Interpoint Distance Methods for the Analysis of High Dimensional Data

DISSEETATION

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Graduate School of The Ohio State University

By

John Lawrence, B. A., M. S., M. A. S.

*****
The Ohio State University
1996

Dissertation Committee:

Professor Dennis K. Pearl, Advisor
Professor Robert Bartoszyński
Professor Haikady Nagaraja

Approved By

Advisor
Department of Statistics
Acknowledgements

Thanks to all my friends in the Department of Statistics at The Ohio State University.
VITA

1988 ............................................................... B. A. Ohio Wesleyan University,
                                          Delaware, Ohio

1992 ............................................................... M. S. The Ohio State University
                                          Columbus, Ohio

1994 ............................................................... M. A. S. The Ohio State University
                                          Columbus, Ohio

Field of Study

Major Field: Statistics

          The comparison of high dimensional distributions
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ................................................................. ii

VITA ....................................................................................... iii

TABLE OF CONTENTS ............................................................... iv

LIST OF TABLES ....................................................................... vi

LIST OF FIGURES ...................................................................... viii

CHAPTER I. Introduction .......................................................... 1

CHAPTER II. The Triangle Goodness-of-fit Test ....................... 3

2.1 Review of the Literature on Goodness-of-fit Tests .......... 3

2.2 Asymptotic Distribution of the Statistic ......................... 15

2.3 Examples ........................................................................ 48

CHAPTER III. Two-sample Tests Based on Interpoint Distances ... 77

3.1 A Class of Consistent Conditional Test Statistics .......... 80

3.2 A Class of Unconditional Statistics ................................. 83

3.3 Monte Carlo Study .......................................................... 90

3.4 Asymptotic Power and Efficiency of the Unconditional Tests ................................................................. 96
CHAPTER IV. Goodness-of-fit Tests and Two-sample Tests for Very High Dimensional Data ........................................ 101

4.1 Motivation and Discussion .................................................. 101

4.2 Examples ............................................................................. 108

CHAPTER V. Conclusion ............................................................ 125

REFERENCES ............................................................................. 130
# LIST OF TABLES

Table 2.1  Values of parameters for various null distributions ..........  18  
Table 2.2  Covariance parameters under various alternatives ..........  21  
Table 2.3  Random variates from the Beta distribution ..................  23  
Table 2.4  Covariance parameters under null distributions ..............  41  
Table 2.5  Covariance parameters under null distributions ..............  42  
Table 2.6  Covariance parameters under null distributions ..............  46  
Table 2.7  Rejection region ..............................................  50  
Table 2.8  Power of the test for uniformity ................................  56  
Table 2.9  Critical values ..................................................  60  
Table 2.10 Power of the test for uniformity ................................  63  
Table 2.11 Covariance parameters under the alternative .................  64  
Table 2.12 Power of the test for normality ...............................  69  
Table 2.13 Power of the test for normality ...............................  71  
Table 3.1  Power of the test for normal populations ......................  94  
Table 3.2  Power of the test for uniform populations .....................  95  
Table 4.1  Summary statistics .............................................. 120  
Table 4.2  Conditional p-values for independence test of prime gaps ... 124
LIST OF FIGURES

Figure 2.1 Partition of the unit cube ........................................... 17
Figure 2.2 Scatterplot of parameter values ..................................... 19
Figure 2.3 Scatterplot of parameter values ..................................... 22
Figure 2.4 Graphs of power ......................................................... 25
Figure 2.5 Graphs of power ......................................................... 26
Figure 2.6 Graphs of the null distribution of the ump test ................. 33
Figure 2.7 Graphs of the null distribution of the Triangle test ............. 34
Figure 2.8 Graphs of the alternative distribution of the Triangle test ... 35
Figure 2.9 Scatterplot of parameter values ..................................... 47
Figure 2.10 Acceptance region ..................................................... 49
Figure 2.11 Graphs of power ......................................................... 61
Figure 2.12 Graphs of power ......................................................... 66
Figure 2.13 Depth of points along the x-axis for an angularly symmetric
distribution ................................................................. 76
Figure 3.1 Geometric interpretation of Q .......................................... 82
Figure 3.2 Graphs of power ......................................................... 99
Figure 4.1 Sample moments of T .................................................... 115
Figure 4.2 Sample moments of T under the null hypothesis ............... 117
Figure 4.3 Data for one week ....................................................... 118
Figure 4.4 Graphs of observed interpoint distances ......................... 119
Figure 4.5 Histogram of standardized gaps .................................... 123
Chapter I

Introduction

Tests of distributional assumptions are at the heart of nearly all statistical procedures. Many of the classical statistical methods are ineffective for high-dimensional data because of two main problems: i) a well-ordering for Euclidean n-space has not been found for n>1 and ii) estimating multivariate densities or distribution functions, even when a parametric model is assumed, is notoriously difficult. Methods based on interpoint distances dispose of both of these problems because interpoint distances are always one-dimensional regardless of the dimension of the original data.

Chapter II reviews some of the classical goodness-of-fit tests and a new goodness-of-fit procedures based on interpoint distances. The versatility and power of the proposed test is demonstrated through a Monte Carlo study and numerous examples.

Chapter III reviews some of the classical two-sample tests and introduce two classes of two-sample tests based on interpoint distances. A theorem provides the asymptotic distribution of one of the tests and results concerning the consistency of the tests are proven. The traditional and
proposed two-sample tests are compared through a Monte Carlo study and several examples are given.

In Chapter IV, a goodness-of-fit test and a two-sample test for very high dimensional data is proposed. These tests make use of the observation that, under mild conditions on the underlying distributions, the interpoint distances are themselves approximately normal.

The Monte Carlo studies and examples throughout Chapters II-IV are summarized and an outline of future plans for research are given in Chapter V.
Chapter II

The Triangle Goodness-of-fit Test

Introduction

For each pair of points in a random sample, consider a random triangle formed with vertices at these two points and a hypothetical point from the hypothesized distribution. Under the null hypothesis, all three points come from the same distribution, so the leg joining the two sample points is equally likely to be the shortest, longest, or the middle leg in the triangle. By averaging over all pairs of sample points, a goodness-of-fit test can be performed by measuring if the leg joining two sample points is shorter or longer on average than would be expected under the null hypothesis. The motivation for this statistic is provided by the theorem in Maa, Pearl, and Bartoszyński (1996).

2.1 Review of the Literature on Goodness-of-fit Tests

There is an often told joke about a policeman who sees a man crawling around on his knees under a streetlight. The policeman asks him what he's
doing and the man replies "I'm looking for my wallet". The policeman asks him if he lost it here and the man replies, "No, I lost it in the alley, but the light is better here". This is similar to what sometimes happens when data from an experiment is analyzed. Goodness-of-fit testing is an attempt to find out if the wallet was lost near the streetlight. To be more precise, probability models are often required to analyze data from an experiment and goodness-of-fit tests are performed to assess the adequacy of these models. For example, in the linear model $\tilde{Y} = X\beta + \tilde{e}$, if the errors, $\tilde{e}$, are multivariate normal, then it is possible to find the maximum likelihood estimator of $\beta$ and it can be shown that this estimator is itself multivariate normal. So, it is possible to find the covariance matrix for $\hat{\beta}$ along with the point estimate and obtain confidence ellipsoids for the parameter $\beta$. These confidence ellipsoids have the correct coverage probability provided the assumption of multivariate normality of $\tilde{e}$ is true. So, a test for multivariate normality should be performed on the estimated error vectors. Indeed, whenever a maximum likelihood estimator is found, the likelihood first needs to be constructed and that requires a probability model for the data. The importance of goodness-of-fit is also illustrated by the computationally intensive simulation-based techniques that have become popular in the last 20 years such as the bootstrap and the Markov chain Monte Carlo methods. These techniques require a pseudo random number generator that produces a sequence of variates which behave like a sequence of independent uniform random variables.

Sometimes, it is known that certain types of analyses are robust to specific types of departures from the model assumptions. For example, least squares estimates may not be maximum likelihood estimates, but will still be
UMVUE if the errors are assumed to be symmetric about the regression line in the simple linear regression model. However, scientists are usually interested in making the most efficient use of their data and if the a priori assumptions are not satisfied for a particular type of analysis, then the model should be changed or the data should be transformed so that the assumptions hold on the transformed data.

The criteria for judging a test differ for each problem. Sometimes the sample is known to come from a particular class $\Omega$ and the null hypothesis is that the sample comes from a subset of this class, $\Omega_0$. Other times, there is a class of alternatives we want to protect against and other alternatives that we don't really care about. To put it succinctly, "all models are wrong, but some models are useful" [George Box]. Hence, it may not do any harm to assume the null hypothesis is true if the true distribution is in some class of alternatives. One instance of this is computer random number generation. We know that the numbers generated are not independent uniform random variables but what we care about is if it does any harm to use them as if they were. In the case where the class of alternatives that we want to protect ourselves against is identified, tests can often be found that are powerful against these particular alternatives (the likelihood ratio test, for example). However, it is often the case that very little is known about the distribution of the sample and any alternative is potentially harmful. In this case, a test which is robust against many types of alternatives, even though it may not have the highest power against a particular alternative, may be desired. Above all other considerations, the test with the highest power against the possible alternatives would be preferred (if no test is uniformly most
powerful, then some sort of weighted combination of the powers against the alternatives of interest should be used to pick the best test).

Another consideration of practical importance is that the test statistic should be easy to calculate; that is, a formula should exist that defines the statistic as a function of the sample and some finite number of parameters that can be quickly calculated or published in a table. If such a formula does not exist, a major hurdle will have to be crossed to convince people that the test is worthwhile. One final consideration is that statistics which can be computed quickly have some advantage. For example, after finding that the sample does not come from a particular distribution, one may wish to try to transform the data so that it does appear to come from a certain distribution [see Bozdogan and Ramirez (1986)]. The criteria for choosing the best transformation may involve finding the transformation that optimizes the test statistic, so it is necessary to calculate the statistic many times. It would not be practical to do this if the statistic takes a very long time to calculate. However, the rate at which numerical calculations can be performed seems to be always increasing, so this is not a long-term concern for statistics which can be computed in polynomial time.

**General Multivariate Goodness-of-fit Testing**

The most commonly used omnibus tests of fit for univariate distributions are Pearson’s chi-square test and the Kolmogorov-Smirnov test. Pearson’s chi-square test is directly applicable to multivariate data by specification of a partition of the support of the sample space $\mathbf{X} = \bigcup_{i=1}^{m} \mathbf{X}_i$. The
family of power-divergence statistics generalizes this statistic. The family, which is indexed as follows
\[ I^\lambda = \frac{1}{\lambda(\lambda + 1)} \sum_{i=1}^{m} O_i \left( \left( \frac{O_i}{E_i} \right)^\lambda - 1 \right); \lambda \in \mathbb{R} \], is discussed in Cressie and Read (1984) and it includes Pearson's statistic ($\lambda = 1$) and the log likelihood ratio statistic ($\lambda = 0$). The authors recommend a compromise by taking $\lambda = \frac{2}{3}$.

Tests based on measures of distances between the empirical distribution function and the hypothesized distribution function do not readily extend to distribution free tests in the multivariate case [Simpson (1953)]. It is possible to transform the data via the Lévy transformation to data which are uniform on the hypercube (under the null hypothesis) [Rosenblatt (1952)] and then use a test for uniformity. One such test is the straightforward multivariate extension of the Kolmogorov-Smirnov test for densities concentrated on the unit hypercube [Saunders and Laud (1980)]. If the null distribution has support on the unit hypercube, the test statistic has the same distribution as the one dimensional Kolmogorov test. The test is not consistent versus all alternatives but it is consistent against all alternatives with the same support as the null distribution. A major weakness of this statistic is that information about the distribution is lost when the data is transformed here. Also, the Kolmogorov-Smirnov test is not particularly sensitive to differences in the tails. In particular, if the difference between the hypothesized distribution and the true distribution is in the tails, this test will have very low power.

A test based on the difference between an empirical probability measure and the hypothesized probability measure is given in Foutz (1980).
The empirical probability measure is defined by first partitioning the sample space into statistically significant blocks [Anderson (1966)]. The measure of difference is defined to be the largest difference between the two probability functions over all Borel sets. However, the main result of the paper is that once the statistically significant blocks $B_1, B_2, ..., B_n$ are found, then the statistic, which at first glance appears to be extremely difficult to compute, is simply

$$\max_{i=1,...,n} \left| \frac{i}{n} - D_i \right|$$

where the $D_i$ are the ordered values of the hypothesized measures of the $B_i$.

Properties of tests based on nearest neighbors can be found in Schilling (1979), Bickel and Breiman (1983) and Schilling (1983). For a random sample $X_1, X_2, ..., X_n$ the nearest neighbor distance from $X_i$ is $R_j = \min_{i \neq j} \|X_i - X_j\|$. If $f(x)$ is the underlying density, then the univariate variables $U_j = \exp \left\{ -n \int_{|x-x_j|<R_j} f(x) \, dx \right\}$ have a distribution which is nearly independent of $f(x)$ and have a joint distribution that is approximately that of independent uniform random variables. Since the nearest neighbor distances are small for large samples, a goodness-of-fit test can be based on testing the variables

$$W_j = \exp \left\{ -nf_0(X_j) \int_{|x-x_j|<R_j} \, dx \right\} = \exp \left\{ -n \int_{|x-x_j|<R_j} f_0(x) \, dx \right\}$$

for deviations from uniformity. They are led to analyzing the sequence $Z_n(t) = \sqrt{n} \left( \hat{H}(t) - \mathbf{E} \hat{H}(t) \right)$ where $\hat{H}(t)$ is the empirical distribution function of the $W_i$. This sequence of processes converges to a zero mean Gaussian process whose covariance depends on the hypothesized density and the true density. Tests of the deviation of the $W_i$ from uniformity can be computed from this sequence.
The biggest weakness of this statistic is that most of the information from the sample is thrown away, so it will have low power.

Tests of Goodness-of-fit for Uniformity

In some sense, testing for uniformity is the most important testing situation in one dimension since a test of $X \sim F$ can be regarded as a test of $F^{-1}(X) \sim \text{Uniform}(0,1)$. However, tests applied to the transformed data may have less power (this depends on the test and the alternative). In this section, literature on testing uniformity is reviewed along with literature on tests which are not designed to test the marginal uniformity of the data, but rather the independence. Both of these aspects are important in random number generator testing.

A test which is designed for alternatives which are important in experimental physics is presented in Swanepoel and de Beer (1990). The class of statistics presented is indexed by $m$ where $1 < m < n$. The statistic is found by first computing the range of all subsamples of size $m$. The statistic is then defined as the $c^{th}$ ordered range where $c$ is the greatest integer less than or equal to $rac{N}{(m-1)! \left( \frac{m-1}{m} \right)^{m} (\text{ml})^{2n^{-1}_{m-1}}}$. The statistic is shown to perform better than the scan statistic [Ajne 1968, Cressie 1977] and a test based on logarithms of high order spacings [Cressie 1978]. The scan statistic is defined as the maximal number of points that can be covered by some semicircle.

