\mathbf{x} - \boldsymbol{\mu}\|^2)^{-\frac{p+1}{2}}.$$

An i.i.d. sample from each of these distributions can be easily generated by MATLAB code (available from the authors upon request). We summarize the empirical type I error rates ($\boldsymbol{\mu} = \mathbf{0}$), which are the percentages of rejection under 2,000 replications of simulation, where the four statistics of F_q defined by (4) are chosen to see the impact of the choice of q . The number q is actually the number of principal component (PC) directions defined by (2).

- (1) Statistic $F_1 \sim F(1, p - 1)$ with $q = 1$, only one PC direction is chosen;
- (2) Statistic $F_{[p/3]} \sim F([p/3], p - [p/3])$, where $[\cdot]$ stands for the integer part of a real number;
- (3) Statistic $F_{[p/2]} \sim F([p/2], p - [p/2])$;
- (4) Statistic $F_r \sim F(r, p - r)$ with $r = \min(n, p) - 1$ being the maximum possible number of PC directions.

Table 1. Empirical type I error rates for the multivariate t-distribution

	The multivariate t-distribution								
	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
	$p = 10$	$p = 20$	$p = 30$	$p = 10$	$p = 20$	$p = 30$	$p = 10$	$p = 20$	$p = 30$
F_1	0.0070	0.0075	0.0105	0.0460	0.0485	0.0510	0.1025	0.0985	0.1000
$F_{[p/3]}$	0.0095	0.0065	0.0080	0.0520	0.0560	0.0390	0.0940	0.1080	0.0835
$F_{[p/2]}$	0.0080	0.0080	0.0080	0.0530	0.0545	0.0470	0.0980	0.0995	0.1010
F_r	0.0100	0.0060	0.0095	0.0475	0.0395	0.0510	0.0955	0.1000	0.0970
	The Kotz-type distribution								
F_1	0.0120	0.0115	0.0095	0.0540	0.0565	0.0500	0.1060	0.1080	0.0985
$F_{[p/3]}$	0.0075	0.0080	0.0120	0.0420	0.0470	0.0510	0.1050	0.1055	0.0990
$F_{[p/2]}$	0.0115	0.0075	0.0090	0.0550	0.0515	0.0515	0.0970	0.0990	0.1020
F_r	0.0090	0.0095	0.0110	0.0530	0.0415	0.0490	0.0950	0.0830	0.1035
	The Pearson type II distribution								
F_1	0.0060	0.0085	0.0070	0.0455	0.0445	0.0475	0.0845	0.0955	0.1030
$F_{[p/3]}$	0.0080	0.0075	0.0110	0.0460	0.0470	0.0465	0.0970	0.0975	0.0975
$F_{[p/2]}$	0.0100	0.0125	0.0100	0.0570	0.0535	0.0485	0.1055	0.1110	0.1010
F_r	0.0115	0.0120	0.0140	0.0460	0.0420	0.0490	0.0880	0.0795	0.0995
	The Cauchy-type distribution								
F_1	0.0080	0.0120	0.0065	0.0510	0.0430	0.0505	0.0975	0.0935	0.0980
$F_{[p/3]}$	0.0100	0.0155	0.0090	0.0580	0.0600	0.0465	0.1090	0.1120	0.0955
$F_{[p/2]}$	0.0105	0.0120	0.0105	0.0560	0.0640	0.0520	0.1080	0.1015	0.0970
F_r	0.0100	0.0090	0.0135	0.0480	0.0545	0.0450	0.0980	0.1045	0.0990

Table 1 summarizes the empirical type I error rates for the four chosen statistics and the four chosen subfamilies $ECD_p(\mathbf{0}, \mathbf{I}_p)$. The results show that the statistics can control the type I error very well, even for the dimension as high as $p = 30$ and the sample size as small as $n = 20$ ($n < p$).

