ASSESSING THE DISTRIBUTIONAL ASSUMPTIONS
IN ONE-WAY REGRESSION MODEL

Dhanuja Kasturiratna

A Dissertation
Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2006

Committee:
Arjun K. Gupta, Coadvisor
Truc T. Nguyen, Coadvisor
John T. Chen
Patricia E. Sharp
Graduate Faculty Representative
ABSTRACT

Arjun K. Gupta, Truc T. Nguyen Advisors

Characterization of the normal distribution related to $k$ samples based on regression is given. This characterization has been transformed to a characterization based on Student’s law. With the help of these characterization results, the composite hypothesis of testing the distributional assumptions in one-way regression model can be replaced by an equivalent simple hypothesis. This simple hypothesis can then be tested using any of the EDF goodness-of-fit tests. The powers of these tests are studied using Monte Carlo methods.

Multivariate normal distribution, with covariance matrix of the form $\sigma^2 \Sigma_0$, is characterized based on UMVU estimator of the density function. Using this result with the transformation proposed by Rincon-Gallardo, Quesenberry and O’Reilly (1979), the composite hypothesis of testing $k$-variate normality with covariance matrix $\sigma^2 \Sigma_0$ has been transformed to an equivalent simple hypothesis. It is shown that the transformation proposed here can also be used in changing the above composite hypothesis to an equivalent simple hypothesis. These transformations are compared using Monte Carlo methods.

Approximate EDF goodness of fit tests for testing the distributional assumptions in one-way regression model are studied using the Monte Carlo simulations.
ACKNOWLEDGMENTS

I would like to thank Professors Arjun K. Gupta and Truc T. Nguyen, my advisors, for their constant support and many suggestions throughout this research. I thank the other members of my committee, Professors John T. Chen and Patricia Sharp, for their valuable time and advice. I also extend special thanks to all my professors in the Department of Mathematics and Statistics at BGSU and Dr. W. N. Wickramasinghe, Dr. Chandanie Perera, who inspired me to continue studying mathematics and statistics.

I am grateful to the Department of Mathematics and Statistics for the teaching assistantship and non-service fellowship awarded to me during my study at BGSU, which were crucial for the completion of this work.

My deepest gratitude goes to my parents for their encouragement and to my family Manori, Manisha and Dunil for their love and patience.

Finally, I wish to thank my colleagues and friends Mirror, Carmelita and Igor for all the help and valuable comments.

Bowling Green, Ohio

Dhanuja Katuriratna

August, 2006
# TABLE OF CONTENTS

## CHAPTER 1: INTRODUCTION

1.1 Motivation ......................................................... 1

1.2 Overview of EDF goodness-of-fit tests .......................... 1

1.3 Literature .......................................................... 4

1.4 Summary .......................................................... 5

## CHAPTER 2: CHARACTERIZATION OF NORMAL DISTRIBUTION RELATED TO TWO SAMPLES BASED ON REGRESSION

2.1 Introduction ....................................................... 7

2.2 Characterization: Two sample case ............................ 8

2.3 Characterization: \( k \) sample case .......................... 18

2.4 Application to goodness-of-fit tests ........................... 20

## CHAPTER 3: EDF GOODNESS-OF-FIT TESTS FOR TESTING THE DISTRIBUTIONAL ASSUMPTIONS IN ONE-WAY REGRESSION MODEL

3.1 Introduction ....................................................... 22

3.2 Characterization of normal distribution related to two samples based on Student’s law ........................................... 23

3.3 Characterization of normal distribution related to \( k \) samples based on Student’s law ........................................... 29
CHAPTER 4: CHARACTERIZATION OF $k$-VARIATE NORMAL DISTRIBUTION WITH COVARIANCE STRUCTURE, $\sigma^2 \Sigma_0$ AND EDF GOODNESS-OF-FIT TESTS

4.1 Introduction .................................................. 47
4.2 Characterization result ...................................... 48
4.3 Application to goodness-of-fit tests ....................... 55
4.4 Power study .................................................. 56

CHAPTER 5: APPROXIMATE EDF GOODNESS-OF-FIT TESTS FOR TESTING THE DISTRIBUTIONAL ASSUMPTIONS OF ONE-WAY REGRESSION MODEL

5.1 Introduction .................................................. 68
5.2 Procedure of computing percentage points for approximate EDF statistics . 69
5.3 Percentage points for approximate EDF statistics ............ 70
5.4 Application to goodness-of-fit tests ....................... 75
5.5 Power study .................................................. 76

REFERENCES .................................................. 79

Appendix: CALCULATING THE PERCENTAGE POINTS OF APPROXIMATE EDF STATISTICS 82
LIST OF TABLES

3.1 Estimated Type-I error of the tests, out of 10000 iterations, $\mu_1 = 1, \mu_2 = 5$. 35
3.2 Estimated Type-I error of the tests, out of 10000 iterations, $\mu_1 = 1, \mu_2 = 10$. 36
3.3 Number of rejections out of 10000 - alternative is exponential distribution,
   $\mu_1 = 1, \mu_2 = 5$ ......................................................... 37
3.4 Number of rejections out of 10000 - alternative is exponential distribution,
   $\mu_1 = 1, \mu_2 = 10$ ......................................................... 38
3.5 Number of rejections out of 10000 - alternative is lognormal distribution,
   $\mu_1 = 1, \mu_2 = 5$ ......................................................... 39
3.6 Number of rejections out of 10000 - alternative is lognormal distribution,
   $\mu_1 = 1, \mu_2 = 10$ ......................................................... 40
3.7 Number of rejections out of 10000 - alternative is location shifted Student’s $t$
   distribution, $\mu_1 = 1, \mu_2 = 5$ ........................................... 41
3.8 Number of rejections out of 10000 - alternative is location shifted Student’s $t$
   distribution, $\mu_1 = 1, \mu_2 = 10$ ........................................... 42
3.9 Number of rejections out of 10000 - samples are from different distributions
   $\mu_1 = 1, \mu_2 = 5, \sigma^2 = 10$ .......................................... 43
3.10 Number of rejections out of 10000 - samples are from normal distribution but
     having different variances, $\mu_1 = 1, \mu_2 = 5$ ...................... 44
3.11 Number of rejections out of 10000 - unequal sample sizes, $\mu_1 = 1, \mu_2 = 5, \sigma^2 = 10$ 45
3.12 Number of rejections out of 10000 - three samples case $\mu_1 = 1, \mu_2 = 5, \mu_3 = 10, \sigma^2 = 10$.

4.1 Estimated Type-I error of the tests, out of 10000 iterations, when $k=2$.

4.2 Estimated Type-I error of the tests, out of 10000 iterations, when $k=3$.

4.3 Number of rejections out of 10000 - alternative is multivariate $t$ distribution, $k=2$.

4.4 Number of rejections out of 10000 - alternative is multivariate $t$ distribution, $k=3$.

4.5 Number of rejections out of 10000 - alternative is mixture of multivariate normal distributions, mixing proportion $p = 0.25, \mu_1 = [1 \ 3]', \mu_2 = [20 \ 60]'$, $k=2$.

4.6 Number of rejections out of 10000 - alternative is mixture of multivariate normal distributions, mixing proportion $p = 0.25, \mu_1 = [1 \ 3 \ 5]', \mu_2 = [20 \ 60 \ 100]'$, $k=3$.

4.7 Number of rejections out of 10000 - alternative is Exponential distribution, $k=2$.

4.8 Number of rejections out of 10000 - alternative is Exponential distribution, $k=3$.

4.9 Number of rejections out of 10000 - alternative is multivariate normal distribution with $\Sigma \neq \sigma^2 \Sigma_0$, for some $\Sigma_0$, $k=2$.

4.10 Number of rejections out of 10000 - alternative is multivariate normal distribution with $\Sigma \neq \sigma^2 \Sigma_0$, for some $\Sigma_0$, $k=3$. 


5.1 Percentage points for testing normality and homogeneity of two samples when
the parameters are unknown. ........................................... 72

5.2 Number of rejections out of 10000 iterations ............................... 78
CHAPTER 1
INTRODUCTION

1.1 Motivation

In the one-way classification with \( k \) treatments, we assume that the data are observed according to the additive model, where the error random variables are independently and identically distributed normally with mean zero and constant variance. To test whether a set of observed data are coming from the above regression model, we need to construct a test for testing the hypothesis that the observations of the treatment groups are normally distributed with same variance. A common practice of testing this hypothesis is to first test the normality of each sample separately and then test the equality of variance. In this case, it would be inherently flawed because in each test we accept 5% chance of our conclusion being wrong (when we test for \( \alpha = 0.05 \)). So in many tests we would expect (by probability) that one test would give us a false result. Therefore, proposing a test to test the normality and homogeneity of variance of the samples simultaneously would be important since it keeps the Type I error of the test as desired.

1.2 Overview of EDF goodness-of-fit tests

Suppose a given random sample of size \( n \) is \( X_1, \ldots, X_n \) and let \( X_{(1)}, \ldots, X_{(n)} \) be the order statistics. Suppose further that the distribution of \( X \) is \( F(x) \). The empirical distribution
function (EDF), \( F_n(x) \) is defined by
\[
F_n(x) = \begin{cases} 
0 & x < X_{(1)}, \\
\frac{i}{n} & X_{(i)} \leq x < X_{(i+1)}, \ i = 1, \ldots, n-1, \\
1 & X_{(n)} \leq x.
\end{cases}
\]

Thus \( F_n(x) \) is a step function, calculated from the sample, as \( x \) increases it takes a step up of height \( 1/n \) as each sample observation is reached. We can expect \( F_n(x) \) to estimate \( F(x) \), and it is in fact a consistent estimator of \( F(x) \); as \( n \to \infty \), \( |F_n(x) - F(x)| \) decreases to zero with probability one.

A statistic measuring the distance between \( F_n(x) \) and \( F(x) \) is called an EDF statistic. Some examples are

Kolmogorov-Smirnov statistic, Kolmogorov (1933)
\[
D = \sup_x | F_n(x) - F(x) |
\]

Kuiper statistic, Kuiper (1966)
\[
V = \sup_x \{F_n(x) - F(x)\} + \sup_x \{F(x) - F_n(x)\}
\]

Cramer-von Mises statistic
\[
W^2 = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 dF(x)
\]

Anderson-Darling statistic
\[
A^2 = n \int_{-\infty}^{\infty} \{F_n(x) - F(x)\}^2 [\{F(x)\}[1 - F(x)]]^{-1} dF(x)
\]

Computing formulas for the above statistics can be found by using the Probability Integral Transformation (PIT), \( Z = F(X) \); when \( F(x) \) is the true distribution of \( X \), the new random
variable $Z$ is uniformly distributed between 0 and 1. Then $Z$ has distribution function $F^*(z) = z$, $0 \leq z \leq 1$. Suppose that a sample $X_1, \ldots, X_n$ gives values $Z_i = F(X_i)$, $i = 1, \ldots, n$, and let $F_n^*(z)$ be the EDF of the values $Z_i$. EDF statistics can now be calculated from a comparison of $F_n^*(z)$ with $F^*(z)$. It can be shown that the following relationship holds:

$$F_n(x) - F(x) = F_n^*(z) - F^*(z) = F_n^*(z) - z.$$

This leads to the following computing formulas;

$$D = \max\{\max_i\{i/n - Z(i)\}, \max_i\{Z(i) - (i - 1)/n\}\}$$

$$V = \max_i\{i/n - Z(i)\} + \max_i\{Z(i) - (i - 1)/n\}$$

$$W^2 = \sum_i\{Z(i) - (2i - 1)/(2n)\}^2 + 1/(12n)$$

$$A^2 = -n - (1/n) \sum_i(2i - 1)[\ln Z(i) + \ln\{1 - Z(n+1-i)\}]$$

All these formulas are very straightforward to calculate.

The general test of fit is a test of

$$H_0 : \text{a random sample } X_1, \ldots, X_n \text{ comes from } F(x; \theta)$$

where $\theta$ is a vector of parameters. For the case that $\theta$ is fully specified, the distribution theory of EDF statistics is well developed (see D’Agostino and Stephens (1986)). When $\theta$ contains one or more unknown parameters, these parameters may be replaced by estimates, to give $\hat{\theta}$ as the estimate of $\theta$. Then the above computation formulas may still be used, with $Z(i) = F(X(i); \hat{\theta})$. However, even when $H_0$ is true, the $Z(i)$ will now not be an ordered uniform sample, and the distribution of EDF statistics will be very different from those cases for which the $\theta$ is fully specified. They will depend on the distribution tested, the parameters estimated, method of estimation as well as on the sample size. The exact distributions of
EDF statistics are very difficult to find.

When $\theta$ is not fully specified, if we can find a transformation which transforms the observed random variables $X_1, \ldots, X_n$ to $Y_1, \ldots, Y_m, m < n$, in such a way, that the joint distribution of $Y$’s is completely known, then the exact EDF goodness-of-fit tests can be performed on these transformed variables. Just finding a transformation is not enough. We need to show $X_1, \ldots, X_n$ follow $F_0(x, \theta)$ if and only if $Y_1, \ldots, Y_m$ follow $G(y_1, \ldots, y_m)$, where $G$ is fully specified. This enables us to replace the composite hypothesis by an equivalent simple hypothesis. This problem is a characterization problem in statistics.

1.3 Literature

Normal distribution has been characterized based on the conditional moments by several authors. Rao (1967) gave characterization of normal distribution based on conditional expected value of one linear statistic given two other linear statistics. Singh and Oliker (1979) studied minimum variance unbiased estimation and characterization of densities of several popular distributions, and proposed a conjecture on how to characterize a normal distribution based on the UMVU estimate of its density function. Gupta and Varga (1990) studied the relationship between joint density and conditional densities. Gupta, Nguyen and Wang, (1997) also gave characterization of normal and gamma distributions via conditional structure. Nguyen and Dinh (1998) gave characterization of normal distribution based on the third and fourth conditional moments.

Characterizations supporting exact EDF goodness-of-fit tests have been given by several authors. These characterizations can be used to replace the composite hypothesis of testing the pre-specified distribution by an equivalent simple hypothesis. Kotlarsky (1966)
characterized the normal distribution by Student’s law. Csargo, Seshadri and Yalovsky (1973) proposed exact EDF goodness-of-fit tests for normality in the presence of unknown parameters. Nguyen and Dinh (2003) gave characterization of normal distribution based on the UMVU estimator of the density function which can be used as a transformation in the procedure to construct an EDF goodness-of-fit test for testing normality of a distribution.

For testing multivariate normality, Mardia (1970) proposed tests based on multivariate skewness and kurtosis. Rincon-Gallardo, Quesenberry and O’Reilly (1979) gave transformations which can be used to construct exact EDF goodness-of-fit tests for multivariate normality. Characterization of $k$-variate normal distribution based on the conditional moments is studied by many authors in recent years. Gupta and Varga (1992) gave characterization of matrix variate normality through conditional distributions. Nguyen, Nguyen and Dinh (2004) gave a characterization of $k$-variate normal distribution based on the UMVU estimate of the density function.

Approximate EDF tests (estimating the unknown parameters by sample estimates) for testing the normality have been studied by several authors. The percentage points for EDF statistics, Cramer-von Mises ($W^2$), and Anderson-Darling ($A^2$), were given by Stephens (1974, 1976). Monte carlo studies for Kolmogorov-Smirnov statistic ($D$) were given by Van Soest (1967), by Lilliefors (1967), and by Stephens (1974). The percentage points for the Kuiper statistics ($V$) were given by Loueter and Koerts (1970).

**1.4 Summary**

In this dissertation, in Chapter 2, we give characterization of normal distribution related two samples based on second conditional moments. Then the characterization results are
extended to \( k \) samples. In Chapter 3, the above characterization results have been changed to characterizations based on Student’s law. The applications of these characterization results in assessing the distributional assumptions in one-way regression model are discussed. Monte Carlo simulations are used to compute the power of EDF goodness-of-fit tests for testing the normality and homogeneity of variance of the samples.

In Chapter 4, we give a characterization of \( k \)-variate normal distribution, with partially known covariance, based on UMVU estimator of the density function. Then this characterization has been changed to a characterization based on Student’s law. An alternative method is also proposed based on the results given in chapters 2 and 3. EDF goodness-of-fit tests for testing \( k \)-variate normality with partially known covariance are studied. Finally the power of the proposed tests for several alternatives are computed.

In Chapter 5, we studied the problem of testing the normality and homogeneity of variance of samples using the approximate EDF goodness-of-fit tests where the unknown parameters are estimated using the samples. Monte Carlo simulations are used to compute the power of the tests for several alternatives.
CHAPTER 2
CHARACTERIZATION OF NORMAL DISTRIBUTION RELATED TO TWO SAMPLES BASED ON REGRESSION

2.1 Introduction

In the one-way classification with two treatments, we assume that the data, $Y_{ij}$, are observed according to the model

$$Y_{ij} = \xi_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, 2,$$

where the error random variables, $\epsilon_{ij}$, are independent and identically normally distributed with mean 0 and variance $\sigma^2$. To test whether a set of observed data, $Y_{ij}$, come from the above regression model, we need to construct a test to test the hypothesis that $Y_{ij}$ are from $N(\xi_i, \sigma^2), \quad j = 1, \ldots, n_i, \quad i = 1, 2$. The motivation of characterization given in this chapter is to find a transformation in the procedure to construct an exact EDF goodness-of-fit test for testing the above hypothesis.

A generalization of the above model is one way classification with $k$ treatments, where we need to test whether the $k$ independent samples, $Y_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k$, can be considered as observed from the linear regression model

$$Y_{ij} = \xi_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k.$$

Characterizations of normal distribution based on the conditional moments have been studied by several authors. Rao(1967) gave characterization of normal distribution based on conditional expected value of one linear statistic given two other linear statistics. Singh

In this chapter, in section 2, we give characterization of normal distribution in two samples based on second conditional moments. The characterization results are extended to $k$ samples in section 3. Applications of these characterization results to construct empirical distribution function (EDF) goodness-of-fit tests for detecting normality of a regression model are discussed in Section 4.

### 2.2 Characterization: Two sample case

Let $X_i, \quad i = 1, \ldots, n$ and $Y_j, \quad j = 1, \ldots, m, \quad m, n \geq 2$, be two independent random samples from two normal distributions with means $\xi$ and $\eta$, respectively, and with common variance $\sigma^2$. Then $Z_1 = \frac{\sum_{i=1}^{n} X_i}{n}, \quad Z_2 = \frac{\sum_{j=1}^{m} Y_j}{m}, \quad Z_3 = \sum_{i=1}^{n} (X_i - Z_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2$, are jointly complete and sufficient statistics for $\xi, \eta$, and $\sigma^2$.