A general class of statistics based on spacings is discussed in Pyke (1965). Spacings are defined as the distances between successive order statistics of the sample. Distance based tests (including nearest neighbor tests) are functions
of spacings in one dimension. In my opinion, tests based on spacings are attractive in one dimension since no information is lost by using spacings because the original data can be retrieved if one knows the spacings. Unfortunately, it is difficult to extend general tests based on spacings to multivariate data since there must be some way to order the data and the additive distance property \( \left\| \mathbf{X}_{(2)} - \mathbf{X}_{(1)} \right\| + \left\| \mathbf{X}_{(3)} - \mathbf{X}_{(2)} \right\| = \left\| \mathbf{X}_{(3)} - \mathbf{X}_{(1)} \right\| \) does not hold in higher dimensions.

There are several power studies of various tests for uniformity including those in Locke and Spurrier (1978), Quesenberry and Miller (1977) and Miller and Quesenberry (1979). In these papers Monte Carlo studies are performed to assess the power of many tests for uniformity versus wide classes of alternatives. Miller and Quesenberry recommend a test proposed by Neyman as a general test as a result of their studies. The particular test they recommend is found by taking \( k=4 \) in the general class of statistics

\[
P_k = \sum_{r=1}^{k} \left( \sum_{j=1}^{n} \pi_r(x_j) \right)^2
\]

where \( \pi_r(x_j) \) is the \( r \)th Legendre polynomial evaluated at the \( j \)th data point.

It is not difficult to find random number generators that produce variates which are nearly identical to uniform random variables marginally. In light of this, there is more interest in theoretical results and empirical tests for the independence of successive pseudo-random variates. If \( k \)-dimensional \((k>1)\) vectors are formed by taking \( k \) successive variates from a linear congruential random number generator, these points lie on equally spaced parallel hyperplanes of lower dimension [Marsaglia (1968)]. Thus, it is desirable that these hyperplanes are as close together as possible.
The many tests for independence of $X_1, X_2, \ldots, X_n$ include

a) a test which for a fixed $N$ uses $X_1$ to obtain an integer, $N_1$, chosen
uniformly among 1, 2, ..., $N$ then inductively uses $X_i$ to obtain an integer, $N_i$,
chosen uniformly among 1, 2, ..., $N_{i-1}$. The test statistic is $N_n$ [Savir 1983,
Marsaglia 1985].

b) the runs test which for a fixed integer $k$ is defined as the number of runs
(either up or down) of length $k$ where a run up (down) of length $k$ is defined
as an occurrence of $k$ successive observations $X_i, X_{i+1}, \ldots, X_{i+k-1}$ which are
increasing (decreasing) [Dagnupar (1988)].

c) the poker test which is defined by using the variates to find hands of 5
cards, say, with each card being between 1 and 13. The number of hands of
each poker type are computed and compared to the expected number
[Dagnupar (1988)].

d) the coupon collector's test which is defined as the number of draws
required until all $N$ tickets are drawn from a box with the tickets 1, 2, ..., $N$
where the draws are made with replacement [Dagnupar (1988), Knuth (1968)].

The common problem with all of these tests is both that it is possible that a
generator could fail one of these tests but still be adequate for particular uses
and that a generator could pass a test but work poorly for the problem at hand.
It is already known that the numbers generated are not independent uniform
random variables. What is important is whether it is harmful to act as if they are.
Tests of Goodness-of-fit for Normality

There have been many tests proposed for testing that data $X_1, X_2, \ldots, X_n$ arise from a one dimensional normal distribution. Many have extensions which have also been proposed as tests of the hypothesis that a multivariate sample $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n$ arises from a multivariate normal distribution. In principle, any test statistic, $S(X_1, X_2, \ldots, X_n)$ which rejects the hypothesis of univariate normality for small values of $S$ can be adapted to a test for multivariate normality via Roy's union-intersection principle [Roy (1953)]. Suppose $X$ is the matrix with $n$ rows whose columns are the $d$-dimensional vectors which are to be tested for multivariate normality, then an extension of the test statistic is $\min_{c \in \mathbb{R}^d} S(c'X)$. By definition of multivariate normality, the random variables which make up $c'X$ are projections to univariate normality if the columns of $X$ are multivariate normal. By minimizing over all $c$, the projection which makes the data most non-normal is chosen. The drawback of this method is that it is computationally infeasible even for small $d$.

However, for the case when $S$ is the Shapiro-Wilk statistic, there are two approximations for this minimum which lead to two tests of similar character [Malkovich and Afifi (1973), Fattorini (1986)]. Let $\bar{X}$ denote the sample mean of the data. Malkovich and Afifi observed that $S(c'X)$ has a lower bound if $c'\left(\bar{X}_i - \bar{X}\right) = \frac{n-1}{n}$ and $c'\left(\bar{X}_j - \bar{X}\right) = -\frac{1}{n}$ for $i \neq j$ and searched for the value of $c$ that minimized the sum of squares

$$
\left(c'\left(\bar{X}_i - \bar{X}\right) - \frac{n-1}{n}\right)^2 + \sum_{j \neq i} \left(c'\left(\bar{X}_j - \bar{X}\right) + \frac{1}{n}\right)^2.
$$

There are $n$ values of $c$ that minimize this expression, namely $c^{(i)} = \frac{1}{n} S_n^{-1}(\bar{X}_i - \bar{X})$ for $i=1,\ldots,n$ where $S_n$ is the sample covariance matrix. Malkovich and Afifi chose to use the statistic

12
\[ S\left(c^{(i)}X\right) \] where \( c^{(i)} \) minimizes the sum of squares expression above while Fattorini suggested \( \min_{i=1,\ldots,n} S\left(c^{(i)}X\right) \). Monte Carlo studies show that Fattorini's procedure is superior [Henze and Zirkler (1990)].

Tests based on multivariate measures of skewness and kurtosis [Mardia (1975)] have more popularity as omnibus tests than they deserve. Let \( \tilde{Y}_i = S_n^{-\frac{1}{2}}(\tilde{X}_i - \tilde{X}) \) denote the \( i \)th sample vector which has been standardized by the sample mean and covariance matrix (the \( i \)th residual), then Mardia's measure of skewness and kurtosis are \( \frac{1}{n^2} \sum_{j,k=1}^{n} (\tilde{Y}_j \tilde{Y}_k)^3 \) and \( \frac{1}{n} \sum_{j=1}^{n} (\tilde{Y}_j \tilde{Y}_i)^2 \). These tests are not consistent against some alternatives and cannot be recommended as general tests. However, they may be useful as tests when specific types of alternatives are of interest. Bozdogan and Ramirez (1986) propose a multivariate Box-Cox transformation and then use these tests to assess how well the transformations perform.

Henze and Zirkler (1990) propose a test statistic which is defined as a weighted integral of the squared modulus of the distance between the empirical characteristic function of the residuals and its pointwise limit \( e^{-\frac{1}{2}t^2} \). This class of statistics is consistent and certain choices of the weights yield tests which are competitive with the best in the Monte Carlo study. Csörgő (1986) presents a different approach of measuring the discrepancy between the empirical characteristic function and its expectation.

Quiroz and Dudley (1991) modify Moore and Stubblebine's (1981) chi-square statistic with elliptical cell boundaries because it has no power against purely angular departures from normality. Quiroz and Dudley's class of
statistics is defined by specifying real valued functions \( g_1, g_2, ..., g_k \) which are functions of the radial variable, \( r \), and real valued functions \( h_1, h_2, ..., h_k \) which are functions of the angular variable, \( u \). The test statistic is

\[
S^2 = n \left\| V^{-\frac{1}{2}} \left( \tilde{f} - E\tilde{f} \right) \right\|^2
\]

where \( \tilde{f} = \left\{ \frac{1}{n} \sum_{i=1}^{n} g_i(r_i)h_i(u_i) \right\} \) is the average of the functions \( f_i(x) = g_i(r) h_i(u) \) over the scaled residuals and \( E\tilde{f} \) and \( V \) are its mean vector and covariance matrix respectively. By taking the angular functions to be 1 and choosing the radial functions to be indicator variables of the form \( g_i(r) = I_{(r_{i0}, r_{ik})}(r) \) where \( r_{00} = 0 \) and \( r_{ik} = \infty \), this statistic is Moore and Stubblebine's chi-square statistic. This is a very broad class of tests. A moderate size Monte Carlo study for two members of the class is presented below which demonstrates that these two tests have reasonable power against alternatives with independent margins as well as alternatives with angular-radial correlation. The statistics in this class are difficult to compute because it is typically not trivial to find \( E\tilde{f} \) and \( V \).

Looney (1995) discusses the extensions of any test for univariate normality to multivariate data by applying the univariate test to the marginals of the principal components. The multivariate statistic is a real valued function of the vector of univariate statistics.

Comparative power studies of several tests appear in Bera and John (1983), Henze and Zirkler (1990) and in Romeu and Ozturk (1993).
2.2 Asymptotic Distribution of the Triangle Statistics

The validity of distributional assumptions that underlie many statistical procedures can be examined with a general multidimensional goodness-of-fit test, such as those discussed in the previous section. Such procedures include simulation-based techniques which rely on an algorithm's ability to produce uniform and independent random variables. In this section, a new test is discussed which is conceptually simple, has an appealing logic, has easily accessible asymptotic properties, is generalizable to a variety of important problems, and appears to have good power against a broad class of alternatives. For the problem of testing whether a \(d\)-dimensional random sample follows a hypothesized distribution, consider a triangle formed by two randomly selected data points and a variable from the hypothesized distribution. If the null hypothesis is true then the three sides of this random triangle have the same distribution. The proposal is to base a goodness-of-fit test on a statistic which estimates the chance that the side formed by the line segment joining the data points is the smallest, middle, or largest side of the triangle.

More precisely, let \(X_1, X_2, \ldots, X_n \overset{iid}{\sim} F\). To test \(H_0: F = F_0\), let \(Y \sim F_0\) and define the statistics

\[
U_i = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P[|Y - X_i| > |X_i - X_j|, |Y - X_i| > |X_i - X_j|].
\]

This statistic estimates the chance that the side joining the two data points is the smallest side of a triangle formed from two randomly selected data points and the hypothetical \(Y\). Similarly, define the statistics
\[ U_2 = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Pr[|Y - X_i| > |X_i - X_j| < |Y - X_j| \text{ or } |Y - X_i| > |X_i - X_j| > |Y - X_j|] \] and
\[ U_3 = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Pr[|Y - X_i| < |X_i - X_j| < |Y - X_j| < |X_i - X_j|]. \] It will be shown below in Theorem 2.1 that under \( H_0 \), the statistic \( U_1 \), for example, is asymptotically normal with mean \( \frac{1}{3} \) and variance \( \frac{\beta}{n} + \frac{2(n-2)}{n^2} \alpha \) where
\[ \alpha = \text{Cov}(|F(2X_3 - X_1) - F(2X_1 - X_3)|, |F(2X_3 - X_2) - F(2X_2 - X_3)|) \] and \[ \beta = \text{Var}(|F(2X_3 - X_1) - F(2X_1 - X_2)|). \]

Each of the three statistics \( U_1, U_2, \) or \( U_3 \) could be used individually as a goodness-of-fit test statistic. However, there are particular alternatives where one may exhibit very high power while another has very low power. Therefore, an omnibus test statistic which can be expected to have moderately high power against a wide variety of alternatives can be obtained by combining the individual statistics together. For the sake of simplicity, the discussion here will first focus on the properties of an individual statistic, \( U_1 \), and then discuss how to combine the statistics together and the properties of this combined statistic.

First, consider the univariate case where \( F \) is the distribution of \( \text{Unif}(0,1) \). To calculate \( \alpha \) and \( \beta \) we partition the unit square into twenty segments so that we can integrate piecewise (see Figure 2.1).
Figure 2.1 The function $|F(2X_2 - X_1) - F(2X_1 - X_2)|$ induces this partition of the square so that the function is simple on each region.

Over each segment, $|F(2X_2 - X_1) - F(2X_1 - X_2)|$ is a simple function, so we can integrate over each segment separately and find $\beta = \frac{1}{9}$. If we partition a cube this way in the $xy$ plane and the $xz$ plane we find $\alpha = \frac{1}{720}$.

Table 2.1 shows the values of $\alpha$ and $\beta$ for various univariate distributions. For those distributions where a closed form solution is not possible, 1,000,000 simulated values of $X_1$, $X_2$ and $X_3$ were obtained. The value in the table is then the method of moment estimators of $\alpha$ and $\beta$. In
those cases where the exact value is known, the simulated values of $\alpha$ agree
with the estimated values in the third decimal place, but the estimates of $\beta$
are not as close to the true values. This shows that larger sample sizes are
needed in estimating $\beta$ to obtain the same accuracy as in estimating $\alpha$.

**Table 2.1** Values of $\alpha$ and $\beta$ for various null distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact Value (Rounded)</td>
<td>Simulated Value</td>
</tr>
<tr>
<td>Unif</td>
<td>0.001389</td>
<td>0.001268</td>
</tr>
<tr>
<td>B(1,2)</td>
<td>0.002927</td>
<td>0.002930</td>
</tr>
<tr>
<td>B(1,5)</td>
<td>0.004883</td>
<td>0.004108</td>
</tr>
<tr>
<td>B(3,3)</td>
<td>0.003283</td>
<td>0.003444</td>
</tr>
<tr>
<td>B(5,3)</td>
<td>0.003696</td>
<td>0.003390</td>
</tr>
<tr>
<td>B(15,4)</td>
<td>0.004392</td>
<td>0.004056</td>
</tr>
<tr>
<td>B(1/2, 1/2)</td>
<td></td>
<td>0.000753</td>
</tr>
<tr>
<td>B(1/10, 1/10)</td>
<td></td>
<td>0.005994</td>
</tr>
<tr>
<td>B(1/100, 1/100)</td>
<td></td>
<td>0.056087</td>
</tr>
<tr>
<td>Exp</td>
<td>0.006349</td>
<td>0.005555</td>
</tr>
<tr>
<td>G(5,1)</td>
<td>0.004905</td>
<td>0.004326</td>
</tr>
<tr>
<td>$\chi^2_6$</td>
<td>0.005115</td>
<td>0.004406</td>
</tr>
<tr>
<td>$\chi^2_{20}$</td>
<td>0.004761</td>
<td>0.004231</td>
</tr>
<tr>
<td>Cauchy</td>
<td></td>
<td>0.012692</td>
</tr>
<tr>
<td>N(0,1)</td>
<td></td>
<td>0.004764</td>
</tr>
</tbody>
</table>
Where the exact values are known, $\alpha$ and $\beta$ are highly correlated ($r=-0.964$), as shown in Figure 2.2. Thus, when it is not possible to find exact values, it may be possible to use the estimate of $\beta$, which can be estimated more easily, to improve the estimate of $\alpha$.