3.2 Empirical Power

The empirical power is computed by percentage of rejection under 2,000 replications of simulation by using the same four subfamilies of ECD as in Table 1 above. Without loss of generality, the mean vector $\boldsymbol{\mu}$ is chosen as $\boldsymbol{\mu} = c\mathbf{1}_p$ and let the constant c increase from $c = 0$ to $c=1$ with an increment of 0.05, which means that taking the step length of 0.05, let c increase from 0 to 1 so that there are 21 c -values: 0, 0.05, 0.1, ..., 1 in computing the percentage of rejections by each statistic as used in Table 1. The power is computed by taking the significance level $\alpha = 0.05$. In order to compare the impact of the choice of q on the power performance, we plot the power values versus the c -values: 0, 0.05, 0.1, ..., 1 for each statistic in Figure 1 below.

The following two empirical conclusions can be summarized.

- (1) The power performance of the two statistics $F_{[q/3]}$ and $F_{[q/2]}$ seems to be the best in most cases. One could expect to have better power performance when choosing the number of PC-directions between $[p/3]$ and $[p/2]$ to construct the exact F -test. Choosing the maximum possible number of PC-directions simply results in loss of power as indicated by the line "*" in most cases;
- (2) Under the same sample size n , an increase of the dimension p leads to an increase of the power. This is because the construction of the F -tests in Theorem 1 reverses the traditional $n \times p$ observation matrix into a $p \times n$ observation matrix. As a result, the dimension p acts as the sample size in the construction of the F -tests. Therefore, these F -tests are especially applicable to the case of high dimension with a small sample size in real problems.

4 An Illustrative Example

The value-weighted New York Stock Exchange return data are available from the Center for Research in Security Prices (CRSP) at the University of Chicago. We choose a partial data set (available from the authors upon request) that contains the market monthly returns between the year of 1966 and 1975 (120 months) from a portfolio with 32 stocks. First, we run the CAPM (capital asset pricing model, see [6] and [10]) to get the monthly risk-adjusted returns. These risk-adjusted returns constitute a sample with a sample size $n = 120$ and dimension $p = 32$. Second, we run a 2-factor analysis model to get another set of monthly risk-adjusted returns. These risk-adjusted returns also constitute a sample with a sample size $n = 120$ and dimension $p = 32$. Now we have two populations and two sets of samples:

- (1) Population #1: risk-adjusted returns computed by CAPM, we have sample #1 with $n = 120$ and $p = 32$;
- (2) Population #2: risk-adjusted returns computed by a 2-factor model, we have sample #2 with $n = 120$ and $p = 32$.

In the theory of CAPM, an investment portfolio is said to be efficient if it satisfies two conditions: 1) it reaches the smallest possible variance given its expected return; and 2) it reaches the largest possible expected return given its variance. The verification of these two conditions is finally reduced to testing the null hypothesis that the intercept parameter in the CAPM is equal to zero (vector) versus the alternative hypothesis that the intercept parameter in the CAPM is not equal zero. In the CAPM, the observations are the risk-adjusted returns. The purpose of constructing populations #1 and #2 is to double check the efficiency of the portfolio consisting of 32 stocks. Assuming population #1 has an $ECD(\boldsymbol{\mu}_1, \sigma_1^2 \mathbf{I}_p)$ and population #2 has an $ECD(\boldsymbol{\mu}_2, \sigma_2^2 \mathbf{I}_p)$, we set up the hypotheses:

$$H_0 : \boldsymbol{\mu}_1 = \mathbf{0}, \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \mathbf{0}, \quad (5)$$

and

$$H_0 : \boldsymbol{\mu}_2 = \mathbf{0}, \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_2 \neq \mathbf{0}. \quad (6)$$