We find the conditional joint distribution of $X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}$ given $Z_1, Z_2, Z_3$. Consider the transformation from $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ to $(X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}, Z_1, Z_2, Z_3)$. 

We have
\[ x_{n-1} + x_n = nz_1 - \sum_{i=1}^{n-2} x_i, \quad (2.2.1) \]
\[ y_m = mz_2 - \sum_{j=1}^{m-1} y_j. \quad (2.2.2) \]

From (2.2.1) and (2.2.2), we obtain
\[ x_{n-1}^2 + x_n^2 = \sum_{i=1}^{n} (x_i - z_1)^2 + nz_1^2 - \sum_{i=1}^{n-2} x_i^2 \]
\[ = z_3 - \sum_{j=1}^{m} (y_j - z_2)^2 + nz_1^2 - \sum_{i=1}^{n-2} x_i^2. \quad (2.2.3) \]

From (2.2.3), we obtain
\[
x_{n-1}x_n = \frac{1}{2} \left[ \left( nz_1 - \sum_{i=1}^{n-2} x_i \right)^2 - x_{n-1}^2 + x_n^2 \right] \\
= \frac{1}{2} \left[ \left( nz_1 - \sum_{i=1}^{n-2} x_i \right)^2 - z_3 + \sum_{j=1}^{m} (y_j - z_2)^2 - nz_1^2 + \sum_{i=1}^{n-2} x_i^2 \right] \\
= \frac{1}{2} \left[ \left( nz_1 - \sum_{i=1}^{n-2} x_i \right)^2 - z_3 + \sum_{j=1}^{m-1} y_j^2 + y_m - mz_2 - nz_1^2 + \sum_{i=1}^{n-2} x_i^2 \right] \\
= \frac{1}{2} \left[ \left( nz_1 - \sum_{i=1}^{n-2} x_i \right)^2 - z_3 + \sum_{j=1}^{m-1} y_j^2 + \left( mz_2 - \sum_{j=1}^{m-1} y_j \right)^2 - mz_2 - nz_1^2 + \sum_{i=1}^{n-2} x_i^2 \right]. \]

Now we have a system of equations in \( x_{n-1} \) and \( x_n \):
\[ x_{n-1} + x_n = a \]
\[ x_{n-1}x_n = b \]

where
\[ a = nz_1 - \sum_{i=1}^{n-2} x_i, \]
\[ b = \frac{1}{2} \left[ \left( nz_1 - \sum_{i=1}^{n-2} x_i \right)^2 - z_3 + \sum_{j=1}^{m-1} y_j^2 + \left( mz_2 - \sum_{j=1}^{m-1} y_j \right)^2 - mz_2 - nz_1^2 + \sum_{i=1}^{n-2} x_i^2 \right]. \]
Solving the above system of equations, we obtain

\[ x_{n-1} = A + B, \]
\[ x_n = A - B \]

where

\[ A = \frac{n z_1 - \sum_{i=1}^{n-2} x_i}{2} \]
\[ B = \frac{1}{2} \sqrt{2 \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( m z_2 - \sum_{j=1}^{m-1} y_j \right)^2 + m z_2^2 + n z_1^2 - \sum_{i=1}^{n-2} x_i^2 \right] \left( n z_1 - \sum_{i=1}^{n-2} x_i \right)^2} \]

Note that the solutions of \( x_n \) and \( x_{n-1} \) can be interchanged. Therefore the jacobian has to be multiplied by 2. Now we obtain the Jacobian of the transformation \((X_1, \ldots, X_n, Y_1, \ldots, Y_m)\) to \((X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}, Z_1, Z_2, Z_3)\):

\[
|J| = 2 \begin{vmatrix} 0 & m & 0 \\ \frac{\partial A}{\partial z_1} + \frac{\partial B}{\partial z_1} & \frac{\partial A}{\partial z_2} + \frac{\partial B}{\partial z_2} & \frac{\partial A}{\partial z_3} + \frac{\partial B}{\partial z_3} \\ \frac{\partial A}{\partial z_1} - \frac{\partial B}{\partial z_1} & \frac{\partial A}{\partial z_2} - \frac{\partial B}{\partial z_2} & \frac{\partial A}{\partial z_3} - \frac{\partial B}{\partial z_3} \end{vmatrix} \\
= -2m \left[ \left( \frac{\partial A}{\partial z_1} + \frac{\partial B}{\partial z_1} \right) \left( \frac{\partial A}{\partial z_2} - \frac{\partial B}{\partial z_2} \right) - \left( \frac{\partial A}{\partial z_3} + \frac{\partial B}{\partial z_3} \right) \left( \frac{\partial A}{\partial z_1} - \frac{\partial B}{\partial z_1} \right) \right] \\
= -2m \left[ -2 \frac{\partial A}{\partial z_1} \frac{\partial B}{\partial z_3} + 2 \frac{\partial B}{\partial z_1} \frac{\partial A}{\partial z_3} \right] \\
= 4m \frac{\partial A}{\partial z_1} \frac{\partial B}{\partial z_3} \\
= mn \frac{1}{\sqrt{2 \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( m z_2 - \sum_{j=1}^{m-1} y_j \right)^2 + m z_2^2 + n z_1^2 - \sum_{i=1}^{n-2} x_i^2 \right] \left( n z_1 - \sum_{i=1}^{n-2} x_i \right)^2}}. \\
\]

Let \( X_{n-2} = (X_1, \ldots, X_{n-2}) \) and \( Y_{m-1} = (Y_1, \ldots, Y_{m-1}) \), then the joint density function of \( X_{n-2}, Y_{m-1}, Z_1, Z_2, Z_3 \) is
We know $Z_1 \sim N(\xi, \sigma^2)$, $Z_2 \sim N(\eta, \sigma^2)$ and $Z_3 \sim \sigma^2 \chi^2_{m+n-2}$ are independent. The conditional joint density function of $X_{n-2}, Y_{m-1}$ given $Z_1, Z_2, Z_3$ is

$$f_{X_{n-2}Y_{m-1}|Z_1Z_2Z_3}(x_{n-2}, y_{m-1} | z_1, z_2, z_3) = K_1(n) \frac{1}{z_3} \left( \frac{m+n-2}{2} \right)^{\frac{m+n-2}{2}} e^{-\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} (x_i - \xi)^2 + \sum_{j=1}^{m} (y_j - \eta)^2 + \sum_{k=1}^{n-2} x_k^2 \right)} \cdot |J|$$

for $(x_{n-2}, y_{m-1}, z_1, z_2, z_3) \in D$ (2.2.4)

where $D = \left\{ (x_{n-2}, y_{m-1}, z_1, z_2, z_3) : \right\}$

$$2 \left( z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( m z_2 - \sum_{j=1}^{m-1} y_j \right)^2 + m z_2^2 + n z_1^2 - \sum_{i=1}^{n-2} x_i^2 \right) > \left( n z_1 - \sum_{i=1}^{n-2} x_i \right)^2 \right\}$$

We know $Z_1 \sim N(\xi, \sigma^2)$, $Z_2 \sim N(\eta, \sigma^2)$ and $Z_3 \sim \sigma^2 \chi^2_{m+n-2}$ are independent. The conditional joint density function of $X_{n-2}, Y_{m-1}$ given $Z_1, Z_2, Z_3$ is

$$f_{X_{n-2}Y_{m-1}|Z_1Z_2Z_3}(x_{n-2}, y_{m-1} | z_1, z_2, z_3) = K_1(n) \frac{1}{z_3} \left( \frac{m+n-2}{2} \right)^{\frac{m+n-2}{2}} e^{-\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} (x_i - \xi)^2 + \sum_{j=1}^{m} (y_j - \eta)^2 + \sum_{k=1}^{n-2} x_k^2 \right)} \cdot |J|$$

for $(x_{n-2}, y_{m-1}, z_1, z_2, z_3) \in D$ (2.2.5)

where $K_1(n) = \frac{\sqrt{mn}}{\pi} \Gamma\left(\frac{m+n-2}{2}\right)$.

Fix $X_1$ and let

$$Z'_1 = \frac{nZ_1 - X_1}{n-1}$$

$$Z'_3 = \sum_{i=2}^{n} (X_i - Z'_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2.$$
Then $Z_3'$ can be written in terms of $Z_1$, $Z_2$ and $Z_3$ as,

$$Z_3' = \sum_{i=2}^{n} (X_i - Z_1')^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2$$

$$= \sum_{i=2}^{n} \left( X_i - \frac{(nZ_1 - X_1)}{n - 1} \right)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2$$

$$= \sum_{i=2}^{n} \left( X_i - Z_1 + \frac{X_1 - Z_1}{n - 1} \right)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2$$

$$= \sum_{i=1}^{n} (X_i - Z_1)^2 - \frac{n}{n - 1} (X_1 - Z_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2$$

$$= Z_3 - \frac{n}{n - 1} (X_1 - Z_1)^2.$$

We obtain the Jacobian of the transformation $(X_2, \ldots, X_n, Y_1, \ldots, Y_m)$ to $(X_2, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}, Z_1', Z_2, Z_3')$:

$$|J| = m(n-1) \frac{1}{\sqrt{2 \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( m z_2 - \sum_{j=1}^{m-1} y_j \right)^2 + m z_2^2 + n z_1^2 - \sum_{i=1}^{n-2} x_i^2 \right] - \left( n z_1 - \sum_{i=1}^{n-2} x_i \right)^2}}$$

Let $X_{n-2}' = (X_2, \ldots, X_{n-2})$ and $Y_{m-1}' = (Y_1, \ldots, Y_{m-1})$, then the conditional joint density function of $X_{n-2}', Y_{m-1}'$ given $Z_1', Z_2, Z_3'$ is,

$$f_{X_{n-2}', Y_{m-1}'}(x_{n-2}', y_{m-1}' | z_1', z_2, z_3') = \frac{K_1^*(n)}{[z_3 - \frac{n}{m-1} (x_1 - z_1)^2]^{\frac{m+n(n-1)-2}{2}}} \sqrt{2 \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( m z_2 - \sum_{j=1}^{m-1} y_j \right)^2 + m z_2^2 + n z_1^2 - \sum_{i=1}^{n-2} x_i^2 \right] - \left( n z_1 - \sum_{i=1}^{n-2} x_i \right)^2},$$

for $(x_{n-2}', y_{m-1}', z_1, z_2, z_3) \in \mathcal{D}$

where $K_1^*(n) = \frac{\sqrt{m(n-1)}}{\pi^{\frac{(m+n(n-1)-2)}{2}}}.$

Multiplying and dividing $f_{X_{n-2}', Y_{m-1}'}(x_{n-2}', y_{m-1}' | z_1, z_2, z_3)$ by $f_{X_{n-2}', Y_{m-1}'}(x_{n-2}', y_{m-1}' | z_1', z_2, z_3')$, we obtain

$$f_{X_{n-2}', Y_{m-1}'}(x_{n-2}', y_{m-1}' | z_1, z_2, z_3) =$$
\[
\sqrt{n} \frac{r\left(\frac{m+n-2}{2}\right)}{\Gamma\left(\frac{m+n-2}{2}\right)} \left[ z_1 - \frac{n-1}{n} (x_1 - z_1)^2 \right] \left( \frac{m+n-2}{2} \right) \int X_{n-2}^X Y_{m-1} | Z_1, Z_2, Z_3 \]

By integrating out \( X_2, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1} \), we obtain the UMVUE of the density function of \( X_1 \) at a given point \( x_1 \):

\[
f_{X_1|Z_1, Z_2, Z_3} (x_1 | z_1, z_2, z_3) = K_2(n) \cdot \frac{\left[ 1 - \frac{n}{n-1} (x_1 - z_1)^2 / z_3 \right]^{\frac{m+n-5}{2}}}{\sqrt{z_3}}
\]

for \( (x_1 - z_1)^2 < \frac{n-1}{n} z_3 \), \hspace{1cm} (2.2.6)

where \( K_2(n) = \frac{\sqrt{\pi} r\left(\frac{m+n-2}{2}\right)}{\Gamma\left(\frac{m+n-2}{2}\right)} \).

Similarly, fixing \( Y_1 \), we obtain the UMVUE of the density function of \( Y_1 \) at a given point \( y_1 \):

\[
f_{Y_1|Z_1, Z_2, Z_3} (y_1 | z_1, z_2, z_3) = K_3(n) \cdot \frac{\left[ 1 - \frac{m}{m-1} (y_1 - z_2)^2 / z_3 \right]^{\frac{m+n-5}{2}}}{\sqrt{z_3}}
\]

for \( (y_1 - z_2)^2 < \frac{m-1}{m} z_3 \), \hspace{1cm} (2.2.7)

where \( K_3(n) = \frac{\sqrt{m} r\left(\frac{m+n-2}{2}\right)}{\Gamma\left(\frac{m+n-2}{2}\right)} \).

Consider the transformation \( V_1 = \sqrt{\frac{n}{n-1}} \frac{(X_1 - Z_1)}{\sqrt{Z_3}} \) and \( V_2 = \sqrt{\frac{m}{m-1}} \frac{(Y_1 - Z_2)}{\sqrt{Z_3}} \). Using (2.2.6), the density function of \( V_1 \) given \( Z_1, Z_2, Z_3 \) is obtained as

\[
f_{V_1|Z_1, Z_2, Z_3} (v_1 | z_1, z_2, z_3) = K_2(n) \sqrt{\frac{n-1}{n}} \left[ 1 - v_1^2 \right]^{\frac{m+n-5}{2}}
\]

for \( v_1^2 < 1 \). \hspace{1cm} (2.2.8)

Similarly, the density function of \( V_2 \), given \( Z_1, Z_2, Z_3 \), by (2.2.7), is obtained as

\[
f_{V_2|Z_1, Z_2, Z_3} (v_2 | z_1, z_2, z_3) = K_3(n) \sqrt{\frac{m-1}{m}} \left[ 1 - v_2^2 \right]^{\frac{m+n-5}{2}}
\]

for \( v_2^2 < 1 \). \hspace{1cm} (2.2.9)

The conditional distribution of \( V_1 \) is symmetric, and does not depend on \( Z_1, Z_2, Z_3 \), hence \( V_1 \) and \( (Z_1, Z_2, Z_3) \) are independent. Therefore, the odd order conditional moments of \( \frac{X_1 - Z_1}{\sqrt{Z_3}} \)
are
\[
\mathbb{E}\left[ \left( \frac{X_1 - Z_1}{\sqrt{Z_3}} \right)^{2r-1} \mid Z_1, Z_2, Z_3 \right] = 0, \quad r = 1, 2, \ldots, \text{ and}
\]
\[
\mathbb{E}\left[ \frac{n}{n-1} \left( \frac{X_1 - Z_1}{\sqrt{Z_3}} \right)^2 \mid Z_1, Z_2, Z_3 \right] = \frac{1}{m + n - 2}. \quad (2.2.10)
\]

Similarly, considering the conditional distribution (2.2.9) of \( V \), we get
\[
\mathbb{E}\left[ \left( \frac{Y_1 - Z_2}{\sqrt{Z_3}} \right)^{2r-1} \mid Z_1, Z_2, Z_3 \right] = 0, \quad r = 1, 2, \ldots, \text{ and}
\]
\[
\mathbb{E}\left[ \frac{m}{m-1} \left( \frac{Y_1 - Z_2}{\sqrt{Z_3}} \right)^2 \mid Z_1, Z_2, Z_3 \right] = \frac{1}{m + n - 2}. \quad (2.2.11)
\]

A characterization of normal distribution related to two samples based on second conditional moment is given in the following theorem.

**Theorem 2.2.1.** Let \( F \) and \( G \) be two nondegenerate distributions with finite second moments and let \( X_i, i = 1, \ldots, n, \ n \geq 2, \ Y_j, j = 1, \ldots, m, \ m \geq 2 \) be two independent random samples from \( F \) and \( G \), respectively. Then \( F \) is a \( N(\xi, \sigma^2) \) distribution and \( G \) is a \( N(\eta, \sigma^2) \) distribution if and only if
\[
\mathbb{E}\left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1}(Y_1 - Z_2)^2 \mid Z_1, Z_2, Z_3 \right] = 0. \quad (2.2.12)
\]

**Proof.** If \( X_1, \ldots, X_n \) is a random sample from \( N(\xi, \sigma^2) \) and \( Y_1, \ldots, Y_m \) is a random sample of \( N(\eta, \sigma^2) \), then by (2.2.10) and (2.2.11),
\[
\mathbb{E}\left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1}(Y_1 - Z_2)^2 \mid Z_1, Z_2, Z_3 \right] = 0.
\]

Conversely, if
\[
\mathbb{E}\left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1}(Y_1 - Z_2)^2 \mid Z_1, Z_2, Z_3 \right] = 0,
\]
then
\[
\mathbb{E} \left[ \frac{n}{n-1} (X_1 - Z_1)^2 \right] = \mathbb{E} \left[ \frac{m}{m-1} (Y_1 - Z_2)^2 \right],
\]
which implies \( \text{Var}(X) = \text{Var}(Y) = \sigma^2 \).