![Figure 2.2 Scatterplot of $\beta$ versus $\alpha$.](image)

**Figure 2.2** Scatterplot of $\beta$ versus $\alpha$.

**Power (Uniform Null Hypothesis, Beta Alternative)**

An approximate $\alpha$-level test of $H_0: X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ Uniform rejects $H_0$ if

$$\left| \frac{U_1 - \frac{1}{3}}{\sqrt{\text{Var}_0(U_1)}} \right| > z_{\frac{\alpha}{2}},$$

where $\text{Var}_0(U_1)$ denotes the variance under the null hypothesis and $z_{\frac{\alpha}{2}}$ denotes the upper $\frac{\alpha}{2}$ quantile of a standard Normal random variable. Now consider the alternative hypothesis
\( H_1: \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \overset{iid}{\sim} F_1 \) and let \( F_0 \) denote the distribution of a Uniform random variable. Under this alternative, \( U_1 \) is asymptotically normal with mean 

\[
E_{\mathcal{P}}[U_1] = E\left[ P(\mathbf{Y} - \mathbf{X}_1 > |\mathbf{X}_1 - \mathbf{X}_2|, |\mathbf{Y} - \mathbf{X}_2| > |\mathbf{X}_1 - \mathbf{X}_2|) \right]
\]

\[
= E\left[ 1 - |F_0(2\mathbf{X}_2 - \mathbf{X}_1) - F_0(2\mathbf{X}_1 - \mathbf{X}_2)| \right] = 1 - E\left[ |F_0(2\mathbf{X}_2 - \mathbf{X}_1) - F_0(2\mathbf{X}_1 - \mathbf{X}_2)| \right]
\]

and variance \( \frac{\beta'}{n} + \frac{2(n-2)}{n} \alpha' \) where \( \alpha' = \text{Cov}(\|F_0(2\mathbf{X}_3 - \mathbf{X}_1) - F_0(2\mathbf{X}_1 - \mathbf{X}_3)\|^2, |F_0(2\mathbf{X}_3 - \mathbf{X}_2) - F_0(2\mathbf{X}_2 - \mathbf{X}_3)|) \) and \( \beta' = \text{Var}(|F_0(2\mathbf{X}_2 - \mathbf{X}_1) - F_0(2\mathbf{X}_1 - \mathbf{X}_2)|). \) Table 2.2 shows the values of \( E_{\mathcal{P}}[U_1], \alpha' \) and \( \beta' \) in the special case when the alternatives are members of the Beta family of distributions [decimal approximations to exactly computed parameters are in bold, others were found by simulation]. Various relationships between these three parameters for the Beta(\( \theta, \theta \)) distributions are illustrated in Figure 2.3.
Table 2.2 Covariance parameters under various alternative parameters
(decimal approximations to exact values in bold).

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$E_1{U_1}$</th>
<th>$\alpha'$</th>
<th>$\beta'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta($\frac{1}{10}, \frac{1}{10}$)</td>
<td>0.402237</td>
<td>0.006877</td>
<td>0.196620</td>
</tr>
<tr>
<td>Beta($\frac{1}{3}, \frac{4}{3}$)</td>
<td>0.356631</td>
<td>0.005019</td>
<td>0.164600</td>
</tr>
<tr>
<td>Beta($\frac{1}{3}, \frac{1}{3}$)</td>
<td>0.328214</td>
<td>0.001556</td>
<td>0.139369</td>
</tr>
<tr>
<td>Beta($\frac{1}{2}, \frac{1}{2}$)</td>
<td>0.317593</td>
<td>0.000428</td>
<td>0.123798</td>
</tr>
<tr>
<td>Beta(1,1)</td>
<td>0.333333</td>
<td>0.001389</td>
<td>0.111111</td>
</tr>
<tr>
<td>Beta(2,2)</td>
<td>0.392857</td>
<td>0.006877</td>
<td>0.108582</td>
</tr>
<tr>
<td>Beta(3,3)</td>
<td>0.444535</td>
<td>0.010615</td>
<td>0.105397</td>
</tr>
<tr>
<td>Beta(4,4)</td>
<td>0.486827</td>
<td>0.012735</td>
<td>0.100440</td>
</tr>
<tr>
<td>Beta(5,5)</td>
<td>0.521891</td>
<td>0.013802</td>
<td>0.094694</td>
</tr>
<tr>
<td>Beta(6,6)</td>
<td>0.551444</td>
<td>0.014206</td>
<td>0.088779</td>
</tr>
<tr>
<td>Beta(7,7)</td>
<td>0.576714</td>
<td>0.014197</td>
<td>0.083023</td>
</tr>
<tr>
<td>Beta(8,8)</td>
<td>0.598585</td>
<td>0.013936</td>
<td>0.077587</td>
</tr>
<tr>
<td>Beta(9,9)</td>
<td>0.617713</td>
<td>0.013526</td>
<td>0.072540</td>
</tr>
<tr>
<td>Beta(10,10)</td>
<td>0.634589</td>
<td>0.013035</td>
<td>0.067901</td>
</tr>
<tr>
<td>Beta(15,15)</td>
<td>0.696091</td>
<td>0.010546</td>
<td>0.050179</td>
</tr>
<tr>
<td>Beta(20,20)</td>
<td>0.735182</td>
<td>0.008352</td>
<td>0.039014</td>
</tr>
<tr>
<td>Beta(25,25)</td>
<td>0.762501</td>
<td>0.006876</td>
<td>0.031671</td>
</tr>
<tr>
<td>Beta(1,2)</td>
<td>0.416666</td>
<td>0.012847</td>
<td>0.11116</td>
</tr>
<tr>
<td>Beta(1,5)</td>
<td>0.643962</td>
<td>0.019116</td>
<td>0.070652</td>
</tr>
<tr>
<td>Beta(5,3)</td>
<td>0.511473</td>
<td>0.013511</td>
<td>0.093567</td>
</tr>
<tr>
<td>Beta(15,4)</td>
<td>0.718528</td>
<td>0.008095</td>
<td>0.038134</td>
</tr>
</tbody>
</table>
Figure 2.3 Scatterplots between the three parameters under the symmetric Beta alternative hypothesis with parameter $\theta$.

The value of $E_i\{U_i\}$ as $\theta \to 0$ should be close to 0.5 because the limiting distribution is the discrete uniform distribution on the points 0 and 1. However, the simulated value of $E_i\{U_i\}$ was found to be 0.717857 when $\theta=0.01$. Hence, there is evidence that the IMSL subroutine for generating Beta(0.01,0.01) random variables does not function properly. Table 2.3 shows...
the values of 100 consecutive variates generated by this subroutine. Clearly, the unusual number of occurrences of values close to 1 relative to the number of values close to 0 indicates that the subroutine is indeed not performing properly. The parameters for the discrete uniform distribution are \( E_i(U_i) = \frac{1}{2}, \alpha' = 0 \) and \( \beta' = \frac{1}{4} \).

Table 2.3 Random Variates from the Beta(0.01,0.01) distribution generated by IMSL.

<table>
<thead>
<tr>
<th>Variate 1</th>
<th>Variate 2</th>
<th>Variate 3</th>
<th>Variate 4</th>
<th>Variate 5</th>
<th>Variate 6</th>
<th>Variate 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2553855</td>
<td>0.9999999</td>
<td>0.9999615</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.0128741</td>
<td>0.9999999</td>
<td>0.9012460</td>
<td>0.9999999</td>
<td>0.9774647</td>
<td>0.023958</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9940792</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.0001453</td>
<td>0.0154066</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.0004906</td>
<td>0.9999999</td>
<td>0.0000020</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.1026579</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999997</td>
<td>0.9999999</td>
<td>0.9999884</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9661210</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999934</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.6617660</td>
</tr>
<tr>
<td>0.1138200</td>
<td>0.1694333</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.8320880</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
<tr>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
<td>0.9999999</td>
</tr>
</tbody>
</table>

Now, the approximate power of the test versus a specific alternative is

\[
P_1 \left[ \frac{U_1 - \frac{1}{3}}{\sqrt{\text{Var}_0(U_1)}} > z_\alpha \right] = P_1 \left[ U_1 < \frac{1}{3} - \frac{z_\alpha}{2} \sqrt{\text{Var}_0(U_1)} \right] + P_1 \left[ U_1 > \frac{1}{3} + \frac{z_\alpha}{2} \sqrt{\text{Var}_0(U_1)} \right]
\]
\[\begin{align*}
\text{Graphs of the power for different alternatives and sample sizes are shown in Figures 2.4 and 2.5.}
\end{align*}\]
Figure 2.4 Graphs of the power versus the sample size for various alternatives ($\alpha=0.05$).
Figure 2.5 Graphs of the power versus the logarithm of the alternative parameter for \( n=20 \) and \( n=200 \).

In order to judge the usefulness of the proposed test, we will compare it to a test designed specifically for this class of alternatives. The most powerful test against the alternative \( \text{Beta}(\theta, \theta) \) rejects when \( \sum_{i=1}^{n} \log x_i + \log(1-x_i) > k_0 \) if \( \theta > 1 \) and when \( \sum_{i=1}^{n} \log x_i + \log(1-x_i) < k_0 \) if \( \theta < 1 \). The statistic

\[ T_n = \sum_{i=1}^{n} \log X_i + \log(1-X_i) \]

is asymptotically normal. Under the null hypothesis, \( E(T_n) = -2n \) and \( \text{Var}(T_n) = n \left( 4 - \frac{\pi^2}{3} \right) \) since if \( X \sim \text{Uniform} \), we know that \( -\log(1-X) \xrightarrow{\text{dist}} -\log X \sim \text{Exp}(1) \), so \( E[\log X + \log(1-X)] = -2 \) and \( \text{Var}[\log X + \log(1-X)] = \text{Var}[\log X] + \text{Var}[\log(1-X)] + 2\text{Cov}[\log X, \log(1-X)] \)

\[ = 1 + 2 \left\{ E[\log X \log(1-X)] - (E[\log X])^2 \right\} = 2 + 2 \left( 2 - \frac{\pi^2}{6} - 1 \right) = 4 - \frac{\pi^2}{3} . \]
In particular, \( E[\log X \log(1-X)] = \int_0^1 \log x \log(1-x) \, dx = 1 + \int_0^1 \frac{(1-x) \log(1-x)}{x} \, dx \\
= 1 - \int_0^1 \frac{(1-x)}{x} \sum_{k=1}^{\infty} x^k \, dx = 1 - \int_0^1 \sum_{k=1}^{\infty} x^{k-1} \sum_{k=1}^{\infty} x^k \, dx = 1 - \sum_{k=2}^{\infty} \sum_{k=1}^{\infty} \frac{x^k}{k} \\
= -\sum_{k=1}^{\infty} \frac{x^k}{k+1} + \sum_{k=1}^{\infty} \frac{x^k}{k} \\
= 1 - \frac{1}{(k+1)^2} = \sum_{k=1}^{\infty} \frac{1}{k} - \frac{1}{k+1} - \frac{1}{(k+1)^2} = 2 - \frac{\pi^2}{6}.

To find the mean and variance of \( T_n \) under the alternative hypothesis, let \( X \sim \text{Beta}(\theta, \theta) \) and let \( f \) and \( F \) be the density and distribution of \( X \). The distribution of \( Y = -\log X \) is \( P[Y \leq y] = P[X \geq e^{-y}] = \int_y^{\infty} \text{Beta}(\theta, \theta; x) \, dx \). The density is \( e^{-y} \text{Beta}(\theta, \theta; e^{-y}) \). Now, for integer values of \( \theta \), we can use the binomial expansion to calculate the mean and variance of \( Y \). In this case,

\[
E[Y] = \sum_{i=0}^{\theta-1} \frac{(2\theta-1)!}{[\theta-1]!} \int_0^{\infty} y^i e^{-y} (1-e^{-y})^{\theta-1-i} (e^{-y})^{\theta-i} \, dy \\
= \frac{(2\theta-1)!}{[\theta-1]!} \sum_{i=0}^{\theta-1} \left( \frac{i}{\theta-1} \right) \int_0^{\infty} y^i (1-e^{-y})^{\theta-1-i} (e^{-y})^{\theta-i} \, dy \\
= \frac{(2\theta-1)!}{[\theta-1]!} \sum_{i=0}^{\theta-1} \left( \frac{i}{\theta-1} \right) \int_0^{\infty} y^i e^{-y(i+\theta)} \, dy \\
= \frac{(2\theta-1)!}{[\theta-1]!} \sum_{i=0}^{\theta-1} \left( \frac{i}{\theta-1} \right) \frac{\theta-1}{(i+\theta)^2} = \sum_{i=0}^{\theta-1} \frac{1}{i^2}.
\]

Similarly, \( E[Y^2] = \sum_{i=0}^{\infty} \frac{(2\theta-1)!}{[\theta-1]!} \int_0^{\infty} y^{2i} e^{-y} (1-e^{-y})^{\theta-1-i} (e^{-y})^{\theta-i} \, dy \\
= \frac{(2\theta-1)!}{[\theta-1]!} \sum_{i=0}^{\infty} \int_0^{\infty} y^{2i} e^{-y(i+\theta)} \, dy \\
= \frac{(2\theta-1)!}{[\theta-1]!} \sum_{i=0}^{\theta-1} \left( \frac{i}{\theta-1} \right) \frac{2(-1)^i}{(i+\theta)^3} = \sum_{i=0}^{\theta-1} \frac{1}{i^3} + \sum_{i=\theta}^{\infty} \frac{1}{i^3},
\]

so that \( \text{Var}[Y] = \sum_{i=\theta}^{\infty} \frac{1}{i^3} \). There is no simple formula for \( E[\log X \log(1-X)] \) here,

although \( E[\log X \log(1-X)] = \int_0^1 \left( \sum_{m=1}^{\infty} \frac{-(1-x)^m}{m} \right) \left( \sum_{n=1}^{\infty} \frac{x^n}{n} \right) \frac{(2\theta-1)!}{[\theta-1]!} x^{\theta-1}(1-x)^{\theta-1} \, dx \).
\[
(2\theta - 1)! \left[ \sum_{m=1}^{n} \sum_{n=1}^{\infty} \frac{1}{mn} x^{\theta+n-1} (1-x)^{\theta+n-1} \right] dx = (2\theta - 1)! \left[ \sum_{m=1}^{n} \sum_{n=1}^{\infty} \frac{(\theta+n)(\theta+m)!}{m(2\theta+m+n-1)!} \right]
\]

\[
= (2\theta - 1)! \sum_{n=1}^{\infty} \frac{(\theta-n)! (\theta+n-1)!}{n(2\theta+n-1)!} \sum_{k=n+\theta}^{\infty} \frac{1}{k!} = (2\theta - 1)! \sum_{n=1}^{\infty} \frac{(n+\theta-1)!}{n(n+2\theta-1)!} \sum_{k=n+\theta}^{\infty} \frac{1}{k!}.
\]