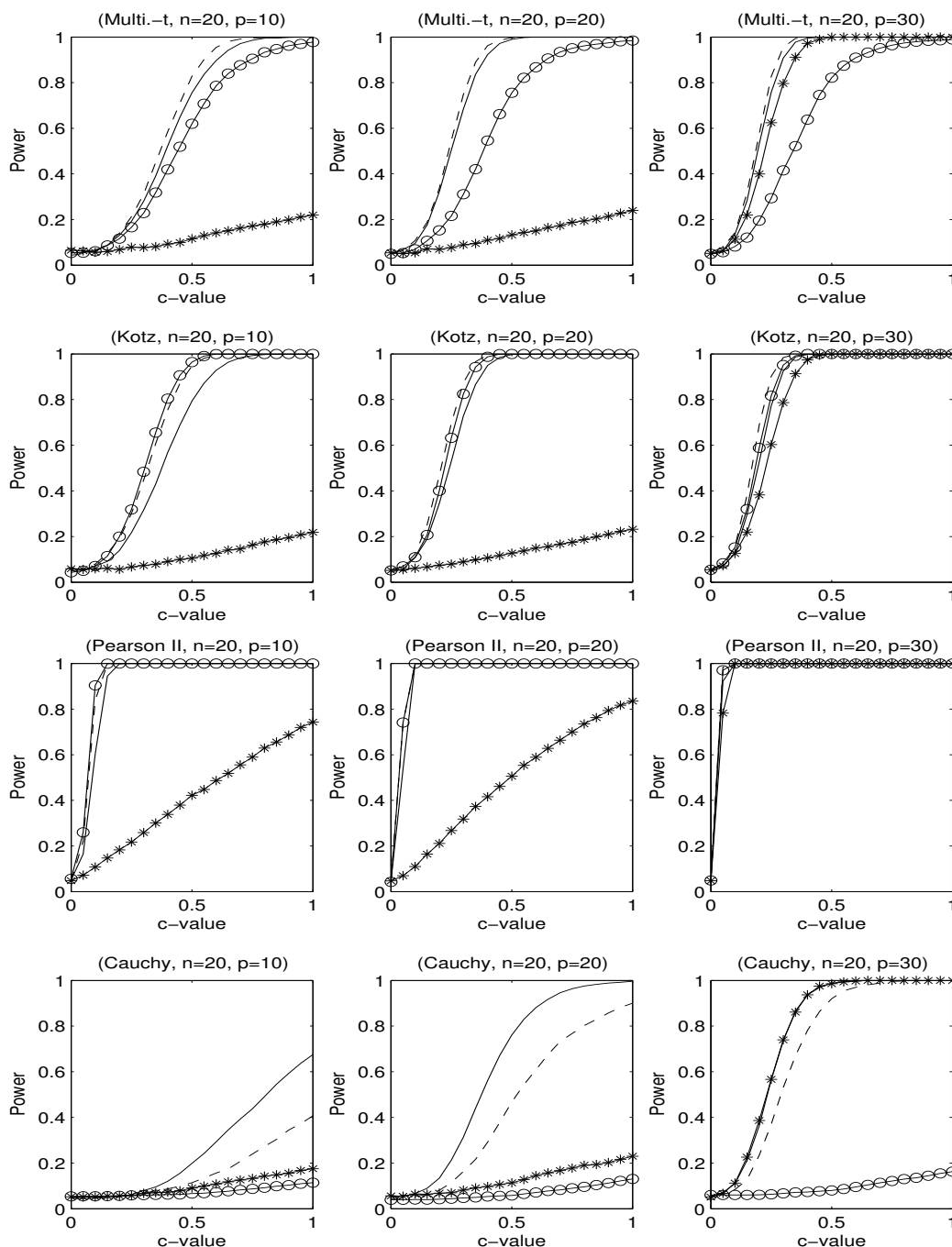


Figure 1. Power comparison between different choices of q : 1) the “-o-” line is for F_1 ($q = 1$); 2) the dashed line “- -” is for $F_{[q/3]}$ ($q = [q/3]$); 3) the real line “-” is for $F_{[q/2]}$ ($q = [q/2]$); and 4) the “*-” line is for F_r (r =the maximum possible of the number of positive eigenvalues for each case).

The F_q -statistics for all possible q -values are chosen for the F_q -tests. The p -values of the tests are summarized in Table 2. Most of the p -values are greater than 0.10, implying most of F_q -tests are insignificant. Therefore, we should not reject the two null hypotheses in (5) and (6). That is, the portfolio consisting of 32 stocks can be considered efficient during the 10 years 1966-1975.