Further, taking the expectation of (2.2.12) w.r.t. \( Z_3 \)
\[
\mathbb{E} \left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1} (Y_1 - Z_2)^2 \right] | Z_1, Z_2
\]
\[= \mathbb{E}_{Z_3} \left[ \mathbb{E} \left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1} (Y_1 - Z_2)^2 \right] | Z_1, Z_2, Z_3 \right] = 0 \tag{2.2.13}
\]
Multiplying both sides of (2.2.13) by \( e^{it_1 Z_1 + it_2 m Z_2} \), and then taking the expectation, we have
\[
\mathbb{E}_{Z_1, Z_2} \left[ e^{it_1 Z_1 + it_2 m Z_2} \mathbb{E} \left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1} (Y_1 - Z_2)^2 \right] | Z_1, Z_2 \right] \]
\[= \mathbb{E} \left[ e^{it_1 Z_1 + it_2 m Z_2} \left[ \frac{n}{n-1} (X_1 - Z_1)^2 - \frac{m}{m-1} (Y_1 - Z_2)^2 \right] \right] = 0 \tag{2.2.14}
\]
Let \( \phi_1(t_1) = \mathbb{E} \left[ e^{it_1 X} \right] \) and \( \phi_2(t_2) = \mathbb{E} \left[ e^{it_2 Y} \right] \) be the characteristic functions of the distributions of \( X \) and \( Y \) respectively. Then from (2.2.14)
\[
\frac{n}{n-1} \phi_2^m(t_2) \mathbb{E} \left[ e^{it_1 Z_1} (X_1 - Z_1)^2 \right] - \frac{m}{m-1} \phi_1^n(t_1) \mathbb{E} \left[ e^{it_2 m Z_2} (Y_1 - Z_2)^2 \right] = 0. \tag{2.2.15}
\]
Consider the first expectation part of (2.2.15), which is
\[
\mathbb{E} \left[ e^{it_1 Z_1} (X_1 - Z_1)^2 \right] = \mathbb{E} \left[ e^{it_1 Z_1} (X_1^2 - 2X_1 Z_1 + Z_1^2) \right]. \tag{2.2.16}
\]
From (2.2.16), we have
\[
\mathbb{E} \left[ e^{it_1 Z_1} X_1^2 \right] = \mathbb{E} \left[ X_1^2 e^{it_1 X_1} e^{it_1 X_2} \ldots e^{it_1 X_n} \right] \]
\[= - \phi_1''(t_1) \phi_1^{n-1}(t_1) \tag{2.2.17}
\]
\[-2 \mathbb{E} \left[ e^{it_1 Z_1} X_1 Z_1 \right] = - \frac{2}{n} \mathbb{E} \left[ e^{it_1 Z_1} X_1 \sum_{i=1}^n X_i \right] \]
\[= \mathbb{E} \left[ X_1^2 e^{it_1 \sum X_i} + X_1 X_2 e^{it_1 \sum X_i} + \ldots + X_1 X_n e^{it_1 \sum X_i} \right] \]
\[= - \frac{2}{n} \left[ -\phi_1''(t_1) \phi_1^{n-1}(t_1) - (n-1) (\phi_1'(t_1))^2 \phi_1^{n-2}(t_1) \right]. \tag{2.2.18}
\]
\[ \mathbb{E} \left[ e^{it_1nZ_1} Z_1^2 \right] = \frac{1}{n^2} \mathbb{E} \left[ e^{it_1nZ_1} \left( \sum X_i^2 + 2 \sum_{i<j} X_i X_j \right) \right] \]

\[ = \frac{1}{n^2} \mathbb{E} X_1^2 e^{it_1 \sum X_i} + \ldots + X_n^2 e^{it_1 \sum X_i} + 2X_n X_{n-1} e^{it_1 \sum X_i} \]

\[ + 2X_n e^{it_1 \sum X_i} = \frac{1}{n^2} \left[ -n\phi''_1(t_1) \phi''_1(t_1) - n(n - 1) \left( \phi'_1(t_1) \right)^2 \phi''_1(t_1) \right] \]  
(2.2.19)

Considering the second expectation part of (2.2.15),

\[ \mathbb{E} \left[ e^{it_2mZ_2} (Y_1 - Z_2)^2 \right] = \mathbb{E} \left[ e^{it_2mZ_2} (Y_1^2 - 2Y_1Z_2 + Z_2^2) \right], \]

we can obtain

\[ \mathbb{E} \left[ e^{it_2mZ_2} Y_1^2 \right] = -\phi''_2(t_2) \phi''_2(t_2), \]  
(2.2.20)

\[ -2\mathbb{E} \left[ e^{it_2mZ_2} Y_1 Z_2 \right] = -\frac{2}{m} \left[ -\phi''_2(t_2) \phi''_2(t_2) - (m - 1) \left( \phi'_2(t_2) \right)^2 \phi''_2(t_2) \right], \]  
(2.2.21)

\[ \mathbb{E} \left[ e^{it_2mZ_2} Z_2^2 \right] = \frac{1}{m^2} \left[ -m\phi''_2(t_2) \phi''_2(t_2) - m(m - 1) \left( \phi'_2(t_2) \right)^2 \phi''_2(t_2) \right]. \]  
(2.2.22)

Substituting from (2.2.17) to (2.2.22) in (2.2.15) and simplifying, we get the following differential equation.

\[ \phi''_2(t_2) \left[ -\phi'_1(t_1) \phi''_1(t_1) + (\phi'_1(t_1))^2 \phi''_1(t_1) \right] \]

\[ + \phi'_1(t_1) \left[ \phi''_2(t_2) \phi''_2(t_2) - (\phi'_2(t_2))^2 \phi''_2(t_2) \right] = 0 \]  
(2.2.23)

Let \( t_2 = 0 \) in (2.2.23), we get

\[ -\phi''_1(t_1) \phi''_1(t_1) + (\phi'_1(t_1))^2 \phi''_1(t_1) + \phi''_1(t_1) \left[ \phi''_1(t_1) - (\phi'_1(t_1))^2 \right] = 0 \]  
(2.2.24)

Substituting \( \phi'_1(t_1) = i\eta \) and \( \phi''_1(t_1) = -(\sigma^2 + \eta^2) \) in (2.2.24), and simplifying, we get

\[ -\phi''_1(t_1) \phi_1(t_1) + (\phi'_1(t_1))^2 - \sigma^2 \phi_1(t_1) = 0. \]

i.e. \( \left( \frac{\phi'_1(t_1)}{\phi_1(t_1)} \right)' = -\sigma^2. \)
Solving the above differential equation we get,

\[ \phi_1(t_1) = e^{-\frac{1}{2} \sigma^2 t_2 + ct + d} \]  

(2.2.25)

Using the conditions \( \phi_1(0) = 1 \) and \( \phi_1'(0) = i\xi \) we get \( d = 0 \) and \( c = i\xi \) respectively. Now (2.2.25) becomes \( \phi_1(t_1) = e^{i\xi t_1 - \frac{1}{2} \sigma^2 t_2^2} \), which implies \( F \) is \( N(\xi, \sigma^2) \).

Similarly, letting \( t_2 = 0 \) in (2.2.17), we get \( G \) is \( N(\eta, \sigma^2) \).

The results of Corollaries 2.2.1, 2.2.2 and 2.2.3 below are obtained directly from Theorem 2.2.1. Corollary 2.2.1 and Corollary 2.2.3 are used in procedures to construct goodness-of-fit test for two normal distributions with unknown means and common unknown variance.

**Corollary 2.2.1.** Let \( F \) and \( G \) be two nondegenerate distributions with finite second moments and let \( X_i, i = 1, \ldots, n, n \geq 2, Y_j, j = 1, \ldots, m, m \geq 2 \) be two independent random samples from \( F \) and \( G \), respectively. Then \( F \) is a \( N(\xi, \sigma^2) \) distribution and \( G \) is a \( N(\eta, \sigma^2) \) distribution if and only if

\[
\mathbb{E} \left[ \frac{\sum_{i=1}^{n} (X_i - Z_1)^2}{Z_3} \middle| Z_1, Z_2, Z_3 \right] = \frac{n - 1}{m + n - 2}. 
\]  

(2.2.26)

**Corollary 2.2.2.** Let \( F \) and \( G \) be two nondegenerate distributions with finite second moments and let \( X_i, i = 1, \ldots, n, n \geq 2, Y_j, j = 1, \ldots, m, m \geq 2 \) be two independent random samples from \( F \) and \( G \), respectively. Then \( F \) is a \( N(\xi, \sigma^2) \) distribution and \( G \) is a \( N(\eta, \sigma^2) \) distribution if and only if the conditional density function of \( X_1 \) given \( Z_1, Z_2, Z_3 \) and the conditional density function of \( Y_1 \) given \( Z_1, Z_2, Z_3 \) are given by (2.2.6) and (2.2.7) respectively.
Proof. We only need to prove that (2.2.6) and (2.2.7) imply normality. Since (2.2.6) implies (2.2.8), (2.2.8) implies (2.2.10) and (2.2.7) implies (2.2.9), (2.2.9) implies (2.2.11), by Theorem 2.2.1, $F$ is $N(\xi, \sigma^2)$ and $G$ is $N(\eta, \sigma^2)$. □

Corollary 2.2.3. Let $F$ and $G$ be two nondegenerate distributions with finite second moments and let $X_i, i = 1, \ldots, n$, $n \geq 2$, $Y_j, j = 1, \ldots, m$, $m \geq 2$ be two independent random samples from $F$ and $G$, respectively. Then $F$ is a $N(\xi, \sigma^2)$ distribution and $G$ is a $N(\eta, \sigma^2)$ distribution if and only if the joint conditional density function of $X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}$ given $Z_1, Z_2, Z_3$ is given by (2.2.5).

Proof. We only need to prove that (2.2.5) implies normality. If (2.2.5) holds then the conditional density function of $X_1$ given $Z_1, Z_2, Z_3$ is (2.2.6) and the conditional density function of $Y_1$ given $Z_1, Z_2, Z_3$ is (2.2.7). Then by Corollary 2.2.2, $F$ is $N(\xi, \sigma^2)$ and $G$ is $N(\eta, \sigma^2)$. □

2.3 Characterization: $k$ sample case

We can extend the above results to $k$ sample case. Let $X_{i1}, \ldots, X_{in_i}, i = 1, \ldots, k$, $n_i \geq 2$ be $k$ random samples. Let $Z_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, i = 1, \ldots, k$, $Z_{k+1} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - z_i)^2$.

Theorem 2.3.1. Let $X_{i1}, \ldots, X_{in_i}, n_i \geq 2$ for $i = 1, \ldots, k$, be random samples from non-degenerate distributions $F_i, i = 1, \ldots, k$ respectively, with a finite second moment. Then $F_i$ is a $N(\xi_i, \sigma^2)$, $i = 1, \ldots, k$ distributions if and only if

$$
\mathbb{E}
\left[
\frac{n_i}{n_i - 1} (X_{i1} - Z_1)^2
- \frac{n_i}{n_i - 1} (X_{i1} - Z_i)^2|Z_1, \ldots, Z_k, Z_{k+1}
\right] = 0 \quad (2.3.1)
$$

$i = 2, \ldots, k$.

Proof. Consider the sample $X_{1j}, j = 1, \ldots, n_1$ with each $X_{ij}, i = 2, \ldots, k$, $j = 1, \ldots, n_i$, and use Theorem 2.2.1. □
The following corollaries can be obtained directly from Theorem 2.3.1 and the techniques used in obtaining Corollaries 2.2.2 and 2.2.3.

**Corollary 2.3.1.** Let \( X_{i1}, \ldots, X_{in_i}, n_i \geq 2 \) for \( i = 1, \ldots, k \), be random samples from non-degenerate distributions \( F_i, i = 1, \ldots, k \) respectively, with a finite second moment. Then \( F_i \) is a \( N(\xi_i, \sigma^2) \), \( i = 1, \ldots, k \) distributions if and only if the conditional density function of \( X_{i1} \) given \( Z_1, \ldots, Z_k, Z_{k+1} \) is given by

\[
f_{(x_{i1}|z_1, \ldots, z_k, z_{k+1})} = K_8(n) \left( 1 - \frac{n_i - 1}{n_i} \frac{(x_{i1} - z_i)^2}{z_{k+1}} \right)^{n_1 + \ldots + n_k - k - 3} \sqrt{z_{k+1}}
\]

for \( (x_{i1} - z_i)^2 < \frac{n_i - 1}{n_i} z_{k+1}, \) \( i = 1, \ldots, k, \)

where \( K_8(n) = \frac{\sqrt{n_i}}{(n_i - 1)^{\frac{1}{2}}} \Gamma(\frac{n_1 + \ldots + n_k - 2}{2}) \Gamma(\frac{n_1 + \ldots + n_k - k - 1}{2}). \)

**Corollary 2.3.2.** Let \( X_{i1}, \ldots, X_{in_i}, n_i \geq 2 \) for \( i = 1, \ldots, k \), be random samples from non-degenerate distributions \( F_i, i = 1, \ldots, k \) respectively, with a finite second moment. Then \( F_i \) is a \( N(\xi_i, \sigma^2) \), \( i = 1, \ldots, k \) distributions if and only if the joint conditional density function of \( X_{11}, \ldots, X_{1(n_1-2)}, X_{21}, \ldots, X_{2(n_2-1)}, \ldots, X_{k1}, \ldots, X_{k(n_k-1)} \) given \( Z_1, \ldots, Z_k, Z_{k+1} \) is

\[
f_{(x_{11}, \ldots, x_{1(n_1-2)}, x_{21}, \ldots, x_{2(n_2-1)}, \ldots, x_{k1}, \ldots, x_{k(n_k-1)}|z_1, \ldots, z_k, z_{k+1})} = \frac{K_9(n)}{z_{k+1}^{n_1 + \ldots + n_k - k - 1}}
\]

\[
\sqrt{2} \left[ z_{k+1} - \sum_{i=2}^{k} \sum_{j=1}^{n_i-2} x_{ij}^2 - \sum_{i=2}^{k} \sum_{j=1}^{n_i-1} (n_i z_i - \sum_{j=1}^{n_i-1} x_{ij})^2 + \sum_{i=1}^{k} n_i z_i^2 - \sum_{j=1}^{n_i-2} x_{ij}^2 \right] - \left( n z_1 - \sum_{j=1}^{n_1-2} x_{1j} \right)^2
\]

for \( z_{k+1} - \sum_{i=2}^{k} \sum_{j=1}^{n_i-2} x_{ij}^2 - \sum_{i=2}^{k} \sum_{j=1}^{n_i-1} (n_i z_i - \sum_{j=1}^{n_i-1} x_{ij})^2 + \sum_{i=1}^{k} n_i z_i^2 - \sum_{j=1}^{n_i-2} x_{ij}^2 \leq \left( n z_1 - \sum_{j=1}^{n_1-2} x_{1j} \right)^2, \)

where

\[
K_9(n) = \frac{\sqrt{n_i}}{(n_i - 1)^{\frac{1}{2}}} \Gamma(\frac{n_1 + \ldots + n_k - 2}{2}) \Gamma(\frac{n_1 + \ldots + n_k - k - 1}{2}).
\]

(2.3.3)
where \( K_9(n) = \sqrt{\frac{n_1 \cdots n_k}{\pi}} \frac{\Gamma\left(\frac{n_1 + \cdots + n_k - k}{2}\right)}{\Gamma\left(\frac{n_1 + \cdots + n_k - k}{2} + k\right)} \).

### 2.4 Application to goodness-of-fit tests

Let \( F \) and \( G \) be two nondegenerate distributions with finite second moments and let \( X_i, i = 1, \ldots, n, n \geq 2 \), \( Y_j, j = 1, \ldots, m, m \geq 2 \) be two independent random samples from \( F \) and \( G \), respectively. Consider testing the composite hypotheses

\[
H_0 : F \text{ is a } N(\xi, \sigma^2) \text{ and } G \text{ is a } N(\eta, \sigma^2) \quad \text{versus} \quad H_a : \text{not } H_0 \quad (2.4.1)
\]

where \( \xi, \eta \) and \( \sigma^2 \) are unknown.

Using the result of Corollary (2.2.3), testing (2.4.1) is equivalent to testing,

\[
H_0^* : \text{The joint conditional density function of } X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1} \text{ given } Z_1, Z_2, Z_3 \text{ is given by (2.2.5)}.
\]

This simple hypothesis can be tested using any EDF statistic, such as Kolmogorov-Smirnov statistic, Kuiper statistic, Cramer-vonMises statistic, Anderson-Darling statistic (See D’Agostino and Stephens (1986)). Since the computations of the EDF goodness-of-fit test statistics are complicated we will propose a set of one-to-one transformations which transform these variables to Student’s \( t \) variables, in the succeeding chapter.

Using (2.2.3) in Corollary 2.2.1, testing (2.4.1) is equivalent to testing

\[
H_0^{**} : E \left[ \frac{\sum_{i=1}^{n} (X_i - Z_1)^2}{\sum_{i=1}^{n} (X_i - Z_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2} \middle| Z_1, Z_2, Z_3 \right] = \frac{n - 1}{m + n - 2}.
\]
In this case, we can use the equivalent test statistic \( T = \frac{\sum_{i=1}^{n} (X_i - Z_1)^2/(n-1)}{\sum_{j=1}^{m} (Y_j - Z_2)^2/(m-1)} \), and reject \( H_0 \) at level \( \alpha \) if and only if, \( T \leq c_1 \) or \( T \geq c_2 \), where \( c_1 \) and \( c_2 \) are determined by

\[
\begin{align*}
P \left( \frac{\sum_{i=1}^{n} (X_i - Z_1)^2/(n-1)}{\sum_{j=1}^{m} (Y_j - Z_2)^2/(m-1)} \leq c_1 | Z_1, Z_2, Z_3 \right) &= \frac{\alpha}{2} \quad \text{and} \quad P \left( \frac{\sum_{i=1}^{n} (X_i - Z_1)^2/(n-1)}{\sum_{j=1}^{m} (Y_j - Z_2)^2/(m-1)} \geq c_2 | Z_1, Z_2, Z_3 \right) = \frac{\alpha}{2}.
\end{align*}
\]

Since the distribution of \( T \) is independent of \( Z_1, Z_2, \) and \( Z_3 \), it is easy to see that \( c_1 \) and \( c_2 \) are the \( \left( \frac{\alpha}{2} \right)^{th} \) and \( \left( 1 - \frac{\alpha}{2} \right)^{th} \) percentiles of the \( F \) distribution with \( (n-1), (m-1) \) degrees of freedom, respectively.

The study of the goodness-of-fit tests, derived from the characterization results of this chapter will be presented in the next chapter.
CHAPTER 3
EDF GOODNESS-OF-FIT TESTS FOR TESTING THE DISTRIBUTIONAL
ASSUMPTIONS IN ONE-WAY REGRESSION MODEL

3.1 Introduction

In the one-way classification with $k$ treatments, we assume that the samples are independent of each other, the populations from which the samples were obtained are normally distributed and the variances of the populations are equal. Assessing the above assumptions can be done by:

testing the hypothesis,

$$H_0 : Y_{ij} \sim N(\xi_i, \sigma^2) \quad i = 1, \ldots, k \quad j = 1, \ldots, n_i \quad \text{vs} \quad H_1 : \text{not } H_a,$$

where $Y_{ij}$ are the observed values. In this case the location and scale parameters are unknown.

Characterizations supporting goodness-of fit tests for normal distribution, when the parameters are not completely known, are given by several authors. Kotlarsky (1966) characterized the normal distribution by Student’s law. Csargo, Seshadri and Yalovsky (1973) proposed exact EDF goodness-of fit tests for normality in the presence of unknown parameters. Nguyen and Dinh (2003) gave characterization of normal distribution based on the UMVU estimator of the density function which can be used as a transformation in the procedure to construct an EDF goodness-of-fit test for testing normality of a distribution. In chapter 2, we gave a characterization of normal distribution related to $k$ samples based on second conditional moments which can be used in constructing an EDF goodness-of-fit test for testing normality and homogeneity of variance of the samples.
In this chapter, in Section 2 and 3, the location and scale parameters in the distributional assumption in one-way regression model are eliminated by means of suitable transformations on the original observations. The transformed random variables are then shown to follow Student’s $t$ distributions with known parameters under the null hypothesis. Testing the above hypothesis, by performing the EDF goodness-of-fit tests on the transformed variables, is discussed in Section 4. The powers of the tests are estimated by Monte Carlo method for several alternatives in Section 5.