In principle, this expression can always be computed exactly (in fact a formula can be found for the mean and variance of the most powerful test statistic for testing against the alternative Beta(m,n) where m,n \( \in \mathbb{Z} \)). For example, if \( \theta=2 \), we find

\[
E[\log X \log(1-X)] = 6 \sum_{n=1}^{\infty} \frac{(n+1)!}{n(n+3)!} \sum_{k=n+2}^{\infty} \frac{1}{k!} = 6 \sum_{n=1}^{\infty} \frac{5+2n}{n(n+2)! (n+3)!} = \frac{37 - 3\pi^2}{18}.
\]

For \( \theta=3 \), we find

\[
E[\log X \log(1-X)] = 60 \sum_{n=1}^{\infty} \frac{(n+1)!}{n(n+3)!} \sum_{k=n+3}^{\infty} \frac{1}{k!} = 60 \sum_{n=1}^{\infty} \frac{47}{3600n} - \frac{1}{6(n+3)^2} - \frac{1}{18(n+3)} + \frac{1}{4(n+4)^2} + \frac{1}{16(n+4)} - \frac{1}{10(n+5)^2} - \frac{1}{50(n+5)}
\]

\[
= \frac{3739 - 300\pi^2}{1800}.
\]

More generally,

\[
\frac{1}{n(n+1)(n+\theta+1)\cdots(n+2\theta-1)} = \sum_{k=0}^{n-\theta} (-1)^k \frac{(\theta+1)!}{\theta-k-1)!k!n(\theta+k)}
\]

\[
= \sum_{k=0}^{\theta-1} (-1)^k \frac{(\theta+1)!}{\theta-k-1)!k!(\theta+k)n} - \sum_{k=0}^{\theta-1} (-1)^k \frac{(\theta+1)!}{\theta-k-1)!k!(\theta+k)n}
\]

and for \( 0 \leq j \leq \theta - 1 \), we have

\[
= \sum_{k=0}^{\theta-1} (-1)^k \frac{(\theta+1)!}{\theta-k-1)!k!(\theta+k)n} - \sum_{k=0}^{\theta-1} (-1)^k \frac{(\theta+1)!}{\theta-k-1)!k!(\theta+k)n}
\]

28
\[
\begin{align*}
\sum_{k=0}^{\infty} & \frac{(-1)^k}{k!} \left( \frac{1}{(\theta+j)n} - \frac{1}{(\theta+j)(n+\theta+j)} \right) \left( \frac{1}{(\theta+j)(n+\theta+j)} - \frac{1}{j!(\theta+j)(n+\theta+j)^2} \right) \\
& - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(n+\theta+j)} \\
& - \frac{1}{(\theta+j)n} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(n+\theta+j)} \\
& + \frac{1}{(n+\theta+j)} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(j-k)} \\
& - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(j-k)(n+\theta+k)} \\
\end{align*}
\]

Hence, \( E[\log X \log(1-X)] = \frac{(2\theta-1)!}{(\theta-1)!} \sum_{n=0}^{\infty} \frac{(n+\theta-1)!}{n(n+2\theta-1)!} \sum_{k=n+\theta}^{\infty} \frac{1}{k} \)

\[
\begin{align*}
& = \frac{(2\theta-1)!}{(\theta-1)!} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} \frac{1}{n(n+\theta+j)(n+\theta)(n+\theta+1) \cdots (n+2\theta-1)} \\
& = \frac{(2\theta-1)!}{(\theta-1)!} \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} \left\{ \frac{1}{(\theta+j)n} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(n+\theta+j)} - \frac{1}{j!(\theta-j-1)!(\theta+j)(n+\theta+j)^2} \right\} \\
& + \frac{1}{(n+\theta+j)} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(j-k)} \\
& - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(\theta-k-1)!(\theta+k)(j-k)(n+\theta+k)} \\
\end{align*}
\]
\[
\frac{(2\Theta - 1)!}{(\Theta - 1)!} \sum_{n=1}^{\infty} \frac{1}{n} \frac{1}{\sum_{j=0}^{\Theta + j}} \left( \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + k)} \right) - \sum_{j=0}^{\Theta - 1} \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)(n + \Theta + j)^2} \\
+ \sum_{j=0}^{\Theta - 1} \frac{1}{(n + \Theta + j)} \left\{ \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + k)(j - k)} - \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + j)(\Theta + k)} \right\} \\
- \sum_{j=0}^{\Theta - 1} \frac{(-1)^k}{(n + \Theta + j)(j - k)(n + \Theta + k)} \right\} \\
= \frac{(2\Theta - 1)!}{(\Theta - 1)!} \sum_{n=1}^{\infty} \frac{1}{n} \frac{1}{\sum_{j=0}^{\Theta + j}} \left( \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + k)} \right) - \sum_{j=0}^{\Theta - 1} \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)(n + \Theta + j)^2} \\
+ \sum_{j=0}^{\Theta - 1} \frac{1}{(n + \Theta + j)} \left\{ \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + k)(j - k)} - \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)^2} \right\} \\
- \sum_{j=0}^{\Theta - 1} \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)(n + \Theta + j)^2} \right\} \\
= \frac{(2\Theta - 1)!}{(\Theta - 1)!} \sum_{n=1}^{\infty} \frac{1}{n} \frac{1}{\sum_{j=0}^{\Theta + j}} \left( \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + k)} \right) - \sum_{j=0}^{\Theta - 1} \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)(n + \Theta + j)^2} \\
+ \sum_{j=0}^{\Theta - 1} \frac{1}{(n + \Theta + j)} \sum_{k=0}^{\Theta - 1} \frac{(-1)^k}{k!(\Theta - k - 1)!(\Theta + k)(j - k)} - \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)^2} - \sum_{j=0}^{\Theta - 1} \frac{(-1)^j}{j!(\Theta - j - 1)!(\Theta + j)(n + \Theta + j)^2} \right\} \\
\right. 
\]
\[
\frac{(2\theta - 1)!}{(\theta - 1)!} \left( \sum_{n=1}^{\theta - 1} \frac{(-1)^k}{n ! (\theta + n)!} \sum_{j=0}^{n - 1} \frac{(-1)^j}{j!(\theta - j - 1)!(\theta + j)!} \right) - \sum_{j=0}^{\theta - 1} \sum_{n=1}^{\theta - 1} \frac{(-1)^j}{j!(\theta - j - 1)!(\theta + j)!(n + \theta + j)^2}
\]

\[
+ \sum_{j=0}^{\theta - 2} \sum_{n=1}^{\theta - j - 1} \left\{ \frac{(-1)^k}{k!(\theta - k - 1)!} \sum_{j=0}^{n - 1} \frac{(-1)^j}{j!(\theta - j - 1)!(\theta + j)!} \right\} - \sum_{j=0}^{\theta - 1} \left\{ \frac{\pi^2}{6} - \sum_{n=1}^{\theta + j} \frac{(-1)^j}{n^2} \right\}
\]

\[
= \frac{(2\theta - 1)!}{(\theta - 1)!} \left( \sum_{n=1}^{\theta - 1} \frac{(-1)^k}{n ! (\theta + n)!} \sum_{j=0}^{n - 1} \frac{(-1)^j}{j!(\theta - j - 1)!(\theta + j)!} \right) - \sum_{j=0}^{\theta - 1} \sum_{n=1}^{\theta - 1} \frac{(-1)^j}{j!(\theta - j - 1)!(\theta + j)!(n + \theta + j)^2}
\]

\[
+ \sum_{j=0}^{\theta - 2} \sum_{n=1}^{\theta - j - 1} \left\{ \frac{(-1)^k}{k!(\theta - k - 1)!} \sum_{j=0}^{n - 1} \frac{(-1)^j}{j!(\theta - j - 1)!(\theta + j)!} \right\} - \sum_{j=0}^{\theta - 1} \left\{ \frac{\pi^2}{6} - \sum_{n=1}^{\theta + j} \frac{(-1)^j}{n^2} \right\}
\]

So, we have an exact expression for the mean and variance of \(T_n\), but it is more convenient to calculate them numerically. For \(\theta < 1\), we must use numerical integration to find \(E[T_n]\) and \(\text{Var}[T_n]\).

In general, the following formulas may be useful in calculating the moments of \(T_n\). Notice that \(Y\) has mean

\[
E[Y] = \int_0^\infty \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} e^{-y}(1 - e^{-y})^{\theta - 1}(e^{-y})^{\theta - 1} dy = \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} \sum_{i=0}^{\theta - 1} \int_0^{\theta - 1} e^{-y}(e^{-y})^{\theta - 1} dy
\]

\[
= \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} \sum_{i=0}^{\theta - 1} (-1)^i \int_0^{\theta - 1} e^{-y(i + \theta)} dy = \frac{\Gamma(2\theta)}{\Gamma(\theta)^2} \sum_{i=0}^{\theta - 1} \frac{(\theta - 1)^i}{(i + \theta)^2}.
\]

Now, if we let \(F(x)\) denote the distribution of \(X\), we find

\[
\frac{\Gamma(\theta)^2}{\Gamma(2\theta)} F(t) = \int_0^t (1 - x)^{\theta - 1} x^{\theta - 1} dx = \sum_{i=0}^{\theta - 1} \frac{(\theta - 1)^i}{i + \theta} t^{i+\theta}
\]

and

\[
\frac{\Gamma(\theta)^2}{\Gamma(2\theta)} \int_0^t F(t) dt = \sum_{i=0}^{\theta - 1} \frac{(\theta - 1)^i}{(i + \theta)^2} \frac{(-1)^i}{i + \theta} \quad \text{so that} \quad E[Y] = \int_0^1 F(t) dt.
\]

31
Also, \( E[Y^2] = \int_0^\infty y^2 \frac{\Gamma(2\theta)}{[\Gamma(\theta)]^2} e^{-y}(1-e^{-y})^{\theta-1}(e^{-y})^{\theta-1} \, dy \)

\[ = \frac{\Gamma(2\theta)}{[\Gamma(\theta)]^2} \sum_{i=0}^{\infty} \int_0^\infty y^2 e^{-y}(1-e^{-y})^i(e^{-y})^{\theta-1} \, dy \]

\[ = \frac{2\Gamma(2\theta)}{[\Gamma(\theta)]^2} \sum_{i=0}^{\infty} \left( \frac{\theta-1}{i} \right) (-1)^i = \frac{2}{\theta} \int_0^\infty \frac{F(t)}{t} \, dt \]

so \( \text{Var}[Y] = 2\int_0^\infty \frac{F(t)}{t} \, dt - \left( \int_0^\infty \frac{F(t)}{t} \, dt \right)^2. \)

If \( \theta < 1 \) the power of the test is approximately

\[
\Phi \left[ \frac{-2n - z_\alpha \sqrt{n \left( 4 \cdot \frac{\pi^2}{3} - E_\theta \{ T_n \} \right)}}{\sqrt{\text{Var}_\theta \{ T_n \}}} \right]
\]

and if \( \theta > 1 \) the power of the test is approximately

\[
1 - \Phi \left[ \frac{-2n + z_\alpha \sqrt{n \left( 4 \cdot \frac{\pi^2}{3} - E_\theta \{ T_n \} \right)}}{\sqrt{\text{Var}_\theta \{ T_n \}}} \right].
\]

Graphs of the power of the uniformly most powerful (UMP) test and the triangle test are shown in Figure 2.6.
Figure 2.6 Graphs of the power of the UMP test and the triangle test versus the sample size for various alternatives ($\alpha=0.05$)
For $\theta > 1$, it looks like the triangle test is as good as the UMP test. The reason that it looks like it beats the UMP test for small $n$ is that the normal approximation is not very accurate. Figure 2.7 shows histograms and normal probability plots to check the validity of the normal approximation under the null hypothesis for different sample sizes for the UMP test. Figure 2.8 shows the same plots for the triangle test.

**Figure 2.7** Normal probability plots for the UMP test statistic under the null hypothesis
Using the Edgeworth Expansion to improve the Normal Approximation

Since the distribution of the triangle statistic is skewed (see Figure 2.8), a transformation of the statistic may be more rapidly approximated by a normal distribution. The third moment will tell us what transformation will be most useful to make the skewness vanish. As in the calculation of the variance, define $V_n = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left| F(2X_j - X_i) - F(2X_i - X_j) \right|$ and let

$A_{ij} = \left| F(2X_j - X_i) - F(2X_i - X_j) \right| - E \left[ \left| F(2X_j - X_i) - F(2X_i - X_j) \right| \right]$. 

\[ \text{Figure 2.8 Normal probability plots of the triangle test statistic under the null hypothesis} \]
Then, \( E[(V_n - E[V_n])^3] = E\left[ (V_{n-1} - E[V_{n-1}]) + \sum_{i=1}^{n-1} A_{in} \right]^3 \) \\
= \( E[(V_{n-1} - E[V_{n-1}])^3] + E\left[ \sum_{i=1}^{n-1} A_{in} \right]^3 + 3E\left[ (V_{n-1} - E[V_{n-1}]) \left( \sum_{i=1}^{n-1} A_{in} \right)^2 \right] \) \\
+ 3E\left[ (V_{n-1} - E[V_{n-1}])^2 \left( \sum_{i=1}^{n-1} A_{in} \right) \right] \\
= \( E[(V_{n-1} - E[V_{n-1}])^3] + E\left[ \sum_{i=1}^{n-1} A_{in} \right]^3 + 3E\left[ \left( \sum_{i=1}^{n-1} A_{in} \right)^2 \right] \) \\
+ 3E\left[ \left( \sum_{i=1}^{n-1} A_{in} \right) \right] \\
= \( E[(V_{n-1} - E[V_{n-1}])^3] + (n-1)E[A_{12}^3] + 2\binom{n-1}{2}E[A_{12}A_{13}^2] + 3\binom{n-1}{3}E[A_{12}A_{13}A_{14}] \) \\
+ 3\left\{ \left[ \binom{n-1}{2}E[A_{12}A_{13}^2] + 2\binom{n-1}{2}E[A_{12}A_{13}A_{23}] + 12\binom{n-1}{3}E[A_{12}A_{14}A_{23}] \right] \right\} \\
+ 3\left\{ \left[ \binom{n-1}{2}E[A_{12}A_{13}^2] + 6\binom{n-1}{3}E[A_{12}A_{13}A_{14}] + 12\binom{n-1}{3}E[A_{12}A_{14}A_{23}] \right] \right\} \\
= \( E[(V_{n-1} - E[V_{n-1}])^3] + (n-1)E[A_{12}^3] + 18\binom{n-1}{2}E[A_{12}A_{13}^2] + 24\binom{n-1}{3}E[A_{12}A_{13}A_{14}] \) \\
+ 72\binom{n-1}{3}E[A_{12}A_{14}A_{23}] + 6\binom{n-1}{2}E[A_{12}A_{13}A_{23}] \)

Now, if we solve the recurrence subject to the initial values 
\( E[(V_1 - E[V_1])^3] = 0, \ E[(V_2 - E[V_2])^3] = E[A_{12}^3] \) and 
\( E[(V_3 - E[V_3])^3] = 3E[A_{12}^3] + 15E[A_{12}A_{13}^2] + 6E[A_{12}A_{13}A_{23}] \), we find 
\( E[(V_n - E[V_n])^3] = \binom{n}{2}E[A_{12}^3] + 18\binom{n}{3}E[A_{12}A_{13}^2] + 24\binom{n}{4}E[A_{12}A_{13}A_{14}] \) \\
+ 72\binom{n}{4}E[A_{12}A_{14}A_{23}] + 6\binom{n}{3}E[A_{12}A_{13}A_{23}] \)
In this formula, under the uniform null hypothesis, \( E[A_{12}^3] = \frac{1}{54} \), \( E[A_{12}A_{13}^2] = -\frac{11}{38880} \), \( E[A_{12}A_{13}A_{14}] = -\frac{1}{30240} \) but the other parameters cannot be computed exactly. Under the exponential null hypothesis, \( E[A_{12}^3] = \frac{19}{945} \), \( E[A_{12}A_{13}^2] = -\frac{9227}{12612600} \), \( E[A_{12}A_{13}A_{14}] = -\frac{1}{1155} \), \( E[A_{12}A_{13}A_{23}] = \frac{24659}{1323000} \), and \( E[A_{12}A_{14}A_{23}] = \frac{92}{93555} \).