Table 2. p -values from the F_q -tests for all possible q in the real-data example.

F_q	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$	$q = 6$	$q = 7$	$q = 8$
Population #1	0.0000	0.0000	0.0001	0.0003	0.0002	0.0005	0.0012	0.0028
Population #2	0.0857	0.2227	0.1734	0.0769	0.1399	0.2264	0.2831	0.3485
F_q	$q = 9$	$q = 10$	$q = 11$	$q = 12$	$q = 13$	$q = 14$	$q = 15$	$q = 16$
Population #1	0.0045	0.0095	0.0172	0.0306	0.0473	0.0750	0.1052	0.1579
Population #2	0.4388	0.5451	0.6103	0.6985	0.7599	0.7902	0.8564	0.8580
F_q	$q = 17$	$q = 18$	$q = 19$	$q = 20$	$q = 21$	$q = 22$	$q = 23$	$q = 24$
Population #1	0.2195	0.2965	0.3741	0.4774	0.5749	0.6753	0.7680	0.8224
Population #2	0.9061	0.8242	0.8389	0.8942	0.8877	0.8916	0.8922	0.9009
F_q	$q = 25$	$q = 26$	$q = 27$	$q = 28$	$q = 29$	$q = 30$	$q = 31$	
Population #1	0.8877	0.9344	0.9542	0.9782	0.9913	0.9942	0.9985	
Population #2	0.9256	0.9234	0.7271	0.3340	0.4259	0.5439	0.7703	

5 Concluding Remarks

The exact F -tests in this paper provides a way to test the mean of some subfamilies of ECD under i.i.d. samples without assuming existence of the probability density function for the population. This is substantially different from many existing approaches to constructing parametric tests. The construction of the F -tests in Theorem 1 considers the population dimension p as the sample size. This makes the F -tests particularly suitable for the case of high dimension with a small sample size. The Monte Carlo study supports this assertion. The real-data example illustrates possible applications of the F -tests in high-dimensional data analysis where the normal assumption may not be appropriate and the number of observations may be very limited due to high cost or difficulty in obtaining data. The exact F -tests are applicable for all population distributions in $ECD_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$. This includes the normal distribution $N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$ as a special case. Therefore, the exact F -tests in this paper could find more applications in some areas like medical research where it is common to have data of high dimension with a small sample size and the normal assumption may be violated.

References

- 1 S. Cambanis, S. Huang and G. Simons, "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, vol. 11, pp. 368–385, 1981.
- 2 A. P. Dawid, "Spherical matrix distributions and a multivariate model" *Journal of the Royal Statistical Society (Series B)*, vol. 39, pp. 254–261, 1977.
- 3 A. P. Dawid, "Extendibility of spherical matrix distributions," *Journal of Multivariate Analysis*, vol. 8, pp. 559–566, 1978.
- 4 K. T. Fang, S. Kotz and K. W. Ng, *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London and New York, 1990.
- 5 K. T. Fang and Y. Zhang, *Generalized Multivariate Analysis*, Springer-Verlag and Science Press, Berlin/Beijing, 1990.
- 6 M. R. Gibbons, S. Ross and J. Shanken, "A test of efficiency of a given portfolio," *Econometrica*, vol. 57, pp. 1121–1152, 1989.
- 7 A. K. Gupta and T. Varga, *Elliptically Contoured Models in Statistics*, Kluwer Academic Publishers, 1993.
- 8 A. K. Gupta, T. Varga and T. Bodnar, *Elliptically Contoured Models in Statistics and Portfolio Theory (Second Edition)*, Springer, 2013.
- 9 J. Läuter, "Exact t and F tests for analyzing studies with multiple endpoints," *Biometrics*, vol. 52, pp. 964–970, 1996.
- 10 C. A. MacKinlay, "On multivariate tests of the CAPM," *Journal of Financial Economics*, vol. 18, pp. 341–372, 1987.