3.2 Characterization of normal distribution related to two samples based on Student’s law

Let $X_i, \ i = 1, \ldots, n$ and $Y_j, \ j = 1, \ldots, m, \ m, n \geq 2$, be two independent random samples from two normal distributions with means $\xi$ and $\eta$, respectively, and with common variance $\sigma^2$. Then $Z_1 = \frac{\sum_{i=1}^{n} X_i}{n}, \ Z_2 = \frac{\sum_{j=1}^{m} Y_j}{m}, \ Z_3 = \sum_{i=1}^{n} (X_i - Z_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2$, are jointly complete and sufficient statistics for $\xi, \eta, \text{and} \sigma^2$.

Let $X_{n-2} = (X_1, \ldots, X_{n-2})$ and $Y_{m-1} = (Y_1, \ldots, Y_{m-1})$. The conditional joint density function of $X_{n-2}, Y_{m-1}$ given $Z_1, Z_2, Z_3$ is

$$f_{X_{n-2}Y_{m-1}|Z_1Z_2Z_3}(x_{n-2}, y_{m-1}|z_1, z_2, z_3) =$$

$$K_1(n). \frac{1}{\sqrt{\pi^{(m+n-2)/2}} \Gamma \left(\frac{m+n-2}{2}\right)\ z_3^{\frac{m+n-2}{2}}} \sqrt{2 \ z_3} \ z_3 \sum_{j=1}^{m-1} y_j^2 - \left(mz_2 - \sum_{j=1}^{m-1} y_j\right)^2 + mz_2^2 + nz_2^2 - \sum_{i=1}^{n-2} x_i^2 - \left(nz_1 - \sum_{i=1}^{n-2} x_i\right)^2,$$

for $(x_{n-2}, y_{m-1}, z_1, z_2, z_3) \in D$ (3.2.1)

where $K_1(n) = \frac{\sqrt{mn}}{\pi^{(m+n-2)/2}} \Gamma \left(\frac{m+n-2}{2}\right) \text{ and } D = \left\{ (x_{n-2}, y_{m-1}, z_1, z_2, z_3) : \right.$

$$2 \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left(mz_2 - \sum_{j=1}^{m-1} y_j\right)^2 + mz_2^2 + nz_2^2 - \sum_{i=1}^{n-2} x_i^2 \right] > \left(nz_1 - \sum_{i=1}^{n-2} x_i\right)^2 \right\}.$$


The UMVUE of the density function of $X_1$ at a given point $x_1$ is

$$f_{X_1|Z_1,Z_2,Z_3}(x_1|z_1, z_2, z_3) = K_2(n). \frac{\left[1 - \frac{n}{n-1}(x_1 - z_1)^2/z_3\right]^{\frac{m+n-5}{2}}}{\sqrt{z_3}}$$

for $(x_1 - z_1)^2 < \frac{n-1}{n} z_3,$ (3.2.2)

where $K_2(n) = \sqrt{\frac{n}{\pi}} \frac{\Gamma\left(\frac{m+n-2}{2}\right)}{\Gamma\left(\frac{m+(n-1)}{2}\right)}.$

The UMVUE of the density function of $Y_1$ at a given point $y_1$ is

$$f_{Y_1|Z_1,Z_2,Z_3}(y_1|z_1, z_2, z_3) = K_3(n). \frac{\left[1 - \frac{m}{m-1}(y_1 - z_2)^2/z_3\right]^{\frac{m+n-5}{2}}}{\sqrt{z_3}}$$

for $(y_1 - z_2)^2 < \frac{m-1}{m} z_3,$ (3.2.3)

where $K_3(n) = \sqrt{\frac{m}{\pi}} \frac{\Gamma\left(\frac{m+n-2}{2}\right)}{\Gamma\left(\frac{m-(m-1)}{2}\right)}.$

The following characterization of normal distribution related to two samples based on second conditional moments is given in Chapter 2 (Corollary 2.2.3):

Let $F$ and $G$ be two nondegenerate distributions with finite second moments and let $X_i, i = 1, \ldots, n, n \geq 2, Y_j, j = 1, \ldots, m, m \geq 2$ be two independent random samples from $F$ and $G$, respectively. Then $F$ is a $N(\xi, \sigma^2)$ distribution and $G$ is a $N(\eta, \sigma^2)$ distribution if and only if the joint conditional density function of $X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}$ given $Z_1, Z_2, Z_3$ is given by (3.2.1).

This characterization can be changed to a characterization using Student’s law. Given $Z_1, Z_2$ and $Z_3$, the following consecutive one-to-one transformations are considered.

Let

$$U_i = \frac{X_i - Z_1}{\sqrt{Z_3}} \quad i = 1, \ldots, n - 2,$$

$$U_{n-2+j} = \frac{Y_j - Z_2}{\sqrt{Z_3}} \quad j = 1, \ldots, m - 1.$$ (3.2.4)
We find the conditional joint density of \( U_1, \ldots, U_k, \ k = 1, \ldots, n - 2 \), given \( Z_1, Z_2, Z_3 \). Fix \( X_1, \ldots, X_k \) and let, 

\[
Z''_1 = \frac{nZ_1 - \sum_{i=1}^{k} X_i}{n - k}
\]

\[
Z''_3 = \sum_{i=k+1}^{n} (X_i - Z''_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2.
\]

Then the Jacobian of the transformation \((X_{k+1}, \ldots, X_n, Y_1, \ldots, Y_m)\) to \((X_{k+1}, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}, Z'_1, Z_2, Z''_3)\):

\[
|J| = \frac{m(n - k)}{\sqrt{2} \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( mz_2 - \sum_{j=1}^{m-1} y_j \right)^2 + mz_2^2 + nz_1^2 - \sum_{i=1}^{n-2} x_i^2 \right] - \left( nz_1 - \sum_{i=1}^{n-2} x_i \right)^2}.
\]

The conditional joint density of \( X_1, \ldots, X_k \) given \( Z_1, Z_2, Z_3 \) is

\[
f_{(x_1, \ldots, x_k | z_1, z_2, z_3)} = \frac{1 - \left( \frac{n-k+1}{n-k} \sum_{i=1}^{k} (x_i - z_1)^2 - \frac{2}{n-k} \sum_{1 \leq i < i' \leq k} (x_i - z_1)(x_i' - z_1) \right) / z_3}{K_4(n)}
\]

for \( \left( \frac{n-k+1}{n-k} \sum_{i=1}^{k} (x_i - z_1)^2 - \frac{2}{n-k} \sum_{1 \leq i < i' \leq k} (x_i - z_1)(x_i' - z_1) \right) / z_3 < 1 \),

where \( K_4(n) = \frac{\sqrt{n}}{\pi^{\frac{m+n-3}{2}}} \Gamma \left( \frac{m+n-2}{2} \right) \Gamma \left( \frac{m+n-k-2}{2} \right) \).

We obtain the Jacobian of the transformation \( X_1, \ldots, X_k \) to \( U_1, \ldots, U_k \), \(|J| = Z'_3\). Then the conditional distribution of \( U_1, \ldots, U_k \) given \( Z_1, Z_2, Z_3 \) is

\[
f_{(u_1, \ldots, u_k | z_1, z_2, z_3)} = K_4(n) \left[ 1 - n - k + 1 \sum_{i=1}^{k} u_i^2 - \frac{2}{n-k} \sum_{1 \leq i < i' \leq k} u_i u_{i'} \right]^{\frac{m+n-k-4}{2}}
\]

for \( \frac{n-k+1}{n-k} \sum_{i=1}^{k} u_i^2 + \frac{2}{n-k} \sum_{1 \leq i < i' \leq k} u_i u_{i'} < 1 \), \( k = 1, \ldots, n - 2 \).
Similarly, fixing $X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_l$ and letting

$$
Z_1''' = \frac{nZ_1 - \sum_{i=1}^{n-2} X_i}{2},
$$

$$
Z_2''' = \frac{mZ_2 - \sum_{j=1}^{l} Y_j}{m-l},
$$

$$
Z_3''' = \sum_{i=n-1}^{n} (X_i - Z_1''')^2 + \sum_{j=l+1}^{m} (Y_j - Z_2''')^2;
$$

we obtain the jacobian of the transformation $(X_{n-1}, X_n, Y_{l+1}, \ldots, Y_m)$ to $(Y_{l+1}, \ldots, Y_{m-1}, Z_1''', Z_2'''', Z_3''')$:

$$
|J| = \frac{(m-l)(n-k)}{\sqrt{2 \left[ z_3 - \sum_{j=1}^{m-1} y_j^2 - \left( m z_2 - \sum_{j=1}^{m-1} y_j \right)^2 + m z_2^2 + n z_1^2 - \sum_{i=1}^{n-2} x_i^2 \right] - \left( n z_1 - \sum_{i=1}^{n-2} x_i \right)^2}}.
$$

The conditional joint density of $U_1, \ldots, U_{n-2}, U_{n-1}, \ldots, U_{n-2+l}$ given $Z_1, Z_2, Z_3$ is obtained as

$$
f_{(u_1, \ldots, u_{n-2}, u_{n-1}, \ldots, u_{n-2+l}|z_1, z_2, z_3)} = K_5(n) \left[ 1 - \frac{3}{2} \sum_{i=1}^{n-2} u_i^2 - \sum_{1 \leq i < i' \leq n-2} u_i u_{i'} - \frac{m-l+1}{m-l} \sum_{j=1}^{l} u_{n-2+j} - \sum_{1 \leq j < j' \leq l} u_{n-2+j} u_{n-2+j'} \right]^{\frac{m-l-2}{2}},
$$

for

$$
\frac{3}{2} \sum_{i=1}^{n-2} u_i^2 + \sum_{1 \leq i < i' \leq n-2} u_i u_{i'} + \frac{m-l+1}{m-l} \sum_{j=1}^{l} u_{n-2+j} + \sum_{1 \leq j < j' \leq l} u_{n-2+j} u_{n-2+j'} \leq 1 \quad (3.2.6)
$$

$$
l = 1, \ldots, m-1,
$$

where $K_5(n) = \frac{\sqrt{mn} \Gamma(m+n-2)}{\sqrt{(m-l)^2} \Gamma(m-l)} \Gamma\left(\frac{m+n-2}{2}\right) \Gamma\left(\frac{m-l}{2}\right)$.

We can see both (3.2.5) and (3.2.6) do not depend on $Z_1, Z_2$ and $Z_3$. Further, from (3.2.5)
and (3.2.6) the conditional distribution of $U_k$ given $U_1, \ldots, U_{k-1}, Z_1, Z_2, Z_3$ is obtained as

$$f_{(u_k|u_1, \ldots, u_{k-1}, z_1, z_2, z_3)} = \frac{f_{(u_1, \ldots, u_k|z_1, z_2, z_3)}}{f_{(u_1, \ldots, u_{k-1}|z_1, z_2, z_3)}}$$

$$= K_0(n) \left[ \frac{1 - \frac{n-k+1}{n-k} \sum_{i=1}^{k} u_i^2 - \frac{2}{n-k} \sum_{1 \leq i < i' \leq k} u_i u_{i'}}{1 - \frac{n-k+2}{n-k+1} \sum_{i=1}^{k-1} u_i^2 - \frac{2}{n-k+1} \sum_{1 \leq i < i' \leq k} u_i u_{i'}} \right]^{\frac{m+n-k-4}{2}}^{\frac{m+n-k-3}{2}}$$

for $\frac{n-k+1}{n-k} \sum_{i=1}^{k} u_i^2 + \frac{2}{n-k} \sum_{1 \leq i < i' \leq k} u_i u_{i'} < 1$, and

$$n - k + 2 \sum_{i=1}^{k-1} u_i^2 + \frac{2}{n-k+1} \sum_{1 \leq i < i' \leq k} u_i u_{i'} < 1,$$

(3.2.7)

$$k = 1, \ldots, n - 2,$$

where $K_0(n) = \frac{\sqrt{(n-k+1)}}{\sqrt{(n-k)}} \frac{\Gamma\left(\frac{m+n-k-1}{2}\right)}{\Gamma\left(\frac{m+n-k-2}{2}\right)}$,

and the conditional distribution of $U_{n-2+l}$ given $U_1, \ldots, U_{n-2}, U_{n-1}, \ldots, U_{n-2+(l-1)}, Z_1, Z_2, Z_3$ is obtained as

$$f_{(u_{n-2+l}|u_1, \ldots, u_{n-2}, u_{n-1}, \ldots, u_{n-2+(l-1)}, z_1, z_2, z_3)} =$$

$$K_7(n) \left[ \frac{1 - \frac{3}{2} \sum_{i=1}^{n-2} u_i^2 - \sum_{1 \leq i < i' \leq n-2} u_i u_{i'} - \frac{1}{m-1} \sum_{j=1}^{l} u_j^2 - \frac{2}{m-1} \sum_{1 \leq j < j' \leq l} u_{n-2+j} u_{n-2+j'}^{m-1} \right]^{\frac{l-1}{2}}$$

$$\left[ 1 - \frac{3}{2} \sum_{i=1}^{n-2} u_i^2 - \sum_{1 \leq i < j' \leq n-2} u_i u_{j'} - \frac{1}{m-1} \sum_{j=1}^{l-1} u_j^2 - \frac{2}{m-1} \sum_{1 \leq j < j' \leq l-1} u_{n-2+j} u_{n-2+j'} \right]^{\frac{l}{m-1}}$$

for $A + \frac{m-l+1}{m-l} \sum_{j=1}^{l} u_{n-2+j}^2 + \frac{2}{m-l} \sum_{1 \leq j < j' \leq l} u_{n-2+j} u_{n-2+j'} < 1$, and

$$A + \frac{m-l+2}{m-l+1} \sum_{j=1}^{l-1} u_{n-2+j}^2 + \frac{2}{m-l+1} \sum_{1 \leq j < j' \leq l-1} u_{n-2+j} u_{n-2+j'} < 1,$$

(3.2.8)

$$l = 1, \ldots, m - 1,$$

where $K_7(n) = \frac{\sqrt{(m+l-1)}}{\sqrt{(m-l)}} \frac{\Gamma\left(\frac{m+l+1}{2}\right)}{\Gamma\left(\frac{m+l}{2}\right)}$ and $A = \frac{3}{2} \sum_{i=1}^{n-2} u_i^2 + \sum_{1 \leq i < i' \leq n-2} u_i u_{i'}$. 
Now consider the transformation

\[
v_i = \sqrt{\frac{n-i+1}{n-i}} u_i + \frac{\sum_{j=1}^{i-1} u_j}{\sqrt{(n-i+1)(n-i)}} \quad \text{for} \quad i = 1, \ldots, n-2,
\]

\[
v_{n-2+l} = \frac{\sqrt{\frac{m-l+1}{m-l}} u_{n-2+l} + \frac{\sum_{j=1}^{l-1} u_{n-2+j}}{\sqrt{(m-l+1)(m-l)}}}{\sqrt{1 - \frac{3}{2} \sum_{i=1}^{n-2} u_i^2 - \sum_{1 \leq i < i' \leq n-2} u_i u_{i'} - \sum_{j=1}^{l-1} u_{n-2+j}^2 - \frac{(\sum_{j=1}^{l-1} u_{n-2+j})^2}{m-l+1}}}
\]

for \( l = 1, \ldots, m-1. \) \hspace{1cm} (3.2.9)

Using (3.2.7) and (3.2.8), the conditional density function of \( V_k \) given \( U_1, \ldots, U_{k-1}, Z_1, Z_2, Z_3 \) is obtained as

\[
f(v_k|u_1, \ldots, u_{k-1}, z_1, z_2, z_3) = \frac{\Gamma\left(\frac{m+n-k-1}{2}\right)}{\Gamma\left(\frac{m+n-k-2}{2}\right)} \left[1 - v_k^2\right]^{-\frac{m+n-k-4}{2}}
\]

for \( v_k^2 < 1, \quad k = 1, \ldots, m+n-3. \) \hspace{1cm} (3.2.10)

Notice that the density function (3.2.10) does not depend on \( U_1, \ldots, U_{k-1}. \) Now we can obtain the conditional joint density of \( V_1, \ldots, V_{m+n-3} \) given \( Z_1, Z_2, Z_3, \) which is

\[
f(v_1, \ldots, v_{m+n-3}|z_1, z_2, z_3) = \prod_{j=1}^{m+n-3} \frac{\Gamma\left(\frac{m+n-j-1}{2}\right)}{\Gamma\left(\frac{m+n-j-2}{2}\right)} \left[1 - v_j^2\right]^{-\frac{m+n-j-4}{2}}
\]

for \( v_i^2 < 1. \) \hspace{1cm} (3.2.11)

Transforming \( V_i \)'s to \( W_i \)'s using

\[
w_i = \sqrt{m + n - i} - 2 \frac{v_i}{\sqrt{1 - v_i^2}} \quad i = 1, \ldots, m+n-3,
\]

from (3.2.11), we can obtain the conditional density of \( W_1, \ldots, W_{m+n-3} \) given \( Z_1, Z_2, Z_3 \) as
\[
f(w_1, \ldots, w_{m+n-3}|z_1, z_2, z_3) = \prod_{j=1}^{m+n-3} \frac{\Gamma\left(\frac{m+n-j-1}{2}\right)}{\sqrt{\pi(m+n-j-2)}} \frac{\Gamma\left(\frac{m+n-j-2}{2}\right)}{\Gamma\left(\frac{m+n-j-1}{2}\right)} \left[1 + \frac{w_j^2}{m+n-j-2}\right]^{m+n-j-1},
\]
for all real values of \(w_1, \ldots, w_{m+n-3}\).

From (3.2.13), given \(Z_1, Z_2\) and \(Z_3\), \(W_i, i = 1, \ldots, m+n-3\), are conditionally independently distributed according to Student’s \(t\) distribution with \(m+n-i-2\) degrees of freedom, respectively.