Related Statistics

In this section, a method of combining the individual statistics \( U_1, U_2, \) and \( U_3 \) is described and properties of the resulting statistic are discussed. Since the sum of the three statistics is 1, it suffices to only consider two of the statistics. It is shown below that the statistic \( \bar{U} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \) is asymptotically multivariate normal with mean vector \( E\bar{U} = \begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \end{pmatrix} \) under the null hypothesis.

We know one component of the variance matrix, namely \( \sigma_i^2 = \text{Var}(U_i) \). A second component of the variance matrix is found by first noticing

\[
\left( \begin{array}{c} \text{Var}(U_2) \\ \text{Var}(U_3) \end{array} \right) = \text{Var}\left( \sum_{i=1}^{n-1} \sum_{j=1}^{n} \left| F(2X_i - X_j) - F(X_i) + F(X_j) - F(2X_i - X_j) \right| \right)
= \text{Var}\left( \sum_{i=1}^{n-2} \sum_{j=1}^{n-1} \left| F(2X_i - X_j) - F(X_i) + F(X_j) - F(2X_i - X_j) \right| \right) + (n-1)\beta_{22} + 3(n-1)(n-2)\alpha_{22}
\]

where \( \alpha_{22} = \text{Cov}\left(\left| F(2X_2 - X_1) - F(X_2) + F(X_1) - F(2X_1 - X_2) \right|, \left| F(2X_3 - X_1) - F(X_3) + F(X_1) - F(2X_1 - X_3) \right| \right) \)

and \( \beta_{22} = \text{Var}\left(\left| F(2X_2 - X_1) - F(X_2) + F(X_1) - F(2X_1 - X_2) \right| \right) \). Therefore, a recurrence relation (indexed by \( n \)) defines these variance terms. If we solve this recurrence relation subject to the required initial values (\( \text{Var}(U_2) = \beta_{22} \) when
n=2), we find \( \sigma^2 = \text{Var}(U_2) = \frac{\beta_{22}}{n} + \frac{2(n-2)}{n} \alpha_{22} \). If the null hypothesis is that the sample comes from a uniform distribution, for example, the parameters are \( \alpha_{22} = \frac{1}{720} \) and \( \beta_{22} = \frac{1}{36} \).

The third component of the variance matrix is

\[
\sigma_{12} = \text{Cov}(U_1, U_2) = \frac{4(n-2)}{n(n-1)} \alpha_{12} + \frac{\beta_{12}}{n} \text{ where}
\]

\[
\alpha_{12} = \text{Cov}(1 - |F(2X_2 - X_1) - F(2X_1 - X_2)|, |F(2X_2 - X_1) - F(X_2) + F(X_2) - F(2X_1 - X_2)|)
\]

and

\[
\beta_{12} = \text{Cov}(1 - |F(2X_2 - X_1) - F(2X_1 - X_2)|, |F(2X_2 - X_1) - F(X_2) + F(X_1) - F(2X_1 - X_2)|).
\]

For the uniform null hypothesis, the parameters are \( \alpha_{12} = \frac{1}{720} \) and \( \beta_{12} = -\frac{1}{24} \).

From a general theorem, we know that under the null hypothesis,

\[
(2.1) \quad Q = \frac{\sigma^2_1 \sigma^2_2}{|\Sigma|} \left\{ \left( \frac{U_1 - \frac{1}{2}}{\sigma_1} \right)^2 + \left( \frac{U_2 - \frac{1}{2}}{\sigma_2} \right)^2 - 2p \left( \frac{U_1 - \frac{1}{2}}{\sigma_1} \left( \frac{U_2 - \frac{1}{2}}{\sigma_2} \right) \right) \right\} \sim \chi^2_2
\]

where \( \Sigma = \begin{bmatrix} \sigma^2_1 & \sigma_{12} \\ \sigma_{12} & \sigma^2_2 \end{bmatrix} \) is the variance matrix found above. This will be our statistic which will be used to combine information from the individual statistics. Theorem 2.1 demonstrates the bivariate normality of any pair of statistics.
THEOREM 2.1. If $\alpha_{11}, \alpha_{22},$ and $\alpha_{33}$ are positive, then as $n \to \infty$, the asymptotic distribution of the pair $(U_r, U_s)$ is bivariate normal for each pair $(r,s)$.

PROOF.

The fact that $E(U_k) = \frac{1}{3}$ under the null hypothesis follows immediately from a symmetry argument. The asymptotic bivariate normality of $(U_r, U_s)$ follows from the fact that $U$ is a trivariate $U$-statistic [see e.g. Randles and Wolfe (1979)] with kernels having finite (since they are bounded) non-zero variances (by assumption).

To determine when the parameters $\alpha_{11}, \alpha_{22},$ and $\alpha_{33}$ are positive, let $C = \int \int h_i(x,y) dF(y) dF(x)$ where $h_i$; $i=1, 2, \text{ or } 3$ are the kernel functions of the $U$-statistics defining the individual triangle statistics. Notice that

$$\int \left[ \int h_i(x,y) dF(y) - C \right]^2 dF(x)$$

$$= \int \left[ \int h_i(x,y) dF(y) \right]^2 dF(x) - C^2 = \text{Cov}[h_i(X,Y), h_i(X,Y')] \]$$

So,

$\text{Cov}[h_i(X,Y), h_i(X,Y')] = 0$ iff $\int h_i(x,y) dF(y)$ is constant for almost all $x$. If $F$ is a continuous univariate distribution and $F$ has derivative $f$, then

$$\int h_3(x,y) dF(y) = \int F(x) - F(y) dF(y) + \int F(y) - F(x) dF(y)$$

$$= 2(F(x))^2 - F(x) - \int F(y) dF(y) + \int F(y) dF(y).$$

39
So, by the Fundamental Theorem of Calculus, this is constant if and only if 
\[ 0 = 2F(x)f(x) - f(x). \] This happens only when \( F(x) = 1/2 \). So, it doesn't happen 
almost everywhere. A similar argument holds for \( h_1 \) and \( h_2 \).

Now, notice that since \( h_1 + h_2 + h_3 = 1 \), there are some relationships that
make it easier to calculate the parameters of the null distribution. For
instance, we have \( \alpha_{i1} + \alpha_{i2} + \alpha_{i3} = \text{Cov}[h_i(X,Y),1] = 0 \) and
\( \beta_{i1} + \beta_{i2} + \beta_{i3} = \text{Cov}[h_i(X,Y),1] = 0 \) for \( i = 1, 2 \) or 3. Also, in one dimension,

\[
\alpha_{33} = \int \left\{ \int_{-\infty}^{x} \int_{-\infty}^{y} h_3(x,y) f(y) dy \right\}^2 f(x) dx - \frac{1}{9}
\]

\[
= \int \left\{ \int_{-\infty}^{x} \left[ F(x) - F(y) \right] f(y) dy + \int_{x}^{\infty} \left[ F(y) - F(x) \right] f(y) dy \right\}^2 f(x) dx - \frac{1}{9}
\]

\[
= \int \left\{ F(x)^2 - F(x) + \frac{1}{6} \right\}^2 f(x) dx - \frac{1}{9} = \frac{1}{180}
\]

and \( \beta_{33} = \int \left\{ \int_{-\infty}^{x} \int_{-\infty}^{y} h_3(x,y)^2 f(y) dy \right\} f(x) dx - \frac{1}{9} \)

\[
= \int \left\{ \int_{-\infty}^{x} \left[ F(x) - F(y) \right]^2 f(y) dy \right\} f(x) dx - \frac{1}{9} = \int \left\{ F(x)^2 - F(x) + \frac{1}{3} \right\} f(x) dx - \frac{1}{9} = \frac{1}{18}.
\]

Table 2.4 shows values of the parameters under different univariate
null distributions.
Table 2.4 Values of the parameters for various univariate null distributions.

<table>
<thead>
<tr>
<th>Dist.</th>
<th>$\alpha_{11}$</th>
<th>$\beta_{11}$</th>
<th>$\alpha_{22}$</th>
<th>$\beta_{22}$</th>
<th>$\alpha_{12}$</th>
<th>$\beta_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>0.001389</td>
<td>0.1111</td>
<td>0.001389</td>
<td>0.0278</td>
<td>0.00139</td>
<td>-0.04167</td>
</tr>
<tr>
<td>Beta($\frac{1}{2}, \frac{1}{2}$)</td>
<td>0.000269</td>
<td>0.1163</td>
<td>0.003419</td>
<td>0.0349</td>
<td>0.00093</td>
<td>-0.04785</td>
</tr>
<tr>
<td>Beta(1,2)</td>
<td>0.002927</td>
<td>0.1072</td>
<td>0.001736</td>
<td>0.0280</td>
<td>0.00045</td>
<td>0.03984</td>
</tr>
<tr>
<td>Beta(1,5)</td>
<td>0.004883</td>
<td>0.1030</td>
<td>0.002914</td>
<td>0.0308</td>
<td>-0.00112</td>
<td>0.03915</td>
</tr>
<tr>
<td>Beta(3,3)</td>
<td>0.003283</td>
<td>0.1050</td>
<td>0.000392</td>
<td>0.0237</td>
<td>0.00094</td>
<td>0.03655</td>
</tr>
<tr>
<td>Beta(5,3)</td>
<td>0.003696</td>
<td>0.1040</td>
<td>0.000500</td>
<td>0.0239</td>
<td>0.00068</td>
<td>0.03620</td>
</tr>
<tr>
<td>$\Gamma(5,1)$</td>
<td>0.004905</td>
<td>0.1014</td>
<td>0.001016</td>
<td>0.0254</td>
<td>-0.00018</td>
<td>-0.03564</td>
</tr>
<tr>
<td>Exp(1)</td>
<td>0.006349</td>
<td>0.1</td>
<td>0.003968</td>
<td>0.0333</td>
<td>-0.00238</td>
<td>-0.03889</td>
</tr>
<tr>
<td>N(0,1)</td>
<td>0.004632</td>
<td>0.1014</td>
<td>0.000284</td>
<td>0.0231</td>
<td>0.00032</td>
<td>-0.03450</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.005917</td>
<td>0.0983</td>
<td>0.000363</td>
<td>0.0236</td>
<td>-0.00036</td>
<td>-0.03317</td>
</tr>
<tr>
<td>Cauchy</td>
<td>0.012995</td>
<td>0.0827</td>
<td>0.002048</td>
<td>0.0340</td>
<td>-0.00474</td>
<td>-0.03056</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>0.005402</td>
<td>0.1011</td>
<td>0.002714</td>
<td>0.0202</td>
<td>-0.00899</td>
<td>-0.03804</td>
</tr>
<tr>
<td>$\chi^2_6$</td>
<td>0.005115</td>
<td>0.1013</td>
<td>0.001517</td>
<td>0.0269</td>
<td>-0.00054</td>
<td>-0.03633</td>
</tr>
</tbody>
</table>

**Multi-Dimensional Data**

Let $X_1, X_2, \ldots, X_n \sim F$ where $X_i$ is $d$-dimensional. To test $H_0: F=F_0$, let $Y \sim F_0$, let $\rho$ be a metric on $\mathbb{R}^d$. Define the statistic $U_1 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} h_1(X_i, X_j)$

where $h_1(X_i, X_j) = P[\rho(Y-X_i) > \rho(X_i-X_j), \rho(Y-X_j) > \rho(X_i-X_i)]$, for example. The other two statistics are defined similarly by replacing the one dimensional norm with the metric $\rho$. As before, the statistic $U_1$ is asymptotically normal with variance $\frac{\beta}{\alpha} + \frac{2(n-2)}{n} \alpha$ where

$$\alpha = \mathbb{E}\left[\left(h_1(X_1, X_2) - \frac{1}{3}\right)\left(h_1(X_1, X_3) - \frac{1}{3}\right)\right]$$

and $\beta = \text{Var}(h_1(X_1, X_2))$. Under $H_0$, the mean is $\frac{1}{3}$ again. One major difference is that exact calculation of the variance parameters is difficult and simulation is usually needed. As an illustration, $h_1(X_i, X_j)$ can be estimated by simulating a large number of
observations of \( Y \) and counting the proportion of triangles with vertices 
\( X_i, X_j \), and \( Y \) whose shortest leg is the side connecting \( X_i \) and \( X_j \).

As in the one-dimensional data case, a test can be constructed based on 
any of the three one-dimensional U-statistics or a combination of two of 
them. Table 2.5 shows the values of the parameters for different dimensions 
under the null hypothesis that the data is uniform on the hypercube.

**Table 2.5** Parameters for the covariance matrix of \( \tilde{U} \) under the null hypothesis uniform on the cube using the Euclidean norm found by simulation.