We summarize these results in the following theorem:

**Theorem 3.2.1.** Let \(F\) and \(G\) be two nondegenerate distributions with finite second moments and let \(X_i, i = 1, \ldots, n, n \geq 2, Y_j, j = 1, \ldots, m, m \geq 2\) be two independent random samples from \(F\) and \(G\), respectively. In the conditional model given \(Z_1, Z_2\) and \(Z_3\), let \(W_i, i = 1, \ldots, m+n-3\) be obtained from \(X_1, \ldots, X_{n-2}, Y_1, \ldots, Y_{m-1}\) by the sequence of one-to-one transformations defined by (3.2.4), (3.2.9), and (3.2.12). Then \(F\) is a \(N(\xi, \sigma^2)\) distribution and \(G\) is a \(N(\eta, \sigma^2)\) distribution if and only if \(W_1, \ldots, W_{m+n-3}\) are conditionally jointly distributed according to (3.2.13), given \(Z_1, Z_2, Z_3\).

### 3.3 Characterization of normal distribution related to \(k\) samples based on Student’s law

The following characterization of normal distribution related to \(k\) samples based on second conditional moments is given in Chapter 2 (Corollary 2.3.2):

Let \(X_{i1}, \ldots, X_{in_i}, i = 1, \ldots, k, n_i \geq 2\) be \(k\) random samples from non-degenerate distributions \(F_i, i = 1, \ldots, k\) respectively, with a finite second moment. Let \(Z_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n}, i = 1, \ldots, k, Z_{k+1} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - z_i)^2\). Then \(F_i\) is a \(N(\xi_i, \sigma^2)\), \(i = 1, \ldots, k\) distributions if
and only if the joint conditional density function of $X_{11}, \ldots, X_{(n_1-2)}, X_{21}, \ldots, X_{2(n_2-1)}, \ldots, X_{k1}, \ldots, X_{k(n_k-1)}$ given $Z_1, \ldots, Z_k, Z_{k+1}$ is

$$f_{(x_{11}, \ldots, x_{1(n_1-2)}, \ldots, x_{2(n_2-1)}, \ldots, x_{k(n_k-1)})|Z_1, \ldots, Z_{k+1}}(x_{11}, \ldots, x_{1(n_1-2)}, \ldots, x_{2(n_2-1)}, \ldots, x_{k(n_k-1)})$$

$$= \frac{K_9(n)}{z_{k+1}}.$$  

$$2 \left[ z_{k+1} - \sum_{i=2}^{k} \sum_{j=1}^{n_i-1} x_{ij}^2 - \sum_{i=2}^{k} \left( n_i z_i - \sum_{j=1}^{n_i-1} x_{ij} \right)^2 + \sum_{i=1}^{k} n_i z_i^2 - \sum_{j=1}^{n_1-2} x_{1j}^2 \right] - \left( n z_1 - \sum_{j=1}^{n_1-2} x_{1j} \right)^2,$$

$$\times \frac{1}{z_{k+1}}.$$

$$\sqrt{2 \left[ z_{k+1} - \sum_{i=2}^{k} \sum_{j=1}^{n_i-1} x_{ij}^2 - \sum_{i=2}^{k} \left( n_i z_i - \sum_{j=1}^{n_i-1} x_{ij} \right)^2 + \sum_{i=1}^{k} n_i z_i^2 - \sum_{j=1}^{n_1-2} x_{1j}^2 \right] - \left( n z_1 - \sum_{j=1}^{n_1-2} x_{1j} \right)^2}, \quad (3.3.1)$$

where $K_9(n) = \sqrt{n_1 \ldots n_k} \frac{\Gamma \left( \frac{n_1 + \ldots + n_k - k}{2} \right)}{\pi^{\frac{n_1 + \ldots + n_k - k}{2}}}.$

**Corollary 3.3.1.** Let $X_{i1}, \ldots, X_{in_i}, n_i \geq 2$ for $i = 1, \ldots, k$, be random samples from non-degenerate distributions $F_i, i = 1, \ldots, k$ respectively, with a finite second moment. In the conditional model given $Z_1, \ldots, Z_k$ and $Z_{k+1}$, let $W_i, i = 1, \ldots, (n_1 + \ldots + n_k - k - 1)$ be
obtained by the sequence of one-to-one transformations defined by

\[
\begin{align*}
    u_i &= \frac{x_{1i} - z_1}{\sqrt{z_{k+1}}}, \quad i = 1, \ldots, n_1 - 2, \\
    u_{n_1 + \ldots + n_s - s + j} &= \frac{x_{sj} - z_s}{\sqrt{z_{k+1}}}, \quad s = 2, \ldots, k, \quad j = 1, \ldots, n_s - 1. \\
    v_i &= \frac{\sqrt{n_{i-1}+1}}{n_{i-1}} u_i + \frac{\sum_{j=1}^{i-1} u_j}{\sqrt{(n_{i-1}+1)(n_{i-1})}} \quad i = 1, \ldots, n_1 - 2, \\
    v_{n_1 - 2 + l} &= \sqrt{1 - \frac{3}{2} \sum_{i=1}^{n_1-2} u_i^2 - \sum_{1 \leq i < j \leq n_1-2} u_i u_j^2 - \sum_{j=1}^{l-1} u_{n_1 - 2 + j}^2 - \left(\frac{\sum_{j=1}^{l-1} u_{n_1 - 2 + j}}{n_2 - l + 1}\right)^2} \\
    v_{n_1 + \ldots + n_s - s + l} &= \sqrt{1 - \frac{3}{2} \sum_{i=1}^{n_s-2} u_i^2 - \sum_{1 \leq i < j \leq n_s-2} u_i u_j^2 - 2 \sum_{j=3}^{n_j-1} \sum_{i=1}^{n_1+\ldots+n_j-2-(j-1)+i} u_{n_1+\ldots+n_j-2-(j-1)+i}^2 - t_1 - t_2} \\
    &\quad s = 3, \ldots, k, \quad l = 1, \ldots, n_s - 1,
\end{align*}
\]

where

\[
\begin{align*}
    t_1 &= 2 \sum_{j=1, j \neq i}^{s} \sum_{1 \leq i < j \leq n_j-1} u_{n_1 + \ldots + n_j-2-(j-1)+i} u_{n_1 + \ldots + n_j-2-(j-1)+j}, \\
    t_2 &= \sum_{j=1}^{l-1} u_{n_1 + \ldots + n_s - s + j}^2 - \left(\sum_{j=1}^{l-1} u_{n_1 + \ldots + n_s - s + j}\right)^2 \\
    w_i &= \sqrt{n_1 + \ldots + n_k - i - k - \frac{v_i}{\sqrt{1 - v_i^2}}} \quad i = 1, \ldots, n_1 + \ldots + n_k - k - 1.
\end{align*}
\]

Then \( F_i \) is a \( N(\xi_i, \sigma^2) \), \( i = 1, \ldots, k \) distributions if and only if, given \( Z_1, \ldots, Z_k \) and \( Z_{k+1} \),
\[ W_i, \ i = 1, \ldots, n_1 + \ldots + n_k - k - 1 \] are conditionally independently distributed according to the Student’s t distribution with \( n_1 + \ldots + n_k - i - k \) degrees of freedom.

### 3.4 Application to goodness-of-fit tests

Let \( X_i, \ i = 1, \ldots, n, \ n \geq 2, \ Y_j, \ j = 1, \ldots, m, \ m \geq 2 \) be two independent random samples from nondegenerate distributions \( F \) and \( G \) with finite second moments, respectively. Consider testing the composite hypotheses \( H_0 : F \) is a \( N(\xi, \sigma^2) \) and \( G \) is a \( N(\eta, \sigma^2) \) versus \( H_a : \) not \( H_0 \), where \( \xi, \eta \) and \( \sigma^2 \) are unknown. Using the result of Theorem 3.2.1, it is equivalent to testing, \( H_0^* : \) Given \( Z_1, Z_2, Z_3, W_i, \ i = 1, \ldots, m + n - 3 \) are independently distributed according to Student’s t distribution with \( m + n - i - 2 \) degrees of freedom. This simple hypothesis can be tested using any EDF statistic, such as Kolmogorov-Smirnov statistic, Kuiper statistic, Anderson-Darling statistic (See D’Agostino and Stephens (1986)). We proposed another equivalent test for testing \( H_0 \) using \( T = \frac{\sum_{i=1}^{n} (X_i - Z_1)^2/(n-1)}{\sum_{j=1}^{m} (Y_j - Z_2)^2/(m-1)} \), as the test statistic and reject \( H_0 \) at level \( \alpha \) if and only if, \( T \leq c_1 \) or \( T \geq c_2 \), where \( c_1 \) and \( c_2 \) are the \( \left( \frac{\alpha}{2} \right)^{th} \) and \( 1 - \left( \frac{\alpha}{2} \right)^{th} \) percentiles of the \( F \) distribution with \( (n - 1), (m - 1) \) degrees of freedom, respectively (Section 4 in Chapter 2).

### 3.5 Power study

In this section we estimate the power of the tests proposed for testing the normality and the equivariance of two samples. Power computation has been carried out for the following alternatives (a) exponential distribution, (b) lognormal distribution, (c) Student’s t distribution, (d) Laplace distribution, (e) Weibull distribution, and (f) uniform distribution all with the same means and variance as the null distribution. However we present here tables for (a), (b)
and (c) only. Two independent samples from normal distribution with unequal variances are also used as an alternative. The type I error of the tests are also calculated. We study how the tests perform for change in means, change in variance and unequal sample sizes. The statistics to be used in the goodness-of-fit tests are the Kolmogorov-Smirnov statistic ($D$), the Kuiper statistic ($V$), the Anderson-Darling statistic ($A^2$) and the proposed T-statistic ($T$). Ten thousand Monte-Carlo samples of size $n_1$ and $n_2$ were drawn from each alternative and we count the number of times when the statistics quoted were found significant at levels $\alpha = 0.1$ and $\alpha = 0.05$. The values corresponding to $\alpha = 0.05$ are always tabulated in the second line of the tables. Finally we extend our power study to $k$ independent samples. We use the notation, $\mu_1, \mu_2, \ldots$ and $\sigma_1^2, \sigma_2^2, \ldots$ to represent means and variances of the distributions from which the samples are obtained respectively and $\sigma^2$ as the common variance. The results are presented in Tables 3.1 to 3.12, from which the following observations can be made.

(i) The EDF goodness-of-fit tests corresponding to the statistics $D$, $V$, $A^2$, and $T$, maintain very well the nominal level $\alpha = 0.1$ and $\alpha = 0.05$.

(ii) The powers of the tests increases considerably as the sample size increases except for the test based on $T$ statistic.

(iii) Out of the alternatives considered, the power of the tests does not get affected by change in mean or variance in exponential, Weibull and uniform distributions. But the power changes as the variance of Student’s $t$ distribution changes. Power of the tests are sensitive to both mean and variance when the alternative is lognormal distribution, specially in small sample sizes, but it does not change significantly for large sample sizes.

(iv) For unequal sample sizes, the powers of the tests reduce substantially. But for some
alternatives (exponential) the reduction in the power is not that significant.

(v) The Kuiper statistic ($V$) is recommended in testing this hypothesis especially for large sample sizes.

(vi) The $T$-statistic is recommended for small sample sizes.
Table 3.1: Estimated Type-I error of the tests, out of 10000 iterations, $\mu_1 = 1, \mu_2 = 5$

<table>
<thead>
<tr>
<th>$n_1 = n_2$</th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 1$</td>
<td>932 931 991 1087</td>
<td>968 905 994 1070</td>
<td>959 976 975 984</td>
</tr>
<tr>
<td></td>
<td>478 466 499 546</td>
<td>485 454 469 580</td>
<td>500 478 466 505</td>
</tr>
<tr>
<td>$\sigma^2 = 5$</td>
<td>926 937 958 1018</td>
<td>970 943 1019 1045</td>
<td>1027 963 1018 974</td>
</tr>
<tr>
<td></td>
<td>464 489 486 535</td>
<td>504 485 512 533</td>
<td>520 493 536 467</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>963 905 965 998</td>
<td>988 959 1014 994</td>
<td>1013 975 1034 986</td>
</tr>
<tr>
<td></td>
<td>478 442 477 506</td>
<td>499 494 491 478</td>
<td>520 477 509 509</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.2: Estimated Type-I error of the tests, out of 10000 iterations, $\mu_1 = 1, \mu_2 = 10$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D$</td>
<td>$V$</td>
<td>$A^2$</td>
</tr>
<tr>
<td>$\sigma^2 = 1$</td>
<td>1070</td>
<td>1023</td>
<td>1096</td>
</tr>
<tr>
<td></td>
<td>545</td>
<td>517</td>
<td>550</td>
</tr>
<tr>
<td>$\sigma^2 = 5$</td>
<td>965</td>
<td>945</td>
<td>994</td>
</tr>
<tr>
<td></td>
<td>490</td>
<td>472</td>
<td>465</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>1000</td>
<td>975</td>
<td>1024</td>
</tr>
<tr>
<td></td>
<td>518</td>
<td>506</td>
<td>485</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.3: Number of rejections out of 10000 - alternative is exponential distribution, $\mu_1 = 1, \mu_2 = 5$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 1$</td>
<td>1799 3174 2013 3147</td>
<td>3855 6749 4997 3459</td>
<td>8883 9882 9847 3784</td>
</tr>
<tr>
<td></td>
<td>1030 2196 997 2288</td>
<td>2514 5707 2975 2600</td>
<td>7736 9773 9322 2955</td>
</tr>
<tr>
<td>$\sigma^2 = 5$</td>
<td>1837 3229 2043 3044</td>
<td>3718 6657 4914 3539</td>
<td>8905 9890 9338 3808</td>
</tr>
<tr>
<td></td>
<td>1042 2183 1044 2185</td>
<td>2460 5590 2854 2672</td>
<td>7776 9786 9335 2964</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>1822 3277 2078 3176</td>
<td>3908 6789 5023 3526</td>
<td>8880 9882 9838 3768</td>
</tr>
<tr>
<td></td>
<td>1051 2267 1043 2338</td>
<td>2608 5701 3003 2692</td>
<td>7751 9785 9326 2911</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.4: Number of rejections out of 10000 - alternative is exponential distribution, $\mu_1 = 1, \mu_2 = 10$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$V$</td>
<td>$A^2$</td>
<td>$T$</td>
</tr>
<tr>
<td>$\sigma^2 = 1$</td>
<td>1805</td>
<td>3221</td>
<td>2039</td>
</tr>
<tr>
<td></td>
<td>1037</td>
<td>2206</td>
<td>978</td>
</tr>
<tr>
<td>$\sigma^2 = 5$</td>
<td>1863</td>
<td>3295</td>
<td>2026</td>
</tr>
<tr>
<td></td>
<td>1055</td>
<td>2250</td>
<td>997</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>1866</td>
<td>3305</td>
<td>2019</td>
</tr>
<tr>
<td></td>
<td>1043</td>
<td>2254</td>
<td>996</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.5: Number of rejections out of 10000 - alternative is lognormal distribution, $\mu_1 = 1, \mu_2 = 5$

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 1$</td>
<td>1433 2139 1389 3648</td>
<td>2439 4513 2511 4201</td>
<td>6209 8995 7192 4741</td>
</tr>
<tr>
<td></td>
<td>777 1310 629 2706</td>
<td>1505 3313 1303 3313</td>
<td>4619 8326 5322 3870</td>
</tr>
<tr>
<td>$\sigma^2 = 5$</td>
<td>2233 4167 1915 6395</td>
<td>5540 8562 5760 6995</td>
<td>9839 9995 9949 7459</td>
</tr>
<tr>
<td></td>
<td>1287 2984 880 5695</td>
<td>3892 7715 3640 6413</td>
<td>9451 9988 9751 6965</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>2857 5415 2467 7290</td>
<td>6946 9394 7296 7803</td>
<td>9981 10000 9998 8276</td>
</tr>
<tr>
<td></td>
<td>1670 4088 1124 6710</td>
<td>5226 8942 5139 7332</td>
<td>9912 10000 9970 7929</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.6: Number of rejections out of 10000 - alternative is lognormal distribution, $\mu_1 = 1, \mu_2 = 10$

<table>
<thead>
<tr>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 1$</td>
<td>4725</td>
<td>9534</td>
</tr>
<tr>
<td></td>
<td>8251</td>
<td>9991</td>
</tr>
<tr>
<td></td>
<td>3841</td>
<td>9564</td>
</tr>
<tr>
<td></td>
<td>9917</td>
<td>9994</td>
</tr>
<tr>
<td>$\sigma^2 = 5$</td>
<td>2106</td>
<td>5212</td>
</tr>
<tr>
<td></td>
<td>3891</td>
<td>8277</td>
</tr>
<tr>
<td></td>
<td>1595</td>
<td>4689</td>
</tr>
<tr>
<td></td>
<td>65.64</td>
<td>7101</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>1337</td>
<td>3661</td>
</tr>
<tr>
<td></td>
<td>1705</td>
<td>2704</td>
</tr>
<tr>
<td></td>
<td>1232</td>
<td>1625</td>
</tr>
<tr>
<td></td>
<td>2439</td>
<td>801</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.7: Number of rejections out of 10000 - alternative is location shifted Student’s $t$ distribution, $\mu_1 = 1, \mu_2 = 5$