<table>
<thead>
<tr>
<th>( d )</th>
<th>( \alpha_{11} )</th>
<th>( \beta_{11} )</th>
<th>( \alpha_{22} )</th>
<th>( \beta_{22} )</th>
<th>( \alpha_{12} )</th>
<th>( \beta_{12} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3888E-03</td>
<td>0.1111</td>
<td>1.3888E-03</td>
<td>2.7777E-02</td>
<td>1.3888E-03</td>
<td>-4.1666E-02</td>
</tr>
<tr>
<td>2</td>
<td>1.1576E-03</td>
<td>0.1009</td>
<td>3.2165E-04</td>
<td>2.4659E-02</td>
<td>2.1717E-03</td>
<td>-3.1010E-02</td>
</tr>
<tr>
<td>3</td>
<td>2.3239E-03</td>
<td>9.6341E-02</td>
<td>1.1430E-03</td>
<td>2.2590E-02</td>
<td>9.8068E-04</td>
<td>-2.5536E-02</td>
</tr>
<tr>
<td>4</td>
<td>2.4511E-03</td>
<td>9.4946E-02</td>
<td>4.5456E-04</td>
<td>2.1876E-02</td>
<td>9.9473E-04</td>
<td>-2.4246E-02</td>
</tr>
<tr>
<td>7</td>
<td>1.1979E-04</td>
<td>1.5061E-02</td>
<td>3.2817E-04</td>
<td>1.2354E-02</td>
<td>1.9762E-04</td>
<td>-4.8480E-04</td>
</tr>
<tr>
<td>8</td>
<td>2.5569E-04</td>
<td>1.3464E-02</td>
<td>5.0775E-04</td>
<td>1.0296E-02</td>
<td>4.7615E-05</td>
<td>-2.2686E-04</td>
</tr>
<tr>
<td>9</td>
<td>2.6968E-04</td>
<td>1.1964E-02</td>
<td>4.0907E-04</td>
<td>8.2597E-03</td>
<td>-1.3575E-07</td>
<td>-4.2921E-05</td>
</tr>
<tr>
<td>10</td>
<td>3.8931E-04</td>
<td>1.0984E-02</td>
<td>3.1621E-04</td>
<td>7.083E-03</td>
<td>-6.8393E-05</td>
<td>1.0404E-04</td>
</tr>
</tbody>
</table>

**Variations on the Triangle Statistic**

Suppose \( \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1d} \end{pmatrix} \) and \( \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2d} \end{pmatrix} \) are independent \( d \)-dimensional random vectors where all of the components are iid \( \text{Unif}(0,1) \). Since edge effects make it difficult to evaluate the kernel functions which define the triangle statistics,
we may try to ignore the edges by allowing the third vertex of the triangle to lie outside of the unit cube. This idea could also be motivated from the perspective of using the usual triangle test statistic with a "wrap-around" distance. Let \( \begin{pmatrix} X_{31} \\ X_{32} \\ \vdots \\ X_{3d} \end{pmatrix} \) be independent of the aforementioned two vectors where the coordinates are iid Unif(-1,2). Let \( \rho \) denote the euclidean metric and \( F \) denote the distribution of a Unif(-1,2) random variable and

\[
d_{ij} = \rho \left( \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{id} \end{pmatrix}, \begin{pmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jd} \end{pmatrix} \right).
\]

We want to evaluate integrals of the following three functions:

i) \[
h_1 \left( \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1d} \end{pmatrix}, \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2d} \end{pmatrix} \right) = P[d_{12} < d_{13}, d_{12} < d_{23}] = 1 - h_2 - h_3
\]

ii) \[
h_2 \left( \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1d} \end{pmatrix}, \begin{pmatrix} X_{21} \\ X_{22} \\ \vdots \\ X_{2d} \end{pmatrix} \right) = P[d_{13} < d_{12} < d_{23} \text{ or } d_{23} < d_{12} < d_{13}] = \text{twice the volume of a } d\text{-dimensional sphere with radius } d_{12} \text{ minus } h_3 \]

of a \( d \)-dimensional sphere with radius \( d_{12} \) minus \( h_3 = \frac{2 \pi^{d/2} d_{12}^d}{\Gamma\left(\frac{d}{2} + 1\right)} - h_3 \).
iii) $h_3 \left( \begin{array}{c} X_{11} \\ X_{12} \\ \vdots \\ X_{1d} \\ X_{21} \\ X_{22} \\ \vdots \\ X_{2d} \end{array} \right) = \Pr[d_{13} < d_{12}, d_{23} < d_{12}]$ which is proportional to the volume of the $d$-dimensional sphere with radius $d_{12}$. The proportionality constant is not known, but the volume of the sphere is $\frac{\pi^\frac{d}{2} d_3^d}{\Gamma\left(\frac{d}{2} + 1\right)}$.

Since each of the 3 functions are a linear function of $d_{12}^d$, all three statistics are linear functions of each other and they are each equivalent to $U = \frac{1}{\binom{n}{2}} \sum_{i<j} h(X_i, X_j)$ where $h(X_i, X_j) = \left(\frac{(X_{i1} - X_{j1})^2 + \cdots + (X_{id} - X_{jd})^2}{2}\right)^\frac{n}{2}$. To find the expectation of $h$, notice that

\[ \left((x_1 - y_1)^2 + \cdots + (x_d - y_d)^2\right)^n = \sum_{n_1, n_2 \ldots n_d} \left(\begin{array}{c} n \\ n_1 \ n_2 \ldots n_d \end{array}\right) (x_1 - y_1)^{2n_1} \cdots (x_d - y_d)^{2n_d}. \]

Let $a_{d,n} = \int_0^1 \cdots \int_0^1 \left((x_1 - y_1)^2 + \cdots + (x_d - y_d)^2\right)^n dy_d \cdots dx_1$

\[ = \sum_{n_1, n_2 \ldots n_d} \left(\begin{array}{c} n \\ n_1 \ n_2 \ldots n_d \end{array}\right) \prod_{i=1}^d \frac{1}{(2n_i + 1)(n_i + 1)} \frac{1}{(2n_i + 1)(n_i + 1)} \cdots \frac{1}{(2n_d + 1)(n_d + 1)} \]

\[ = \sum_{n_1, n_2 \ldots n_d} \left(\begin{array}{c} n \\ n_1 \ n_2 \ldots n_d \end{array}\right) \pi(n_1, n_2 \ldots n_d) \prod_{i=1}^d \frac{1}{(2n_i + 1)(n_i + 1)} \]

where the sum is taken over all distinct partitions of $n$ and where $\pi(n_1, \ldots, n_d)$ is the number of permutations of the set $\{n_1, \ldots, n_d\}$. Also, $\left((x_1 - y_1)^2 + \cdots + (x_d - y_d)^2\right)^n \left((x_1 - z_1)^2 + \cdots + (x_d - z_d)^2\right)^n$

\[ = \sum \sum \left(\begin{array}{c} n \\ m_1 \ m_2 \ldots m_d \end{array}\right) \left(\begin{array}{c} n \\ n_1 \ n_2 \ldots n_d \end{array}\right) (x_1 - z_1)^{2m_1} (x_1 - y_1)^{2n_1} \cdots (x_d - z_d)^{2m_d} (x_d - y_d)^{2n_d}. \]

44
Now, let \( b_{d,n} = \int_0^1 \int_0^1 \cdots \int_0^1 \left( (x_1 - y_1)^2 + \cdots + (x_d - y_d)^2 \right)^n \left( (x_1 - z_1)^2 + \cdots + (x_d - z_d)^2 \right)^n \, dz_1 \cdots dx_1 \)

\[
= \sum \sum \left( \binom{n}{m_1 \ldots m_d} \binom{n}{n_1 \ldots n_d} \prod_{j=1}^d \frac{2}{(2m_j + 1)(2n_j + 1)(3 + 2m_j + 2n_j)} \left( 1 + \frac{1}{2m_j + 2n_j + 2} \right) \left( 1 + \frac{1}{2n_j + 1} \right) \right)
\]

\[
= 2 \sum \binom{n}{m_1 \ldots m_d} \binom{n}{n_1 \ldots n_d} \pi((m_1, n_1), \ldots, (m_d, n_d)) \prod_{j=1}^d \frac{1}{(2m_j + 1)(2n_j + 1)(3 + 2m_j + 2n_j)} \left( 1 + \frac{1}{2m_j + 2n_j + 2} \right) \left( 1 + \frac{1}{2n_j + 1} \right) \]

\[
+ \sum \binom{n}{n_1 \ldots n_d}^2 \pi(n_1, \ldots, n_d) \prod_{j=1}^d \frac{1}{(2n_j + 1)(3 + 4n_j)} \left( 1 + \frac{1}{4n_j + 2} \right) \left( 1 + \frac{1}{2n_j + 1} \right)
\]

where the first sum is taken over all distinct sets of ordered pairs \( \{(m_1, n_1), \ldots, (m_d, n_d)\} \) where \( m_1, \ldots, m_d \) and \( n_1, \ldots, n_d \) are partitions of \( n \) such that the partition \( m_1, \ldots, m_d \) is less than (is smaller in an arbitrary ordering) the partition \( n_1, \ldots, n_d \) and the latter sum is taken over all partitions of \( n \). For large values of \( n \), there are useful recurrence relations for the two sequences, namely,

\[
a_{d,n} = \int_0^1 \int_0^1 \cdots \int_0^1 \left( (x_1 - y_1)^2 + \cdots + (x_d - y_d)^2 \right)^n \, dy_1 \cdots dx_1 = \sum_{j=0}^n \binom{n}{j} a_{d-1,j} a_{1,n-j}
\]

and by introducing a third subscript on the second sequence,
\[ b_{k,m,n} = \int_0^1 \int_0^1 \left( \left( x_1 - y_1 \right)^2 + \ldots + \left( x_d - y_d \right)^2 \right)^m \left( x_1 - z_1 \right)^2 + \ldots + \left( x_d - z_d \right)^2 \right)^n dz_d \ldots dx_1 \]

\[ = \int_0^1 \int_0^1 \left( \sum_{i=0}^m \binom{m}{i} \left( x_1 - y_1 \right)^2 + \ldots + \left( x_{d-1} - y_{d-1} \right)^2 \right)^i \left( x_d - y_d \right)^{2(m-i)} \right) \left( \sum_{j=0}^n \binom{n}{j} \left( x_1 - z_1 \right)^2 + \ldots + \left( x_{d-1} - z_{d-1} \right)^2 \right)^j \left( x_d - z_d \right)^{2(n-j)} \right) dz_d \ldots dx_1 \]

\[ = \sum_{i=0}^m \sum_{j=0}^n \binom{m}{i} \binom{n}{j} b_{d-1,i,j} b_{1,m-1,n-j}. \]

Now, \( E[h(X, Y)] = a_d \beta, \quad \beta = a_{d,d} - \left( a_{d,d} \right)^2 \) and

\[ \alpha = b_{d,d} \left( a_{d,d} \right)^2. \]

Table 2.6 shows the values of these 3 parameters for even dimensions up to \( d=16 \).

<table>
<thead>
<tr>
<th>( d )</th>
<th>( E[h(X, Y)] )</th>
<th>( \alpha )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.333333333333</td>
<td>.01111111111</td>
<td>.07777777778</td>
</tr>
<tr>
<td>4</td>
<td>.600000000000</td>
<td>.06613756614</td>
<td>.4597883598</td>
</tr>
<tr>
<td>6</td>
<td>1.769841270</td>
<td>.8517546827</td>
<td>6.700512274</td>
</tr>
<tr>
<td>8</td>
<td>7.287407407</td>
<td>19.51362438</td>
<td>178.4085912</td>
</tr>
<tr>
<td>10</td>
<td>38.54834055</td>
<td>699.1611785</td>
<td>7532.682629</td>
</tr>
<tr>
<td>12</td>
<td>249.1446421</td>
<td>36096.94598</td>
<td>462995.4393</td>
</tr>
<tr>
<td>14</td>
<td>1902.751044</td>
<td>.2538057968e07</td>
<td>.3908856207e08</td>
</tr>
<tr>
<td>16</td>
<td>16765.78141</td>
<td>.2332052505e09</td>
<td>.4344875252e10</td>
</tr>
</tbody>
</table>

There is a very strong log-linear relationship between the three parameters and between \( d \) and the log of any one of the parameters. For example, Figure 2.9 shows the graphs of \( n \) vs. the logarithms of the parameters.
Figure 2.9 Scatterplots of d versus logarithms of the three null parameters
2.3 Examples

Example 1. Binomial Data

Suppose $X_i, i=1,2,\ldots, n$ are iid Bernoulli random variables with probability of success $p$. Let $S = \sum_{i=1}^{n} X_i$, then the UMP unbiased test of $H_0: p = p_0$ versus

$$H_1: p \neq p_0 \text{ is } \phi(S) = \begin{cases} \gamma_i & \text{if } S < C_i \text{ or } S > C_2 \\ 0 & \text{otherwise} \end{cases}$$

where the constants are chosen so that

$$\sum_{x=C_i+1}^{C_i-1} \binom{n}{x} p_0^x q_0^{n-x} + \sum_{i=1}^{2} (1-\gamma_i) \binom{n}{C_i} p_0^{C_i} q_0^{n-C_i} = 1 - \alpha \text{ and}$$

$$\sum_{x=C_i+1}^{C_i-1} \binom{n-1}{x-1} p_0^{x-1} q_0^{n-x} + \sum_{i=1}^{2} (1-\gamma_i) \binom{n-1}{C_i-1} p_0^{C_i-1} q_0^{n-C_i} = 1 - \alpha.$$ 

By using the normal approximation, an approximate test rejects when $|S - np_0| > 1.96\sqrt{np_0q_0}$. The triangle test statistic based on the number of smallest legs is

$$U_1 = \frac{1}{n} \sum_{i=1}^{n} h_1(X_i, X_i) \text{ where } h_1(X, Y) = \begin{cases} \frac{1}{3} P[Z = X] + P[Z \neq X] & \text{if } X = Y \text{ where } Z \text{ is}\\ 0 & \text{otherwise} \end{cases}$$

Bernoulli(p), i.e. $h_1(X, Y) = \begin{cases} \frac{3-2p}{3} & \text{if } X = Y = 1 \\ \frac{2p+1}{3} & \text{if } X = Y = 0 \\ 0 & \text{otherwise} \end{cases}$

Thus, $U_1 = \frac{\sum_{i=1}^{n} h_1(X_i, X_i)}{n}$. Under $H_0$, Var $U_1$

$$= \frac{\text{Var}[h_1(X, Y)]}{n} + 2(n-2)\text{Cov}[h_1(X, Y), h_1(X, Y')]$$

where $\text{Var}[h_1(X, Y)]$

$$= \frac{3}{2} (\frac{3-2p}{3}) + \frac{2p+1}{3}$$

and $\text{Cov}[h_1(X, Y), h_1(X, Y')] = \frac{p(1-p)(1-2p)^2}{9}$. Now,
assume that $0 < p_0 < 1/2$. An approximate level 0.05 test rejects the null
hypothesis when

$$\left| \frac{U_1 - \frac{1}{3}}{\sqrt{\text{Var} U_1}} \right| > 1.96.$$ 

Since $U_1$ is a parabola in $S$ with vertex at a point greater than ES, Figure 2.10 shows the acceptance region.

![Figure 2.10 Acceptance region](image)

The lower cutoff, $S_L$, satisfies

$$\left( \frac{S_L}{2} \right) \left[ \frac{3 - 2p}{3} \right] + \left( \frac{n - S_L}{2} \right) \left[ \frac{2p + 1}{3} \right] = \frac{1}{3} + 1.96 \sqrt{\text{Var} U_1}$$

and the upper cutoff satisfies a similar equation. Table 2.7 shows the critical values for the classical test and the Triangle test.
Table 2.7 Rejection region cutoff values

<table>
<thead>
<tr>
<th></th>
<th>standard appr.</th>
<th>exact C1, C2, γ1, γ2</th>
<th>triangle statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=1000, p=0.25</td>
<td>223.162, 276.838</td>
<td>223, 277, 0.99, 0.33</td>
<td>224.903, 279.622</td>
</tr>
<tr>
<td>n=1000, p=0.35</td>
<td>320.437, 379.563</td>
<td>321, 380, 0.13, 0.74</td>
<td>323.397, 387.804</td>
</tr>
<tr>
<td>n=10000, p=0.25</td>
<td>2415.13, 2584.87</td>
<td>2417, 2587, 0.99, 0.16</td>
<td>2417.12, 2587.17</td>
</tr>
<tr>
<td>n=10000, p=0.35</td>
<td>3406.51, 3593.49</td>
<td>3407, 3594, 0.99, 0.04</td>
<td>3410.29, 3598.51</td>
</tr>
</tbody>
</table>

Example 2. General Discrete Data.

Suppose $X_i$ i=1,2, ..., n are iid discrete random variables with support {1, 2, ..., N} and we want to test $H_0: f(i) = P[X_i = i] = p_i$. Let $S_j = \sum_{i=1}^{n} I_{ij}(X_i)$ be the number of observations equal to j. Pearson's $\alpha$-level chi-square test rejects when $\sum_{j=1}^{N} \frac{(S_j - np_j)^2}{np_j} > \chi^2_\alpha (N - 1)$. A natural way to conduct a test based on the triangle statistic with the number of smallest legs is to place N points in N-1 dimensional Euclidean space such that each pair of points are equidistant.

The statistic is $U_i = \frac{1}{\binom{n}{2}} \sum_{j \neq i} h_i(X_i, X_j)$ where

\[
h_i(X, Y) = \begin{cases} 
\frac{1}{3}P[Z = X] + P[Z \neq X] & \text{if } X = Y \\
\frac{1}{3}P[Z \neq X, Z \neq Y] & \text{otherwise}
\end{cases}
\]

Here, Z is a random variable with the same distribution as $X_1$, in other words,

\[
h_i(X, Y) = \begin{cases} 
\frac{3 - 2p_x}{3} & \text{if } X = Y \\
\frac{3 - p_x - p_y}{3} & \text{if } X \neq Y
\end{cases}
\]
\[
U_1 = \sum_{i=1}^{N} \binom{N}{i} \left( \frac{3-2p_i}{3} \right) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} S_i S_j \left( \frac{1-p_i - p_j}{3} \right)
\]

So, \( U_1 = \frac{n}{2} \sum_{i<j} h_2(X_i, X_j) \) where

\[
h_2(X, Y) = \begin{cases} 
\frac{1}{3} P[Z = X] & \text{if } X = Y \\
\frac{1}{3} P[Z \neq X, Z \neq Y] + \frac{1}{2} P[Z = X \text{ or } Z = Y] & \text{otherwise}
\end{cases}
\]

But, \( U_2 = \frac{\sum_{i=1}^{N} \binom{N}{2} \binom{p_i}{3} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_i S_j \left( \frac{2+p_i + p_j}{6} \right)}{\binom{n}{2}} \) and so \( \binom{n}{2} \{ 2U_2 + U_1 \} \)

\[
= \sum_{i=1}^{N} \binom{S_i}{2} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_i S_j = -\frac{n}{2} + \frac{1}{2} \sum_{i=1}^{N} S_i^2 + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_i S_j = -\frac{n}{2} + \frac{1}{2} \left( \sum_{i=1}^{N} S_i \right)^2 = \frac{n^2 - n}{2} .
\]

This implies that the two statistics are equivalent for testing purposes. Under \( H_0, \)

\[
\text{Var} [U_1] = \frac{\text{Var}[h_1(X, Y)]}{\binom{n}{2}} + 2(n-2)\text{Cov}[h_1(X, Y), h_1(X, Y')] \]

where

\[
\text{Var}[h_1(X, Y)] = 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i p_j \left( \frac{1-p_i - p_j}{3} \right)^2 + \sum_{i=1}^{N} \binom{p_i}{2} \left( \frac{3-2p_i}{3} \right)^2 - \frac{1}{9}
\]

so that

\[
9 \text{Var}[h_1(X, Y)] = \sum_{i=1}^{N} \sum_{j=i}^{N} p_i p_j (1-p_i - p_j)^2 + \sum_{i=1}^{N} p_i^2 (3-2p_i)^2 - 1
\]

\[
= \sum_{i=1}^{N} p_i (1-p_i)^2 \sum_{j=1}^{N} p_j - 2 \sum_{i=1}^{N} p_i (1-p_i) \sum_{j=i+1}^{N} p_j^2 + \sum_{i=1}^{N} p_i \sum_{i=1}^{N} p_j^3 + \sum_{i=1}^{N} p_i^2 (3-2p_i)^2 - 1
\]

\[
= \sum_{i=1}^{N} p_i (1-p_i)^3 - 2 \sum_{i=1}^{N} p_i (1-p_i) \left( -p_i^2 + \sum_{j=1}^{N} p_j^2 \right)
\]

\[
+ \sum_{i=1}^{N} p_i \left( p_i^3 + \sum_{j=1}^{N} p_j^3 \right) + \sum_{i=1}^{N} p_i^2 (3-2p_i)^2 - 1
\]

\[51\]
\[
(1 - 1) + (-3 - 2 + 9) \sum_{i=1}^{N} p_i^2 + (3 + 2 + 1 - 12) \sum_{i=1}^{N} p_i^3 + (1 - 2 - 1 + 4) \sum_{i=1}^{N} p_i^4 + 2 \left( \sum_{i=1}^{N} p_i^2 \right)^2 \\
= 4 \sum_{i=1}^{N} p_i^2 - 6 \sum_{i=1}^{N} p_i^3 + 2 \sum_{i=1}^{N} p_i^4 + 2 \left( \sum_{i=1}^{N} p_i^2 \right)^2.
\]

Also, \(\text{Cov}[h_i(X, Y), h_i(X', Y')] = \sum_{j \neq i}^{N} \frac{1}{9} \left\{ \frac{3 - 2p_i}{3} \right\}^2 + \frac{1}{9} \left\{ \frac{1 - p_i - p_j}{3} \right\}^2 - \frac{1}{9} \]

\[
= \frac{1}{9} \left\{ \sum_{i=1}^{N} p_i \left( 3 - 2p_i \right) + \sum_{j \neq i}^{N} p_j \left( 1 - p_i \right) - \sum_{j \neq i}^{N} p_j^2 \right\} - 1
\]

\[
= \frac{1}{9} \left\{ \sum_{i=1}^{N} p_i \left( 3 - 2p_i \right) + (1 - p_i)^2 - \left( \sum_{j \neq i}^{N} p_j^2 - p_i^2 \right) \right\} - 1
\]

\[
= \frac{1}{9} \left\{ 1 - \sum_{j=1}^{N} p_j^2 \right\} + 2 \left( 1 - \sum_{j=1}^{N} p_j^2 \right) \left( \sum_{j=1}^{N} p_j^2 \right) + \sum_{i=1}^{N} p_i^3 - 1
\]

\[
= \frac{1}{9} \left\{ \sum_{i=1}^{N} p_i^3 - \left( \sum_{j=1}^{N} p_j^2 \right)^2 \right\}.
\]

\(U_1\) will be asymptotically normal unless \(\text{Cov}[h_i(X, Y), h_i(X', Y')] = 0\). This occurs when \(p_i = \frac{1}{N}\) but for no other cases provided all \(p_i > 0\). In the situation where \(p_i = \frac{1}{N}\), the statistic is

\[
U_1 = \frac{3N - 2 \sum_{j=1}^{N} S_j \binom{n}{2} + \frac{N - 2}{3N} \sum_{j=1}^{N} \sum_{i=1}^{N} S_i S_j}{3N - 2 \sum_{j=1}^{N} S_j \binom{n}{2} + \frac{N - 2}{6N} \sum_{j=1}^{N} S_j^2 + \frac{N - 2}{6N} \binom{n}{2}}
\]

\[
= \frac{\frac{3N - 2}{6N} \sum_{j=1}^{N} S_j^2 + \frac{N - 2}{n} n^2 - \frac{3N - 2}{6N} n}{\binom{n}{2}}.
\]

Hence, this statistic is a linear function of Pearson's chi-square statistic which reduces to \(\chi^2 = \frac{N}{n} \sum_{j=1}^{N} S_j^2 - n\) in this case.

52
The asymptotic distribution also follows from a theorem in Gregory (1977) which provides the limiting distribution of degenerate U-statistics. To use this theorem, we need to find the eigenvectors of the matrix $H$ where $h_{ij}=h_1(i,j)-1/3$. This is because $x$ is an eigenvector of $H$ iff
\[ \sum_{j=1}^{N} \left( h_1(i,j) - \frac{1}{3} \right) p_j x_j = \lambda x_i \text{ for all } i. \]

For illustration, let's take the example of testing whether a die is fair. In this case, $N=6$ and $p_i=1/6$ for all $i$. Here, $h_{i,i} = \begin{cases} 5/9 & \text{if } i = j \\ -1/9 & \text{otherwise} \end{cases}$ and we find that $0$ is an eigenvalue with eigenvector $(1,1,1,1,1,1)$ and $6$ is an eigenvector of multiplicity $5$ with eigenvectors: $(-1,1,0,0,0,0)$, $(-1,0,1,0,0,0)$, $(-1,0,0,1,0,0)$, and $(-1,0,0,0,1,0)$. Hence, Gregory’s theorem says that the limiting distribution of a linear function of the statistic is $P\left\{ \sum_{k=2}^{6} 6[Z_k^2 - 1] \leq x \right\}$ since the first eigenvalue is $0$ and $a_k = 0$ for each $k>1$.

Therefore, the statistic is asymptotically a linear function of a chi-square r.v. with $5$ degrees of freedom.

**Example 3. Roulette Wheel Testing**

In this example, we consider a testing situation in which the sample consists of discrete ordinal random variables. This differs from the nominal case because with ordinal variables, an interpoint distance can be defined which takes into account the ordering on the support of the variables.

Consider a continuous roulette wheel where a ball is spun on the unit circle and it lands at some point on the circle. The operator of the roulette
wheel is cheating so that the ball does not land uniformly on the circle but
rather if $\phi$ denotes the angle (measured in $2\pi$ radians from the positive $x$ axis)
where the ball lands, then the density of $\phi$ is $f(\phi) = 1 + \tau \sin 2\pi \phi$ where $\tau$ is an
unknown parameter that is between -1 and 1. We wish to estimate $\tau$ based on
a random sample $\phi_1, \phi_2, \ldots, \phi_n$. The log-likelihood is
$$\log \prod_{i=1}^{n} (1 + \tau \sin 2\pi \phi_i) = \sum_{i=1}^{n} \log (1 + \tau \sin 2\pi \phi_i)$$
and its derivative is
$$\sum_{i=1}^{n} \frac{\sin 2\pi \phi_i}{1 + \tau \sin 2\pi \phi_i} = \sum_{i=1}^{n} \frac{1}{\csc 2\pi \phi_i + \tau}$$
and the maximum likelihood estimate is a root
of this function. To find the parameters for the statistic $Q$ defined by equation
2.1 using the distance measured around the circle, notice that

$$h_1(\phi, \phi) = \begin{cases} 
\int_{\phi}^{\phi + 1 - \frac{1}{3}} f(t) dt & \text{if } \phi \leq \phi + \frac{1}{3} \\
0 & \text{otherwise}
\end{cases}$$

and

$$h_3(\phi, \phi) = \begin{cases} 
\int_{\phi}^{\phi + \frac{1}{3}} f(t) dt & \text{if } \phi \leq \phi + \frac{1}{3} \\
\int_{\phi + \frac{1}{3}}^{\phi + \frac{1}{2}} f(t) dt + \int_{\phi + \frac{1}{2}}^{\phi + \frac{1}{3}} f(t) dt & \text{if } \phi + \frac{1}{3} \leq \phi + \frac{1}{2} \\
\int_{\phi + \frac{1}{3}}^{\phi + \frac{1}{2}} f(t) dt + \int_{\phi + \frac{1}{2}}^{\phi + \frac{1}{3}} f(t) dt & \text{if } \phi + \frac{1}{2} \leq \phi + \frac{1}{3} \\
\int_{\phi + \frac{1}{3}}^{\phi + \frac{2}{3}} f(t) dt & \text{if } \phi + \frac{2}{3} \leq \phi + 1 \\
0 & \text{otherwise}
\end{cases}$$
where the inequalities are defined counterclockwise on the circumference of the circle. For example, \( \alpha_1 = \text{Var}[h_1(\varphi, \phi)] = 1/9 \)

\[
= \int_0^1 f(\varphi) \left[ \int_{\varphi/3}^{\varphi+1/3} \left( \int_{\varphi+1/3}^{\varphi+2/3} f(t)dt \right) f(\phi)d\phi + \int_{\varphi+2/3}^{\varphi+3/3} \left( \int_{\varphi+3/3}^{\varphi+4/3} f(t)dt \right) f(\phi)d\phi \right] d\varphi - \frac{1}{9}.
\]

This example demonstrates the flexibility of the triangle statistic.