<table>
<thead>
<tr>
<th>$n_1 = n_2$</th>
<th>$D$</th>
<th>$V$</th>
<th>$A^2$</th>
<th>$T$</th>
<th>$D$</th>
<th>$V$</th>
<th>$A^2$</th>
<th>$T$</th>
<th>$D$</th>
<th>$V$</th>
<th>$A^2$</th>
<th>$T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>3559</td>
<td>5201</td>
<td>4521</td>
<td>6585</td>
<td>7246</td>
<td>8574</td>
<td>8200</td>
<td>7669</td>
<td>9907</td>
<td>9985</td>
<td>9981</td>
<td>8548</td>
</tr>
<tr>
<td></td>
<td>2484</td>
<td>4327</td>
<td>2791</td>
<td>6009</td>
<td>6312</td>
<td>8109</td>
<td>7027</td>
<td>7235</td>
<td>9814</td>
<td>9970</td>
<td>9944</td>
<td>8265</td>
</tr>
<tr>
<td>20</td>
<td>1162</td>
<td>1334</td>
<td>1124</td>
<td>1254</td>
<td>1415</td>
<td>1922</td>
<td>1374</td>
<td>2454</td>
<td>2094</td>
<td>3231</td>
<td>2149</td>
<td>2735</td>
</tr>
<tr>
<td></td>
<td>608</td>
<td>752</td>
<td>553</td>
<td>2427</td>
<td>793</td>
<td>1212</td>
<td>694</td>
<td>1641</td>
<td>1249</td>
<td>2309</td>
<td>1196</td>
<td>1902</td>
</tr>
<tr>
<td>50</td>
<td>1008</td>
<td>1119</td>
<td>1014</td>
<td>1376</td>
<td>1082</td>
<td>1276</td>
<td>1122</td>
<td>1516</td>
<td>1240</td>
<td>1624</td>
<td>1195</td>
<td>1676</td>
</tr>
<tr>
<td></td>
<td>508</td>
<td>575</td>
<td>536</td>
<td>743</td>
<td>570</td>
<td>680</td>
<td>537</td>
<td>885</td>
<td>668</td>
<td>908</td>
<td>601</td>
<td>994</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line. $\nu$ is the degrees of freedom.
Table 3.8: Number of rejections out of 10000 - alternative is location shifted Student’s $t$ distribution, $\mu_1 = 1, \mu_2 = 10$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D$</td>
<td>$V$</td>
<td>$A^2$</td>
</tr>
<tr>
<td>$\nu = 1$</td>
<td>3690</td>
<td>5285</td>
<td>4608</td>
</tr>
<tr>
<td></td>
<td>2569</td>
<td>4362</td>
<td>2559</td>
</tr>
<tr>
<td>$\nu = 5$</td>
<td>1173</td>
<td>1352</td>
<td>1162</td>
</tr>
<tr>
<td></td>
<td>637</td>
<td>740</td>
<td>589</td>
</tr>
<tr>
<td>$\nu = 10$</td>
<td>1075</td>
<td>1137</td>
<td>1032</td>
</tr>
<tr>
<td></td>
<td>523</td>
<td>573</td>
<td>524</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line. $\nu$ is the degrees of freedom.
Table 3.9: Number of rejections out of 10000 - samples are from different distributions $\mu_1 = 1, \mu_2 = 5, \sigma^2 = 10$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = 10$</th>
<th></th>
<th>$n_1 = n_2 = 20$</th>
<th></th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D$</td>
<td>$V$</td>
<td>$A^2$</td>
<td>$T$</td>
<td>$D$</td>
</tr>
<tr>
<td>normal / exponential</td>
<td>1249</td>
<td>1773</td>
<td>1227</td>
<td>2353</td>
<td>1758</td>
</tr>
<tr>
<td>(normal / lognormal)</td>
<td>623</td>
<td>1027</td>
<td>613</td>
<td>1568</td>
<td>994</td>
</tr>
<tr>
<td>exponential)</td>
<td>1175</td>
<td>1645</td>
<td>1270</td>
<td>2282</td>
<td>1627</td>
</tr>
<tr>
<td>(lognormal)</td>
<td>614</td>
<td>975</td>
<td>639</td>
<td>1457</td>
<td>888</td>
</tr>
<tr>
<td>exponential)</td>
<td>1626</td>
<td>2820</td>
<td>1820</td>
<td>3075</td>
<td>3299</td>
</tr>
<tr>
<td>lognormal</td>
<td>904</td>
<td>1837</td>
<td>855</td>
<td>2209</td>
<td>2153</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.10: Number of rejections out of 10000 - samples are from normal distribution but having different variances, $\mu_1 = 1, \mu_2 = 5$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = 10$</th>
<th>$n_1 = n_2 = 20$</th>
<th>$n_1 = n_2 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1^2 = 10, \sigma_2^2 = 20$</td>
<td>898  933  758  2557</td>
<td>1020  1386  870  4315</td>
<td>1386  2570  1417  7679</td>
</tr>
<tr>
<td></td>
<td>434  456  356  1598</td>
<td>501  789  365  3120</td>
<td>706  1601  639  6659</td>
</tr>
<tr>
<td>$\sigma_1^2 = 10, \sigma_2^2 = 50$</td>
<td>1025  1511  662  7444</td>
<td>1755  3796  1428  9656</td>
<td>4807  8398  5773  9999</td>
</tr>
<tr>
<td></td>
<td>459  815  265  6245</td>
<td>873  2624  524  9314</td>
<td>3010  7494  3566  9996</td>
</tr>
<tr>
<td>$\sigma_1^2 = 10, \sigma_2^2 = 100$</td>
<td>1318  2555  730  9508</td>
<td>3315  6715  2643  9990</td>
<td>8615  9905  9142  10000</td>
</tr>
<tr>
<td></td>
<td>614  1496  264  9083</td>
<td>1833  5380  1054  9972</td>
<td>7014  9799  7801  10000</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
Table 3.11: Number of rejections out of 10000 - unequal sample sizes, \( \mu_1 = 1, \mu_2 = 5, \sigma^2 = 10 \)

<table>
<thead>
<tr>
<th></th>
<th>( n_1 = 0.5, n_2 = 15 )</th>
<th>( n_1 = 10, n_2 = 30 )</th>
<th>( n_1 = 20, n_2 = 80 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D )</td>
<td>1825</td>
<td>1766</td>
<td>1766</td>
</tr>
<tr>
<td>( V )</td>
<td>3552</td>
<td>2615</td>
<td>2615</td>
</tr>
<tr>
<td>( A^2 )</td>
<td>2081</td>
<td>1967</td>
<td>1967</td>
</tr>
<tr>
<td>( T )</td>
<td>2804</td>
<td>7152</td>
<td>3236</td>
</tr>
</tbody>
</table>

Exponential 1033 2523 1056 1930 2463 5766 2884 2480 7721 9784 9336 2723

Lognormal 984 1720 979 64.42 2747 5012 2504 7508 6859 9250 8082 7991

Student’s \( t \) 1766 2615 1791 3236 3194 4542 3376 4346 6501 8083 7122 5408

\( \alpha = 0.05 \) are tabulated in the second line.
Table 3.12: Number of rejections out of 10000 - three samples case $\mu_1 = 1, \mu_2 = 5, \mu_3 = 10, \sigma^2 = 10$

<table>
<thead>
<tr>
<th></th>
<th>$n_1 = n_2 = n_3 = 10$</th>
<th>$n_1 = n_2 = n_3 = 20$</th>
<th>$n_1 = n_2 = n_3 = 50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>$V$</td>
<td>$A^2$</td>
<td>$T$</td>
</tr>
<tr>
<td>Exponential</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2276</td>
<td>4229</td>
<td>2555</td>
<td>3245</td>
</tr>
<tr>
<td>1342</td>
<td>3066</td>
<td>1268</td>
<td>2391</td>
</tr>
<tr>
<td>Lognormal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2742</td>
<td>4740</td>
<td>2249</td>
<td>7872</td>
</tr>
<tr>
<td>1597</td>
<td>3468</td>
<td>1059</td>
<td>7384</td>
</tr>
<tr>
<td>Student’s $t$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2131</td>
<td>3089</td>
<td>2324</td>
<td>4182</td>
</tr>
<tr>
<td>1287</td>
<td>2289</td>
<td>1396</td>
<td>3364</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
CHAPTER 4
CHARACTERIZATION OF \( k \)-VARIATE NORMAL DISTRIBUTION WITH COVARIANCE STRUCTURE, \( \sigma^2 \Sigma_0 \) AND EDF GOODNESS-OF-FIT TESTS

4.1 Introduction

In the multivariate linear regression, we fit the model

\[
Y_i = \beta X_i + \epsilon_i, \quad i = 1, \ldots, n,
\]

where \( Y_i \) and \( \epsilon_i \) are \( k \) component response vector variables and \( k \) component error random vectors respectively. In the inference of this model, we assume \( \epsilon_i \), are identically and independently distributed normally with mean \( 0 \) and variance-covariance \( \Sigma \). Suppose \( \Sigma \) is known up to a scalar multiple, i.e. \( \Sigma = \sigma^2 \Sigma_0 \) with \( \Sigma_0 \) known. Then to test whether a set of observed data, \( Y_i \), are coming from the above regression model, we need to construct a test for testing the hypothesis that \( Y_i \) is from \( N(\beta X_i, \sigma^2 \Sigma_0) \), \( i = 1, \ldots, n \).

For testing multivariate normality there is no test that is widely used. Most tests depend on asymptotic distribution theory such as \( \chi^2 \) tests with estimated parameters. Mar-dia (1970) proposed tests based on multivariate skewness and kurtosis. Rincon-Gallardo, Quesenberry and O’Reilly (1979) gave transformations which can be used to construct exact EDF goodness-of-fit tests for multivariate normality. Characterization of \( k \)-variate normal distribution based on the conditional moments is studied by many authors in recent years. Gupta and Varga (1992) gave characterization of matrix variate normality through conditional distributions. Nguyen, Nguyen and Dinh (2004) gave a characterization of \( k \)-variate normal distribution based on the UMVU estimate of the density function.
In this chapter, in Section 2, we give a characterization of $k$-variate normal distribution, with partially known covariance, based on UMVU estimator of the density function. This characterization has been changed to a characterization based on Student’s law, using the transformation given by Rincon-Gallardo, Quesenberry and O’Reilly (1979). An alternative method is also proposed based on the results given in Chapters 2 and 3. Application of these results to EDF goodness-of-fit tests for testing $k$-variate normality with partially known covariance is studied in Section 3. Power comparison of the two methods is carried out in Section 4.

4.2 Characterization result

Let $X_j = (X_{ji})', i = 1, \ldots, k, j = 1, \ldots, n, n \geq k + 2$, be a random sample from a $k$-variate normal distribution with an unknown mean vector $\boldsymbol{\xi} = (\xi_i)', i = 1, \ldots, k,$ and partially known covariance matrix $\sigma^2 \Sigma_0$ where $\Sigma_0$ is a known $k \times k$ positive definite matrix and $\sigma^2$ is an unknown positive scalar. Then $\bar{X} = (1/n) \sum_{j=1}^{n} X_j$ and $s = \sum_{j=1}^{n} (X_j - \bar{X})' \Sigma_0^{-1} (X_j - \bar{X})$ are jointly complete and sufficient statistics for $\boldsymbol{\xi}$ and $\sigma^2$. The UMVUE of the density function of $X_1$ at a given point $x_1$ is (Ghurye and Olkin (1969))

$$f_{X_1|\bar{X},s}(x_1|\bar{X},s) = K_1(n) \frac{1 - \frac{n}{n-1} (x_1 - \bar{x})' \Sigma_0^{-1} (x_1 - \bar{x})/s}{s^{\frac{k}{2}}}$$

for $\frac{n}{n-1} (x_1 - \bar{x})' \Sigma_0^{-1} (x_1 - \bar{x})/s < 1$, (4.2.1)

where $K_1(n) = \frac{n^{\frac{k}{2}} \Gamma\left(\frac{(n-1)k}{2}\right)}{\pi^{\frac{k}{2}} \Gamma\left(\frac{(n-2)k}{2}\right) \Sigma_0^{\frac{k}{2}}}$. Let $U = \sqrt{\frac{n}{n-1}} (s \Sigma_0)^{-\frac{1}{2}} (X_1 - \bar{X})$. In the conditional model (4.2.1) the Jacobian of the...
transformation is \((n-1)s/n \) and

\[
f_{U|X,s}(u|x, s) = \frac{\Gamma\left[\frac{(n-1)k}{2}\right]}{\pi^{k} \Gamma\left[\frac{(n-2)k}{2}\right]} \left(1 - \sum_{i=1}^{k} u_{i}^{2}\right)^{\frac{(n-2)k-2}{2}}
\]

for \( \sum_{i=1}^{k} u_{i}^{2} < 1 \),

where \( U = (U_{1}, \ldots, U_{k})' \).

From (4.2.2), it is noted that the density of \( U \) does not depend on \( X \) and \( s \). Then for any non-zero vector \( a = (a_{1}, \ldots, a_{k})' \) of \( R^{k} \), \( U \) is independent of \( a'X \) and \( s \). Hence, the conditional density function of \( U \) given \( a'X \) and \( s \) is also given by the right hand side of (4.2.2). Without loss of generality, assume that \( a_{k} \neq 0 \), and let \( W = a'U \). The transformation from \((U_{1}, \ldots, U_{k})'\) to \((U_{1}, \ldots, U_{k-1}, W)'\) has the Jacobian \( 1/|a_k| \), and the conditional joint density function of \( U_{1}, \ldots, U_{k-1}, W \) given \( a'X \) and \( s \) is obtained as

\[
f_{U_{1},\ldots,U_{k-1},W|a'X,s}(u_{1}, \ldots, u_{k-1}, w|a'X, s)
= \frac{\Gamma\left[\frac{(n-1)k}{2}\right]}{\pi^{k} \Gamma\left[\frac{(n-2)k}{2}\right] |a_k|} \left(1 - \sum_{i=1}^{k-1} u_{i}^{2} - \left(\frac{w - \sum_{i=1}^{k-1} a_{i}u_{i}}{a_{k}^{2}}\right)^{2}\right)^{\frac{(n-2)k-2}{2}}
\]

for \( \sum_{i=1}^{k-1} u_{i}^{2} \) and \( \left(\frac{w - \sum_{i=1}^{k-1} a_{i}u_{i}}{a_{k}^{2}}\right)^{2} < 1 \).

From (4.2.3), the conditional joint density function of \( U_{2}, \ldots, U_{k-1}, W \) given \( a'X \) and \( s \) is obtained as

\[
f_{U_{2},\ldots,U_{k-1},W|a'X,s}(u_{2}, \ldots, u_{k-1}, w|a'X, s)
\]
\[ f_{W|a'x,s}(a'x, s) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n-1}{2}\right) |a|} \left[ 1 - \frac{w^2}{|a|^2} \right]^\frac{(n-1)k-3}{2} \] for \( w^2 < |a|^2 \). (4.2.5)

A characterization of \( k \)-variate normal distribution based on the third conditional moment is given in the following theorem. In the proof of the theorem, the following lemma is used:

\[ \text{IE} \left[ (a'x_1 - a'\bar{x})^3 | a'x_1, s \right] = 0. \] (4.2.6)
Lemma 4.2.1. (Nguyen and Dinh (1998)) A characteristic function \( \phi \) of a non-degenerate distribution is the characteristic function of a normal distribution if and only if it is a solution of the differential equation \( \phi''' \phi^2 - 3 \phi'' \phi' + 2 (\phi')^3 = 0 \).

Theorem 4.2.1. Let \( X_j = (X_{ji})', \ i = 1, \ldots, k, \ j = 1, \ldots, n, \ n \geq k + 2, \) be a random sample from a non-singular \( k \)-variate distribution \( F \) with all finite third moments. Then \( F \) is a \( N(\xi, \sigma^2 \Sigma_0) \) distribution if and only if the conditional density function of \( X_1 \) given \( X \) and \( s \) is given by (4.2.1).

Proof. We only need to prove that the conditional density function of \( X_1 \) given \( X \) and \( s \), which is (4.2.1), implies normality of \( X_j \)'s. Since (4.2.1) implies (4.2.6), it is sufficient to show that \( \mathbb{E} \left[ (a'X_1 - a'X)^3 | a'X, s \right] = 0 \), implies normality of \( X_j \)'s, for an arbitrary vector \( a = (a_1, \ldots, a_k)' \neq 0 \).

Let \( \mathbb{E} \left[ (a'X_1 - a'X)^3 | a'X, s \right] = 0 \). Taking the expectation w.r.t. \( s \), we get

\[
\mathbb{E} \left[ (a'X_1 - a'X)^3 | a'X \right] = \mathbb{E}_s \left[ \mathbb{E} \left[ (a'X_1 - a'X)^3 | a'X, s \right] \right] = 0.
\]

Now letting \( V_j = a'X_j \) and \( \overline{V} = (1/n) \sum_{j=1}^n V_j \) we have

\[
\mathbb{E} \left[ (V_1 - \overline{V})^3 | \overline{V} \right] = 0. \tag{4.2.7}
\]

Multiplying both sides of (4.2.7) by \( e^{i t \overline{V}} \), and then taking the expectation, we get

\[
\mathbb{E}_\overline{V} \left[ e^{i t \overline{V}} \mathbb{E} \left[ (V_1 - \overline{V})^3 | \overline{V} \right] \right] = \mathbb{E} \left[ e^{i t \overline{V}} (V_1 - \overline{V})^3 \right] = 0
\]

or

\[
\mathbb{E} \left[ e^{i t \overline{V}} \left( V_1^3 - 3 V_1^2 \overline{V} + 3 V_1 \overline{V}^2 - \overline{V}^3 \right) \right] = 0. \tag{4.2.8}
\]

Next let \( \phi(t) = \mathbb{E} \left[ e^{i t V} \right] \) be the characteristic function of the distribution of \( V \). Then from (4.2.8), we get
\[
\begin{align*}
\mathbb{E} \left[ e^{itnV_1^3} \right] &= i\phi''' \phi^{n-1} \quad (4.2.9) \\
-3\mathbb{E} \left[ e^{itnV_1^2 V_1} \right] &= -\frac{3}{n} \mathbb{E} \left[ e^{itnV_1^2} \sum V_i \right] \\
&= -\frac{3i}{n} \left[ \phi''' \phi^{n-1} + (n-1)\phi'' \phi' \phi^{n-2} \right] \quad (4.2.10) \\
3\mathbb{E} \left[ e^{itnV_1 V_1^2} \right] &= \frac{3}{n^2} \mathbb{E} \left[ e^{itnV_1} \left( \sum V_i^2 + 2 \sum_{i<j} V_i V_j \right) \right] \\
&= \frac{3i}{n^2} \left[ \phi''' \phi^{n-1} + 3(n-1)\phi'' \phi' \phi^{n-2} + (n-1)(n-2)(\phi')^3 \phi^{n-3} \right] \quad (4.2.11) \\
-\mathbb{E} \left[ e^{itnV^3} \right] &= -\frac{1}{n^3} \mathbb{E} \left[ e^{itnV} \left( \sum V_i^3 + 3 \sum_{i<j} V_i^2 V_j + 3 \sum_{i<j} V_i V_j^2 + 6 \sum_{i<j<k} V_i V_j V_k \right) \right] \\
&= -\frac{i}{n^3} \left[ n\phi''' \phi^{n-1} + 3n(n-1)\phi'' \phi' \phi^{n-2} + n(n-1)(n-2)(\phi')^3 \phi^{n-3} \right] \\
&= \frac{n^3}{3i} \left[ \phi''' \phi^{n-1} + 3(n-1)\phi'' \phi' \phi^{n-2} + (n-1)(n-2)(\phi')^3 \phi^{n-3} \right] \quad (4.2.12)
\end{align*}
\]

Substituting from (4.2.9) to (4.2.12) in (4.2.8) and simplifying, we get the following differential equation,

\[ \phi''' \phi^2 - 3\phi'' \phi' \phi + 2(\phi')^3 = 0 \quad (4.2.13) \]

Hence, by Lemma 4.2.1, the distribution \( F \) is \( N(\xi, \sigma^2 \Sigma_0) \).

Let \( X_j = (X_{ji})', \ i = 1, \ldots, k, \ j = 1, \ldots, n, \ n \geq k + 2 \), be a random sample from a \( k \)-variate normal distribution with an unknown mean vector \( \xi = (\xi_i)', \ i = 1, \ldots, k \), and partially known covariance matrix \( \sigma^2 \Sigma_0 \) where \( \Sigma_0 \) is a known \( k \times k \) positive definite matrix and \( \sigma^2 \) is an unknown positive scalar. Then the following one-to-one transformations change a random sample from the \( k \)-variate \( N(\xi, \sigma^2 \Sigma_0) \) to a set of independent random variables with Student’s \( t \) distributions.
Let
\[ Z_j = \frac{A (X_j - \bar{X}_j)}{\left[ ((j - 1) s_j/j) - (X_j - \bar{X}_j)' \Sigma_0^{-1} (X_j - \bar{X}_j) \right]^{1/2}} \] (4.2.14)
where \( A' A = \Sigma_0^{-1} \), \( s_j = \text{tr} \Sigma_0^{-1} S_j \), \( \bar{X}_j = (1/j) \sum_{i=1}^{j} X_i \) and \( S_j = \sum_{i=1}^{j} (X_i - \bar{X}_j) \).

\((X_i - \bar{X}_j)'\) for \( j = 3, \ldots, n \),
and
\[ W_{i,j} = Z_{i,j} \left[ \frac{(j - 2) k + i - 1}{1 + Z_{i,j}^2 + \cdots + Z_{i-1,j}^2} \right]^{1/2}. \] (4.2.15)
for \( i = 1, \ldots, k \) and \( j = 3, \ldots, n \). Then \( W_{i,j} \) are independently distributed according to univariate Student-\( t \) distribution with \((j - 2)k + i - 1\) degrees of freedom (Rincon-Gallardo, Quesenberry and O’Reilly (1979)).

We call this set of transformations as Transformation 1.

**Theorem 4.2.2.** Let \( X_j = (X_{ji})' \), \( i = 1, \ldots, k \), \( j = 1, \ldots, n \), \( n \geq k+2 \), be a random sample from a non-singular \( k \)-variate distribution \( F \) with all finite third moments. In the conditional model given \( \bar{X} \) and \( s \), let \( W_{i,j}, i = 1, \ldots, k, j = 3, \ldots, n \), obtained by the sequence of one-to-one transformations defined by (4.2.14) and (4.2.15). Then \( F \) is a \( N(\xi, \sigma^2 \Sigma_0) \) distribution if and only if the conditional density function of \( W_{i,j}, i = 1, \ldots, k, j = 3, \ldots, n \) are independently distributed according to the univariate Student-\( t \) distribution with \((j - 2)k + i - 1\) degrees of freedom.

**Proof.** The result follows from Theorem 4.2.1 and the result given by Rincon-Gallardo, Quesenberry and O’Reilly (1979). \( \square \)

Next an alternative approach for the same problem is given. We know \( X_j = (X_{ji})' \), \( i = 1, \ldots, k \), \( j = 1, \ldots, n \) are distributed \( N(\xi, \sigma^2 \Sigma_0) \) if and only if \( Y_j = (Y_{ji})' \), \( i = 1, \ldots, k \), \( j =
1, ..., n are distributed $N(\xi^*, \sigma^2 I)$ where $Y_j = \Sigma_0^{-\frac{1}{2}} X_j$ and $\xi^* = (\xi_1^*)' = \Sigma_0^{-\frac{1}{2}} \xi$. These $Y_{ji}$'s form k independent samples from normal distributions with different means but same variance. That is $Y_{ji} \sim N(\xi_i^*, \sigma^2)$ for $i = 1, ..., k$, $j = 1, ..., n$. We have dealt with this problem in chapters 2 and 3. We gave characterization results based on second conditional moments. We proposed the following one-to-one transformations which change $Y_j = (Y_{ji})'$, $i = 1, ..., k$, $j = 1, ..., n$ to a set of independent Student’s $t$ random variables.

Let

$$u_j = \frac{y_{j1} - z_1}{\sqrt{z_{k+1}}}, \quad j = 1, ..., n - 2,$$

$$u_{(i-1)n-i+j} = \frac{y_{ji} - z_i}{\sqrt{z_{k+1}}}, \quad i = 2, ..., k, \quad j = 1, ..., n - 1,$$  \hspace{1cm} (4.2.16)

where $z_i = \frac{\sum_{j=1}^{n} y_{ji}}{n}$, $i = 1, ..., k$, $z_{k+1} = \sum_{j=1}^{k+1} \sum_{j=1}^{n} (y_{ji} - z_i)^2$.

$$v_j = \sqrt{\frac{n-j+1}{n-j}} u_j + \frac{\sum_{i=1}^{j-1} u_i}{\sqrt{(n-j+1)(n-j)}}$$

$$\sqrt{1 - \sum_{i=1}^{j-1} u_i^2 - \left(\sum_{i=1}^{j-1} u_i\right)^2} \quad j = 1, ..., n - 2,$$

$$v_{n-2+j} = \sqrt{\frac{n-j+1}{n-j}} u_{n-2+j} + \frac{\sum_{i=1}^{j-1} u_{n-2+i}}{\sqrt{(n-j+1)(n-j)}}$$

$$\sqrt{1 - \frac{3}{2} \sum_{i=1}^{n-2} u_i^2 - \sum_{1 \leq i' \leq n-2} u_i u_i' - \sum_{i=1}^{j-1} u_{n-2+i} - \left(\sum_{i=1}^{j-1} u_{n-2+i}\right)^2} \quad j = 1, ..., n - 1,$$

$$v_{(i-1)n-i+j} = \sqrt{\frac{n-j+1}{n-j}} u_{(i-1)n-i+j} + \frac{\sum_{k=1}^{i-1} u_{(i-1)n-i+k}}{\sqrt{(n-j+1)(n-j)}}$$

$$\sqrt{1 - \frac{3}{2} \sum_{k=1}^{n-2} u_k^2 - \sum_{1 \leq k' \leq n-2} u_k u_k' - 2 \sum_{k=3}^{i} \sum_{l=1}^{n-1} u_{(k-2)n-(k-1)+l} - t_1 - t_2} \quad i = 3, ..., k, \quad j = 1, ..., n - 1,$$  \hspace{1cm} (4.2.17)
where

\[ t_1 = 2 \sum_{k=3}^{i} \sum_{1 \leq l < l' \leq n-1} u_{(k-2)n-(k-1)+l} u_{(k-2)n-(k-1)+l'}; \]

\[ t_2 = \sum_{k=1}^{j-1} u_{(i-1)n-k}^2 = \frac{\left( \sum_{k=1}^{j-1} u_{(i-1)n-k} \right)^2}{n - j + 1}. \]

\[ w_i = \sqrt{n k - i - k} \frac{v_i}{\sqrt{1 - v_i^2}} \quad i = 1, \ldots, nk - k - 1. \]  

(4.2.18)

Then \( W_i, i = 1, \ldots, nk - k - 1 \) are independently distributed according to the Student’s t distribution with \( nk - i - k \) degrees of freedom. In this case we lose less data than in the previous case specially for high dimensional data.

Let’s call this set of transformations as Transformation 2.

### 4.3 Application to goodness-of-fit tests

Let \( X_j = (X_{ji})', i = 1, \ldots, k, j = 1, \ldots, n \), \( n \geq k + 2 \), be a random sample from a non-singular \( k \)-variate distribution \( F \) with all finite third moments. Consider testing the composite hypothesis

\[ H_0 : F \text{ is a } N(\xi, \sigma^2 \Sigma_0) \quad \text{versus} \quad H_a : \text{not } H_0 \]  

(4.3.1)

where \( \xi, \sigma^2 \) are unknown and \( \Sigma_0 \) is known.

Using the Theorem 4.2.2, testing (4.3.1) is equivalent to testing;

\[ H_0^* : W_{i,j}, i = 1, \ldots, k, j = 3, \ldots, n \text{ are conditionally independently} \]

\[ \text{distributed according to univariate Student-t distribution with} \]

\( ((j - 2)k + i - 1) \) degrees of freedom.

Also, using the results of chapters 2 and 3, we obtain the equivalent hypothesis,
\( H_0^{**} \): Given \( Z_1, Z_2, Z_3, \ W_i, \ i = 1, \ldots, m + n - 3 \) are independently distributed according to Student’s \( t \) distribution with 
\( m + n - i - 2 \) degrees of freedom.

These simple hypotheses can be tested using any EDF statistic, such as Kolmogorov-Smirnov statistic, Kuiper statistic, Cramer-vonMises statistic, Anderson-Darling statistic (See D’Agostino and Stephens (1986)).

### 4.4 Power study

In this section we estimate the power of the tests proposed for testing the multivariate normality when the mean \( \mu \) is unknown and the covariance \( \Sigma = \sigma^2 \Sigma_0 \) is partially known up to a scalar multiple. For the bivariate case, the null distribution used is a normal distribution with \( \mu = (1 \ 3)' \), \( \sigma^2 = 3 \), \( \Sigma_0 = \begin{pmatrix} 2 & 3 \ 1 & 5 \end{pmatrix} \) and when the dimension is three \((k = 3)\), we used normal distribution with \( \mu = (1 \ 3 \ 5)' \), \( \sigma^2 = 3 \), \( \Sigma_0 = \begin{pmatrix} 2 & 3 & 5 \ 1 & 5 & 4 \ 2 & 3 & 3 \end{pmatrix} \) as the null distribution. The set of alternatives consists of multivariate \( t \) distribution with same mean and the covariance as in the null distribution and mixture of multivariate normal distributions having different means but the same covariance structure as in the null distribution. We generated independent exponential variates and correlated them using the same covariance structure as in the null distribution. The resulting variates were then used as the alternative. We considered also multivariate normal distribution with \( \Sigma \neq \sigma^2 \Sigma_0 \) as covariance structure, as an alternative.
The statistics to be used are the Kolmogorov-Smirnov statistic \( D \), the Kuiper statistic \( V \), the Cramer-vonMises statistic \( W^2 \) and the Anderson-Darling statistic \( A^2 \). Ten thousand Monte-Carlo samples of size \( n(= 5, 10, 20, 50) \) were drawn from each alternative and using the above test statistics, we counted the number of rejections at levels \( \alpha(= 0.10, 0.05, 0.01) \). The type I error of the tests were also found. These values are tabulated in Tables 4.1 to 4.10. The values corresponding to Transformation 2 are always tabulated in the second line of the tables. The following observations are made based on the simulation.

(i) The EDF statistics \( D, V, W^2, \) and \( A^2 \), maintain very well the nominal level \( \alpha = 0.1, \alpha = 0.05 \) and \( \alpha = 0.01 \).

(ii) Between the two transformations, Transformation 2 performs better, as expected since it loses less data than Transformation 1. Further, Transformation 2 performs substantially better than Transformation 1 for high dimensional data.

(iii) Among the EDF goodness-of-fit test statistics used, the Kuiper \( (V) \) statistic is superior to the others in testing multivariate normality with partially known covariance structure.

(iv) Transformation 1 performs very poorly when the alternative is multivariate normal distribution with covariance structure \( \Sigma \neq \sigma^2 \Sigma_0 \), for some \( \Sigma_0 \).
Table 4.1: Estimated Type-I error of the tests, out of 10000 iterations, when \( k=2 \).

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>( \alpha )</td>
<td>( D )</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>948</td>
</tr>
<tr>
<td></td>
<td></td>
<td>979</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>492</td>
</tr>
<tr>
<td></td>
<td></td>
<td>478</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>925</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>515</td>
</tr>
<tr>
<td></td>
<td></td>
<td>475</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>1026</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>494</td>
</tr>
<tr>
<td></td>
<td></td>
<td>508</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>916</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1023</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>447</td>
</tr>
<tr>
<td></td>
<td></td>
<td>507</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.2: Estimated Type-I error of the tests, out of 10000 iterations, when k=3.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>α</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>964</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1011</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>466</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>516</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>998</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>991</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>977</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>486</td>
</tr>
<tr>
<td></td>
<td></td>
<td>490</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>974</td>
</tr>
<tr>
<td></td>
<td></td>
<td>485</td>
</tr>
<tr>
<td></td>
<td></td>
<td>507</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.3: Number of rejections out of 10000 - alternative is multivariate $t$ distribution, $k=2$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>1275</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1065</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>714</td>
</tr>
<tr>
<td></td>
<td></td>
<td>541</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>1720</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1654</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1079</td>
</tr>
<tr>
<td></td>
<td></td>
<td>976</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>3208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.99</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>2263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2413</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>6783</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7606</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>5744</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6576</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.4: Number of rejections out of 10000 - alternative is multivariate $t$ distribution, $k=3$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>416</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1442</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td></td>
<td>695</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>919</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4784</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>449</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3317</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>2679</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8942</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1840</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8054</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>7340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9999</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6406</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9995</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.5: Number of rejections out of 10000 - alternative is mixture of multivariate normal distributions, mixing proportion $p = 0.25$, $\bm{\mu}_1 = [1 \ 3]'$, $\bm{\mu}_2 = [20 \ 60]'$, $k=2$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$\alpha$</td>
<td>$D$</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>3040</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1942</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>2360</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1359</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>4503</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3324</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>3422</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2241</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>7884</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7819</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6697</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6332</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>9993</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9989</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>9952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9968</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.6: Number of rejections out of 10000 - alternative is mixture of multivariate normal distributions, mixing proportion \( p = 0.25 \), \( \mu_1 = [1\ 3\ 5]' \), \( \mu_2 = [20\ 60\ 100]' \), \( k=3 \).

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>( \alpha )</td>
<td>( D )</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>419</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4676</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1733</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>1617</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9480</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>739</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9444</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>5868</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9968</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>4020</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9967</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>9997</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10000</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>9970</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10000</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.7: Number of rejections out of 10000 - alternative is Exponential distribution, k=2.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>α</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>943</td>
</tr>
<tr>
<td></td>
<td></td>
<td>560</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>405</td>
</tr>
<tr>
<td></td>
<td></td>
<td>209</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>1625</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1222</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>939</td>
</tr>
<tr>
<td></td>
<td></td>
<td>563</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>3484</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7022</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>2306</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2213</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>7924</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9227</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>6773</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8296</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.8: Number of rejections out of 10000 - alternative is Exponential distribution, \( k=3 \).

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>( \alpha )</th>
<th>( D )</th>
<th>( V )</th>
<th>( W^2 )</th>
<th>( A^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.10</td>
<td>2355</td>
<td>2121</td>
<td>2175</td>
<td>2929</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2470</td>
<td>1726</td>
<td>2554</td>
<td>3095</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1442</td>
<td>1318</td>
<td>1340</td>
<td>1797</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1705</td>
<td>1017</td>
<td>1793</td>
<td>2145</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>3180</td>
<td>3920</td>
<td>2763</td>
<td>3931</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2876</td>
<td>2853</td>
<td>2816</td>
<td>3517</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>2185</td>
<td>2847</td>
<td>1793</td>
<td>2526</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2017</td>
<td>1973</td>
<td>1940</td>
<td>2317</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>5301</td>
<td>7271</td>
<td>4733</td>
<td>6421</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4142</td>
<td>5404</td>
<td>3987</td>
<td>5478</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>4082</td>
<td>6165</td>
<td>3288</td>
<td>4517</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3079</td>
<td>4333</td>
<td>2800</td>
<td>3681</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>9091</td>
<td>9912</td>
<td>9146</td>
<td>9796</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7644</td>
<td>9256</td>
<td>7765</td>
<td>9494</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>8339</td>
<td>9827</td>
<td>8142</td>
<td>9406</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6546</td>
<td>8782</td>
<td>6392</td>
<td>8581</td>
<td></td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.9: Number of rejections out of 10000 - alternative is multivariate normal distribution with $\Sigma \neq \sigma^2 \Sigma_0$, for some $\Sigma_0$, $k=2$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>EDF goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$D$</td>
</tr>
<tr>
<td>5</td>
<td>0.10</td>
<td>2689</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1609</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1734</td>
</tr>
<tr>
<td></td>
<td></td>
<td>958</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>2388</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2038</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1527</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1219</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>2401</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3575</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1521</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2251</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>2678</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8321</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6643</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
Table 4.10: Number of rejections out of 10000 - alternative is multivariate normal distribution with $\Sigma \neq \sigma^2 \Sigma_0$, for some $\Sigma_0$, $k=3$.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Significant level</th>
<th>$D$</th>
<th>$V$</th>
<th>$W^2$</th>
<th>$A^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.10</td>
<td>210</td>
<td>1468</td>
<td>131</td>
<td>614</td>
</tr>
<tr>
<td></td>
<td></td>
<td>351</td>
<td>2463</td>
<td>56</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>57</td>
<td>902</td>
<td>22</td>
<td>261</td>
</tr>
<tr>
<td></td>
<td></td>
<td>56</td>
<td>1429</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>165</td>
<td>1364</td>
<td>120</td>
<td>452</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2811</td>
<td>7896</td>
<td>2389</td>
<td>2433</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>35</td>
<td>847</td>
<td>25</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1136</td>
<td>6797</td>
<td>624</td>
<td>629</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>213</td>
<td>1516</td>
<td>180</td>
<td>420</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8522</td>
<td>9951</td>
<td>9127</td>
<td>9295</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>57</td>
<td>963</td>
<td>52</td>
<td>152</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6726</td>
<td>9869</td>
<td>7400</td>
<td>7884</td>
</tr>
<tr>
<td>50</td>
<td>0.10</td>
<td>367</td>
<td>2084</td>
<td>347</td>
<td>532</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
<td>10000</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>127</td>
<td>1404</td>
<td>106</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9999</td>
<td>10000</td>
<td>9999</td>
<td>10000</td>
</tr>
</tbody>
</table>

The values corresponding to Transformation 2 are tabulated in the second line.
CHAPTER 5
APPROXIMATE EDF GOODNESS-OF-FIT TESTS FOR TESTING THE DISTRIBUTIONAL ASSUMPTIONS OF ONE-WAY REGRESSION MODEL

5.1 Introduction

In the one-way classification with \( k \) treatments, we assume that the data are observed according to the additive model, where the error random variables are independent and identically normally distributed with mean zero and constant variance. Then to test whether a set of observed data comes from the above regression model, we need to construct a test for testing the hypothesis that the observations of the treatment groups are normally distributed with same variance. Since this is a composite hypothesis (means and variance of the distribution are unknown) we can not use the exact EDF goodness-of-fit tests for testing the distributional assumptions of the model. It would be natural to use the sample means and the pooled sample variance as estimates of the unknown parameters and then perform probability integral transformation on observed variables. However, even when the null hypothesis is true, the resulting new variables will now not be a sample from the uniform distribution with limits 0, 1, and the critical points of the EDF statistics will be depend on the method of estimation and the sample size. This kind of testing criteria, by estimating the unknown parameters using the sample estimates, is called approximate EDF goodness-of-fit tests in statistics.

Using the EDF statistics for testing the normality of a sample, by estimating the unknown parameters using the sample estimates, has been studied by several authors. The
percentage points for EDF statistics, Cramer-von Mises ($W^2$), and Anderson-Darling ($A^2$), were given by Stephens (1974, 1976). Monte Carlo studies for Kolmogorov-Smirnov statistic ($D$) were given by Van Soest (1967), by Lilliefors (1967), and by Stephens (1974). The percentage points for the Kuiper statistics ($V$) were given by Loueter and Koerts (1970).

In section 2, we discuss the procedure of computing percentage points for approximate EDF statistics for testing normality and homogeneity of variance of samples. In section 3 we tabulate the percentage points for approximate EDF tests, obtained from Monte Carlo simulations. Application to goodness-of-fit tests is discussed in section 4. The powers of the approximate EDF goodness-of-fit tests are computed in section 5.

5.2 Procedure of computing percentage points for approximate EDF statistics

Suppose $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively, where $\mu_1$, $\mu_2$ and $\sigma^2$ are unknown. By replacing the unknown parameters $\mu_1$, $\mu_2$ and $\sigma^2$ by their maximum likelihood estimates (MLE’s), $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\sigma}^2$, we can obtain,

$$Z_i = \Phi \left( \frac{X_i - \hat{\mu}_1}{\hat{\sigma}} \right), \quad i = 1, \ldots, n,$$

$$Z_{n+j} = \Phi \left( \frac{Y_j - \hat{\mu}_2}{\hat{\sigma}} \right), \quad j = 1, \ldots, m,$$  

(5.2.1)

where $\Phi(X)$ denotes the cumulative density function of a standard normal distribution evaluated at $X$. These random variables are no more independent and identically distributed with uniform distribution in the interval (0, 1). Still we use the formulas in chapter 1 with these new variables, to calculate the EDF statistics. But the distributions of EDF statistics are very different from those when the parameters are completely known.
When the true distributions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ respectively, the joint distribution of $Z_1, \ldots, Z_{m+n}$ does not depend on $\mu_1$, $\mu_2$ and $\sigma^2$ (since $\mu_1$, $\mu_2$ and $\sigma^2$ are location and scale parameters). Therefore if we estimate $\mu_1$, $\mu_2$ and $\sigma^2$ by their MLE’s, then the distribution of EDF statistics will not depend on the true values of the unknown parameters. Thus the percentage points for EDF tests depend only on the sample sizes, $n$ and $m$. Hence, following the steps below, we can compute the percentage points for the EDF tests.

a) Simulate $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from $N(0, 1)$.

b) Calculate the $Z$’s using (5.2.1).

c) Calculate EDF statistics using the formulas in chapter 1.

d) Repeat the above three steps for large number of times (say 10,000).

e) Sort the EDF statistics in ascending order.

f) Take the desired percentile points.

This can be generalized to $k$ samples easily.

5.3 Percentage points for approximate EDF statistics

The percentage points for approximate EDF statistics, Kolmogorov-Smirnov ($D$), Kuiper ($V$), Cramer-von Mises ($W^2$), and Anderson-Darling ($A^2$) are computed using Monte Carlo simulations. The unknown parameters are estimated by maximum likelihood estimators (MLE). The critical values are obtained by computing the percentile points from 10,000
iterations. Table 5.1 provides percentage points for approximate EDF statistics for testing normality and equivariance of two samples of equal sizes. The percentage points are tabulated for the sample sizes 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. We can use the interpolation to obtain the percentage points for the sample sizes which are not tabulated but within the range 5 to 50.

We observed the percentage points were heavily dependent on the sample size and vary a lot when there were unequal sample sizes. We also observed the percentage points were depended on the number of samples whose normality and homogeneity of variance have to be checked. Because of these facts, it is impractical to provide tables of percentage points which can cover all the possibilities. Therefore we provide a Matlab program for computing the percentage points for approximate EDF tests when there are two samples of sizes $n$ and $m$. One can easily modify the program for any number of samples.
Table 5.1: Percentage points for testing normality and homogeneity of two samples when the parameters are unknown.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>EDF statistic</th>
<th>0.15</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=m=5</td>
<td>$D$</td>
<td>0.3165</td>
<td>0.3248</td>
<td>0.3375</td>
<td>0.3482</td>
<td>0.3581</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.5863</td>
<td>0.5985</td>
<td>0.6191</td>
<td>0.6393</td>
<td>0.6643</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>0.2684</td>
<td>0.2840</td>
<td>0.3073</td>
<td>0.3262</td>
<td>0.3504</td>
</tr>
<tr>
<td></td>
<td>$A^2$</td>
<td>1.4948</td>
<td>1.5598</td>
<td>1.6500</td>
<td>1.7279</td>
<td>1.8234</td>
</tr>
<tr>
<td>n=m=10</td>
<td>$D$</td>
<td>0.3380</td>
<td>0.3441</td>
<td>0.3530</td>
<td>0.3593</td>
<td>0.3675</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.6388</td>
<td>0.6481</td>
<td>0.6620</td>
<td>0.6740</td>
<td>0.6874</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>0.7219</td>
<td>0.7444</td>
<td>0.7778</td>
<td>0.8036</td>
<td>0.8360</td>
</tr>
<tr>
<td></td>
<td>$A^2$</td>
<td>3.8305</td>
<td>3.9233</td>
<td>4.0574</td>
<td>4.1600</td>
<td>4.2954</td>
</tr>
<tr>
<td>n=m=15</td>
<td>$D$</td>
<td>0.3508</td>
<td>0.3555</td>
<td>0.3627</td>
<td>0.3686</td>
<td>0.3753</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.6726</td>
<td>0.6799</td>
<td>0.6900</td>
<td>0.6993</td>
<td>0.7106</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>1.2515</td>
<td>1.2772</td>
<td>1.3163</td>
<td>1.3481</td>
<td>1.3869</td>
</tr>
<tr>
<td>n=m=20</td>
<td>$D$</td>
<td>0.3608</td>
<td>0.3651</td>
<td>0.3711</td>
<td>0.3758</td>
<td>0.3823</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.6970</td>
<td>0.7039</td>
<td>0.7133</td>
<td>0.7208</td>
<td>0.7286</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>1.8243</td>
<td>1.8547</td>
<td>1.8964</td>
<td>1.9314</td>
<td>1.9692</td>
</tr>
<tr>
<td>n=m=25</td>
<td>$D$</td>
<td>0.3690</td>
<td>0.3728</td>
<td>0.3781</td>
<td>0.3830</td>
<td>0.3879</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.7169</td>
<td>0.7222</td>
<td>0.7304</td>
<td>0.7376</td>
<td>0.7451</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>2.4285</td>
<td>2.4591</td>
<td>2.5055</td>
<td>2.5393</td>
<td>2.5804</td>
</tr>
</tbody>
</table>
Table 5.1: Percentage points for testing normality and homogeneity of two samples when the parameters are unknown. (continued)
<table>
<thead>
<tr>
<th>Sample size</th>
<th>EDF statistic</th>
<th>Significance level $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.15</td>
</tr>
<tr>
<td>$n=m=30$</td>
<td>$D$</td>
<td>0.3759</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.7320</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>3.0500</td>
</tr>
<tr>
<td></td>
<td>$A^2$</td>
<td>15.2738</td>
</tr>
<tr>
<td>$n=m=35$</td>
<td>$D$</td>
<td>0.3811</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.7445</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>3.6867</td>
</tr>
<tr>
<td></td>
<td>$A^2$</td>
<td>18.3556</td>
</tr>
<tr>
<td>$n=m=40$</td>
<td>$D$</td>
<td>0.3859</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.7553</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>4.3375</td>
</tr>
<tr>
<td>$n=m=45$</td>
<td>$D$</td>
<td>0.3899</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.7642</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>5.0041</td>
</tr>
<tr>
<td></td>
<td>$A^2$</td>
<td>24.6899</td>
</tr>
<tr>
<td>$n=m=50$</td>
<td>$D$</td>
<td>0.3936</td>
</tr>
<tr>
<td></td>
<td>$K$</td>
<td>0.7731</td>
</tr>
<tr>
<td></td>
<td>$W^2$</td>
<td>5.6739</td>
</tr>
</tbody>
</table>
5.4 Application to goodness-of-fit tests

Let $X_i, i = 1, \ldots, n, n \geq 2, Y_j, j = 1, \ldots, m, m \geq 2$ be two independent random samples from distributions $F$ and $G$, respectively. Consider testing the composite hypotheses

$$H_0 : F \text{ is a } N(\xi, \sigma^2) \text{ and } G \text{ is a } N(\eta, \sigma^2) \text{ versus } H_a : \text{ not } H_0$$

where $\xi, \eta$ and $\sigma^2$ are unknown.

The approximate EDF goodness-of-fit tests can be carried out using the following steps.

a) Calculate $W_i$ from

$$W_i = (X_i - Z_1) / \sqrt{Z_3} \quad i = 1, \ldots, n,$$

$$W_{n+j} = (Y_j - Z_2) / \sqrt{Z_3} \quad j = 1, \ldots, m,$$

where $Z_1 = \sum_{i=1}^{n} X_i / n, \quad Z_2 = \sum_{j=1}^{m} Y_j / m,$

$$Z_3 = \left( \sum_{i=1}^{n} (X_i - Z_1)^2 + \sum_{j=1}^{m} (Y_j - Z_2)^2 \right) / (n + m - 2),$$

are MLE’s of $\xi, \eta,$ and $\sigma^2$.

b) Calculate $U_i = \Phi(W_i), \quad i = 1, \ldots, (m+n)$, where $\Phi(X)$ denotes the cumulative density function of a standard normal distribution evaluated at $X$.

c) Calculate the EDF statistics using the standard formulas.

d) If the value of the statistic used exceeds the appropriate percentage point at level $\alpha$, $H_0$ is rejected with significance level $\alpha$.

This can be easily generalized to $k$ samples.
5.5 Power study

In this section we estimate the power of the approximate EDF goodness-of-fit tests for testing the normality and the homogeneity of variance of two samples. Power computation has been carried out for the number of symmetric and asymmetric alternatives. The powers of the tests for the following alternatives, exponential distribution, lognormal distribution, and Student’s $t$ distribution are tabulated. We use the notation, $\mu_1, \mu_2,$ and $\sigma^2_1, \sigma^2_2, \ldots$ to represent means and variances of the distributions from which the samples are obtained respectively and $\sigma^2$ as the common variance. $\nu$ represents the degrees of freedom of the distribution. We generate two samples, with $\mu_1 = 1, \mu_2 = 5$ and $\sigma^2 = 5$, from the exponential and lognormal distributions and used as alternatives. The location shifted Student’s $t$ distributions with $(\mu_1 = 1, \mu_2 = 5, \nu = 1)$ and $(\mu_1 = 1, \mu_2 = 5, \nu = 5)$ are also the alternatives considered in the power study. Two independent samples from normal distribution with unequal variances $(\sigma^2_1 = 1, \sigma^2_2 = 100)$ are also used as an alternative. This alternative is indicated as $\text{Normal}^*$ in the table. The type I error of the tests are also calculated. The statistics to be used in the goodness-of-fit tests are the Kolmogorov-Smirnov statistic ($D$), the Kuiper statistic ($V$), the Cramer-von Misses statistic and the Anderson-Darling statistic ($A^2$). Ten thousand Monte Carlo samples of size $n$ and $m$ ($n = m = 10, n = m = 20, n = m = 50$) were drawn from each alternative and we counted the number of times for which the statistics quoted were declared significant at levels $\alpha = 0.1$ and $\alpha = 0.05$. The values corresponding to $\alpha = 0.05$ are always tabulated in the second line of the tables. The results are presented in Tables 5.1 from which the following observations are made about the powers of tests:
(i) The approximate EDF statistics $D$, $V$, $W^2$, and $A^2$, maintain very well the nominal level $\alpha = 0.1$ and $\alpha = 0.05$.

(ii) Kolmogorov-Smirnov statistic has more power compare to the other EDF statistics used in skewed distributions.

(iii) Cramer-von Misses and Anderson-Darling statistics are recommended for symmetric distributions.

(iv) Approximate EDF tests perform very poorly when the samples are from normal distributions with different variances.
Table 5.2: Number of rejections out of 10000 iterations

<table>
<thead>
<tr>
<th></th>
<th>$n = m = 10$</th>
<th></th>
<th></th>
<th>$n = m = 20$</th>
<th></th>
<th></th>
<th>$n = m = 50$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D$</td>
<td>$V$</td>
<td>$W^2$</td>
<td>$A^2$</td>
<td>$D$</td>
<td>$V$</td>
<td>$W^2$</td>
<td>$A^2$</td>
<td>$D$</td>
<td>$V$</td>
</tr>
<tr>
<td>Normal</td>
<td>971</td>
<td>998</td>
<td>1002</td>
<td>997</td>
<td>1060</td>
<td>1024</td>
<td>972</td>
<td>960</td>
<td>983</td>
<td>993</td>
</tr>
<tr>
<td></td>
<td>474</td>
<td>494</td>
<td>481</td>
<td>495</td>
<td>499</td>
<td>496</td>
<td>486</td>
<td>492</td>
<td>499</td>
<td>493</td>
</tr>
<tr>
<td>Exponential</td>
<td>6060</td>
<td>4023</td>
<td>3521</td>
<td>3394</td>
<td>8779</td>
<td>5500</td>
<td>4639</td>
<td>4464</td>
<td>9959</td>
<td>7120</td>
</tr>
<tr>
<td></td>
<td>5235</td>
<td>3147</td>
<td>2781</td>
<td>2703</td>
<td>8443</td>
<td>4622</td>
<td>3974</td>
<td>3827</td>
<td>9946</td>
<td>6411</td>
</tr>
<tr>
<td>Lognormal</td>
<td>2239</td>
<td>1809</td>
<td>1971</td>
<td>1859</td>
<td>2853</td>
<td>2112</td>
<td>2640</td>
<td>2507</td>
<td>4228</td>
<td>2585</td>
</tr>
<tr>
<td></td>
<td>1974</td>
<td>1599</td>
<td>1731</td>
<td>1645</td>
<td>2615</td>
<td>1925</td>
<td>2412</td>
<td>2292</td>
<td>3946</td>
<td>2395</td>
</tr>
<tr>
<td>$t_{(1)}$</td>
<td>5859</td>
<td>5659</td>
<td>5601</td>
<td>5293</td>
<td>7248</td>
<td>6986</td>
<td>7363</td>
<td>6952</td>
<td>8377</td>
<td>8198</td>
</tr>
<tr>
<td></td>
<td>5280</td>
<td>5180</td>
<td>5031</td>
<td>4837</td>
<td>6809</td>
<td>6606</td>
<td>7015</td>
<td>6637</td>
<td>8118</td>
<td>7958</td>
</tr>
<tr>
<td>$t_{(5)}$</td>
<td>3200</td>
<td>3258</td>
<td>3670</td>
<td>3513</td>
<td>3875</td>
<td>3860</td>
<td>4947</td>
<td>4714</td>
<td>4465</td>
<td>4319</td>
</tr>
<tr>
<td></td>
<td>2511</td>
<td>2672</td>
<td>3069</td>
<td>2973</td>
<td>3270</td>
<td>3356</td>
<td>4409</td>
<td>4211</td>
<td>3904</td>
<td>3891</td>
</tr>
<tr>
<td>$Normal^*$</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The values corresponding to $\alpha = 0.05$ are tabulated in the second line.
REFERENCES


Appendix

CALCULATING THE PERCENTAGE POINTS OF APPROXIMATE EDF STATISTICS

Matlab program

Comment: Sizes of the samples

n=10; m=10;

x = []; y = []; D = []; V = []; WSQ = []; ADSQ = [];

for j = 1 : 10000,

Comment: Generating random numbers

x = randn(n, 1);

y = randn(m, 1);

Comment: Computing MLE’s

z1 = mean(x);

z2 = mean(y);

z3 = sum((x - z1).^2) + sum((y - z2).^2)/(m + n - 2);

x = (x - z1)/sqrt(z3);

y = (y - z2)/sqrt(z3);

nor = [];

nor = [x; y];

Comment: Using probability integral transformation

uni = []; orderuni = [];

uni = normcdf(nor, 0, 1);

orderuni = sort(uni);

Comment: Calculating Kolmogorov-Smirnov statistic

D1 = []; D2 = [];
for $i = 1 : (n + m)$,
$D1(i, 1) = i/(n + m) - orderuni(i, 1)$;
$D2(i, 1) = orderuni(i, 1) - (i - 1)/(n + m)$
end
$D(j, 1) = \max(\max(D1), \max(D2))$;
Comment: Calculating Kuiper statistic
$V(j, 1) = (\max(D1) + \max(D2))$;
Comment: Calculating Cramer-von misses statistic
$CWM = []$;
for $i = 1 : (n + m)$,
$CWM(i, 1) = orderuni(i, 1) - (2*i - 1)/(2*(n + m))$;
end
$WSQ(j, 1) = sum(CWM.^2) + 1/(12*(n + m))$;
Comment: Calculating Anderson-Darling statistic
$AD1 = []$;
for $i = 1 : (n + m)$,
if $orderuni(i, 1) == 0$
$orderuni(i, 1) = 0.0001$;
end
if $orderuni((n + m + 1 - i), 1) == 1$
$orderuni((n + m + 1 - i), 1) = 0.9999$;
end
$AD1(i, 1) = (2*i - 1) * (\log(orderuni(i, 1)) + \log(1 - orderuni((n + m + 1 - i), 1)))$;
end
$ADSQ(j, 1) = -(n + m) - (1/(n + m)) * sum(AD1)$
end
Comment: Sorting the observed values of the EDF statistics
\[ DS = []; VS = []; WSQS = []; ADSQS = []; \]

\[ DS = \text{sort}(D); \]
\[ VS = \text{sort}(V); \]
\[ WSQS = \text{sort}(WSQ); \]
\[ ADSQS = \text{sort}(ADSQ); \]

Comment: Finding the percentage points of the EDF statistics

\[ DS(9500, 1) \]
\[ DS(9900, 1) \]
\[ VS(9500, 1) \]
\[ VS(9900, 1) \]
\[ WSQS(9500, 1) \]
\[ WSQS(9900, 1) \]
\[ ADSQS(9500, 1) \]
\[ ADSQS(9900, 1) \]