\textbf{Example 4. Monte Carlo Study of Tests for Uniformity}

A comparison of tests based on \( U_1, U_2 \) and \( Q \) against each other and some other goodness of fit tests can be made versus the alternatives in Families 1 and 3 in Miller and Quesenberry (1979). Both families are indexed by a parameter \( j \) and are univariate. If \( U \) is uniform on \((0, 1)\), then the alternative from Family 1 has the distribution of \( U^{j+1} \) and the alternative from Family 2 has the distribution of \((1-2B)U^{j+1} \) where \( B \) is independent of \( U \) and has a binomial distribution with \( p=1/2 \). The proportion (out of 10,000) of 0.05 level tests that rejected the null hypothesis when the data were simulated from the various alternatives are shown in Table 2.8. In the table, \( N \) is the sample size and in the second part of the table, \( k \) is the dimension of vectors consisting of independent coordinates with the indicated distribution.
<table>
<thead>
<tr>
<th>N</th>
<th>j</th>
<th>Family 1</th>
<th>Family 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$U_1$</td>
<td>$U_2$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.0192000</td>
<td>0.0155000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.0496000</td>
<td>0.0434000</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.0975999</td>
<td>0.0898000</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.1494000</td>
<td>0.1454000</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.1341000</td>
<td>0.1920000</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.2864000</td>
<td>0.3674000</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.4175000</td>
<td>0.5183000</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0.5329000</td>
<td>0.6145000</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.2223000</td>
<td>0.3181000</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.4968000</td>
<td>0.6347000</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>0.6996000</td>
<td>0.8180000</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.8227000</td>
<td>0.9044000</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.2843000</td>
<td>0.4435000</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.6663000</td>
<td>0.8079000</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>0.8581000</td>
<td>0.9405000</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>0.9383000</td>
<td>0.9808000</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.3684000</td>
<td>0.5377000</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0.7788000</td>
<td>0.9023000</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>0.9372000</td>
<td>0.9800000</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>0.9804000</td>
<td>0.9977000</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>0.6774000</td>
<td>0.8821000</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>0.9849000</td>
<td>0.9991000</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>0.9999000</td>
<td>1.0000000</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>1.0000000</td>
<td>1.0000000</td>
</tr>
<tr>
<td>N</td>
<td>k=3 X_1~Beta(2,2)</td>
<td>k=5 X_1~Beta(0.5,0.5)</td>
<td>k=3 X_1~Family 1, j=1</td>
</tr>
<tr>
<td>----</td>
<td>------------------</td>
<td>---------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>3</td>
<td>0.2810000</td>
<td>0.4050000</td>
<td>0.12900000</td>
</tr>
<tr>
<td>4</td>
<td>0.3960000</td>
<td>0.5160000</td>
<td>0.16000001</td>
</tr>
<tr>
<td>5</td>
<td>0.4640000</td>
<td>0.5620000</td>
<td>0.17899999</td>
</tr>
<tr>
<td>6</td>
<td>0.5600000</td>
<td>0.6570000</td>
<td>0.19500000</td>
</tr>
<tr>
<td>7</td>
<td>0.6140000</td>
<td>0.7370000</td>
<td>0.25000000</td>
</tr>
<tr>
<td>8</td>
<td>0.6910000</td>
<td>0.7850000</td>
<td>0.28800000</td>
</tr>
<tr>
<td>9</td>
<td>0.7420000</td>
<td>0.8230000</td>
<td>0.31800002</td>
</tr>
<tr>
<td>10</td>
<td>0.7860000</td>
<td>0.8730000</td>
<td>0.37900001</td>
</tr>
<tr>
<td>11</td>
<td>0.8180000</td>
<td>0.8830000</td>
<td>0.39500002</td>
</tr>
<tr>
<td>12</td>
<td>0.8470000</td>
<td>0.9140000</td>
<td>0.43200001</td>
</tr>
<tr>
<td>13</td>
<td>0.8780000</td>
<td>0.9340000</td>
<td>0.46900000</td>
</tr>
<tr>
<td>14</td>
<td>0.9120000</td>
<td>0.9530000</td>
<td>0.51500000</td>
</tr>
<tr>
<td>15</td>
<td>0.9290000</td>
<td>0.9600000</td>
<td>0.56699990</td>
</tr>
<tr>
<td>16</td>
<td>0.9440000</td>
<td></td>
<td>0.55199999</td>
</tr>
<tr>
<td>17</td>
<td>0.9540000</td>
<td></td>
<td>0.60600001</td>
</tr>
<tr>
<td>18</td>
<td>0.9650000</td>
<td></td>
<td>0.65200001</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>0.64599998</td>
<td>0.75300001</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>0.68599999</td>
<td>0.80300000</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>0.72999999</td>
<td>0.83600000</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td>0.73200002</td>
<td>0.89000002</td>
</tr>
<tr>
<td>23</td>
<td></td>
<td>0.75300001</td>
<td>0.89900002</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>0.76700002</td>
<td>0.98499998</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>0.86999997</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td>0.95700000</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>0.97599998</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 5. Uniform Versus Tent Shaped Alternatives

For the continuous case, we'll compare three tests for the problem of testing if a sample $X_1$, $X_2$, $X_n$ is from a Unif(0,1) distribution vs. the alternative that it's from a "tent shaped" alternative with density

$$f(x) = \begin{cases} 1-t+4tx & \text{if } 0 < x < \frac{1}{2} \\ 1+3t-4tx & \text{if } \frac{1}{2} < x < 1 \end{cases}.$$ The first test is the triangle test, $Q$. The second is the likelihood ratio test, LR. It is easier to compute the LR if we assume the we observe the sample $Y_1, Y_2, \ldots, Y_n$ where $Y_i = \begin{cases} X_i & \text{if } 0 < X_i < \frac{1}{2} \\ \frac{1}{2}-X_i & \text{if } \frac{1}{2} < X_i < 1 \end{cases}$ and we test whether the $Y$'s come from a uniform distribution on $(0,\frac{1}{2})$ versus the alternative that the $Y$'s come from a distribution with density

$$f(y) = \begin{cases} 2-2t+8ty & \text{if } 0 < y < \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases}.$$ To calculate LR, we need to find the maximum likelihood estimate for $t$. Notice that the log-likelihood is

$$\ell(t) = \sum_{i=1}^{n} \log(2-2t+8ty_i).$$ Hence, $\ell'(t) = \sum_{i=1}^{n} \frac{4y_i - 1}{1+t(4y_i - 1)} = \sum_{i=1}^{n} \frac{z_i}{1+tz_i}$ where $z_i = 4y_i - 1$. Now, $\ell'(t) = 0$ provided $\sum_{i=1}^{n} \prod_{j \neq i} (1 + tz_j) = n$. But,

$$\prod_{j \neq i} (1 + tz_j) = 1 + t \sum_{j \neq i} z_j + t^2 \sum_{j \neq i, j \neq k} z_j z_k + \cdots + t^{n-1} \prod_{j \neq i} z_j$$ and $\sum_{i=1}^{n} \prod_{j \neq i} (1 + tz_j) = n + t(n-1) \sum_{j} z_j + t^2(n-2) \sum_{j \neq k} z_j z_k + \cdots + t^{n-2} \sum_{j \neq i} \prod_{j \neq i} z_j$. So, the maximum likelihood estimate for $t$ is a root of $(n-1) \sum_{j} z_j + t(n-2) \sum_{j \neq k} z_j z_k + \cdots + t^{n-2} \sum_{j \neq i} \prod_{j \neq i} z_j = 0$. Once the maximum likelihood estimate for $t$, say $\hat{t}$, is found, LR statistic is

$$-2[\ell(\hat{t}) - \ell(0)]$$ and is asymptotically $\chi^2$ with $n-1$ degrees of freedom. The third test statistic is

$$\int_{\mathbb{R}} \frac{n(\hat{f}(x) - f(x))^2}{f^2(x)} \, dF(x)$$ where $\hat{f}$ is an estimate of the density based
on the sample and \( f \) is the hypothesized density. When the sample is from a
discrete distribution with support \( 1, 2, \ldots, N \) and we use
\[
\hat{f}(x) = \frac{\# \text{ of observations equal to } x}{n},
\]
this statistic is Pearson’s chisquare statistic
\[
\int \frac{n(\hat{f}(x) - f(x))^2}{f^2(x)} \, dF(x) = \sum_{i=1}^{N} \frac{n(\hat{f}(i) - f(i))^2}{f(i)} = \sum_{i=1}^{N} \frac{(\text{obs}_i - \text{exp}_i)^2}{\text{exp}_i}
\]
in this case. One estimator of the density that we could use is
\[
\hat{f}(x) = \begin{cases} 
1 & \text{if } 0 < x < \frac{x_{(0)}}{2} \\
\frac{1 + x_{(0)} - x_{(i)}}{n(x_{(i)} - x_{(i-1)})} & \text{if } \frac{x_{(0)} + x_{(i-1)}}{2} < x < \frac{x_{(i)} + x_{(i-1)}}{2} \\
1 & \text{if } x > \frac{1 + x_{(0)}}{2}.
\end{cases}
\]
ordered sample and we define \( x_{(0)} = 0 \) and \( x_{(n+1)} = 1 \). The third test statistic then
would be
\[
\sum_{i=1}^{n} \frac{(1 + x_{(i-1)} - x_{(0)})^2}{n(x_{(i)} - x_{(i-1)})} = \sum_{i=1}^{n} \frac{(1 + x_{(i)} - x_{(0)})^2}{n} - \frac{(x_{(i+1)} - x_{(i-1)})^2}{2(x_{(i)} - x_{(i-1)})}.
\]
Another estimator that we could use is
\[
\hat{f}(x) = \frac{1}{(n+1)(x_{(i+1)} - x_{(0)})} I_{(x_{(0)}, x_{(i+1)})}(x).
\]
The test statistic then would be
\[
\sum_{i=0}^{n} \left( \frac{1}{n+1} - \frac{x_{(i+1)} - x_{(i)}}{x_{(i+1)} - x_{(0)}} \right)^2 \frac{n}{(x_{(i+1)} - x_{(0)})}.
\]
Both of these statistics belong to a class of statistics based on spacings of order statistics. For example,
the statistic
\[
\sum_{i=0}^{n} \left( \frac{1}{n+1} - \frac{x_{(i+1)} - x_{(i)}}{x_{(i+1)} - x_{(0)}} \right)^2
\]
is discussed in Kimball (1947) who shows
that the asymptotic distribution is normal and was also suggested by Irwin in
the discussion of Greenwood (1946). A general review of statistics based on
spacings of order statistics is in Pyke (1965). The statistic
\[
\sum_{i=0}^{n} \left( \frac{1}{n+1} - \frac{x_{(i+1)} - x_{(i)}}{x_{(i+1)} - x_{(0)}} \right)^2
\]
has low power, so we’ll use the statistic
\[
\chi^2 = \sum_{i=0}^{n} \left( \frac{1}{n+1} - \frac{x_{(i+1)} - x_{(i)}}{x_{(i+1)} - x_{(0)}} \right)^2.
\]
Empirical quantiles of the distribution of this
statistic were found by simulation (for samples of size \( n=50, 100 \) and \( 200 \)) and
are shown in Table 2.9.
Table 2.9 Critical values of $\chi^2 = \sum_{i=0}^{n-1} \left( \frac{1}{\sigma^2} - (x_{(i+1)} - x_{(i)}) \right)^2$

<table>
<thead>
<tr>
<th>n</th>
<th>95th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.1672728</td>
</tr>
<tr>
<td>100</td>
<td>.1143469</td>
</tr>
<tr>
<td>200</td>
<td>7.85120E-02</td>
</tr>
</tbody>
</table>

Figure 2.11 shows the graphs of the power (for a size 0.05 test) of Q, LR, and $\chi^2$ for the continuous case and the corresponding three tests for the discrete case for samples of size 50, 100 and 200.
Figure 2.11 Power of the three tests for the discrete case with N=5 or 10 and the continuous case, N=∞ (brown=LR, pink=Q, green=χ²).
Figure 2.11 (continued)
Power of the three tests when the alternative is a shifted tent shaped distribution

Table 2.10 shows the power of the three tests for a "shifted tent shaped" distribution. This distribution is concentrated on (0, 1) and has the density

\[ f(x) = \begin{cases} 1+4x & \text{if } 0 < x < 1/4 \\ 3-4x & \text{if } 1/4 < x < 3/4 \\ -3+4x & \text{if } 3/4 < x < 1 \end{cases} \]

Here, LR is not the most powerful test because the alternative is not within the class for which it is designed.

Table 2.10 Power of the tests for a shifted tent shaped alternative

Percent of rejections from 1000 samples ($\alpha=0.05$  $N=\infty$  $n=1000$)

<table>
<thead>
<tr>
<th>n</th>
<th>t</th>
<th>Q</th>
<th>Pearson</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1.0</td>
<td>97</td>
<td>84</td>
<td>3</td>
</tr>
<tr>
<td>100</td>
<td>1.0</td>
<td>100</td>
<td>98</td>
<td>5</td>
</tr>
<tr>
<td>200</td>
<td>1.0</td>
<td>100</td>
<td>100</td>
<td>5</td>
</tr>
</tbody>
</table>

ARE of triangle test versus LR

Under the alternative hypothesis, the LR has approximately a noncentral $\chi^2$ distribution with 1 d.f. and noncentrality parameter

\[ \lambda = -t^2E\left(\frac{\partial^2 \log L}{\partial t^2}\right) = -t^2E\left(\sum_{i=1}^{n} \frac{\partial^2 \log(2-2t+8ty)}{\partial t^2}\right) \]

\[ = nt^2E\left(\frac{2-8y}{2-2t+8ty}\right)^2 = nt^2 \int_{0}^{1} \frac{(2-8y)^2}{2-2t+8ty} dy = \frac{n}{2t} \log \left(\frac{1+t}{1-t}\right) - n \]
[see Kendall, Stuart and Ord (1983)]. The expectations and variances of the triangle statistics under the alternative can be found in Table 2.11.

**Table 2.11** Parameters under the alternative

<table>
<thead>
<tr>
<th>$E^*U_1$</th>
<th>$\alpha^{*11}$</th>
<th>$\beta^{*11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{40 + 5t + 6t^2}{120}$</td>
<td>$\frac{560 + 2275t + 2880t^2 - 375t^3 - 1008t^4}{403200}$</td>
<td>$\frac{43200 - 4140t + 4825t^2 - 1620t^3 - 972t^4}{388800}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$E^*U_2$</th>
<th>$\alpha^{*22}$</th>
<th>$\beta^{*22}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{40 + 5t - 4t^2}{120}$</td>
<td>$\frac{560 - 1715t + 1260t^2 + 870t^3 - 448t^4}{403200}$</td>
<td>$\frac{10800 + 720t + 505t^2 + 1080t^3 - 432t^4}{388800}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$E^*U_3$</th>
<th>$\alpha^{*33}$</th>
<th>$\beta^{*33}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{20 - 5t - t^2}{60}$</td>
<td>$\frac{20 + 10t - 5t^2 - 5t^3 - t^4}{3600}$</td>
<td>$\frac{200 - 100t + 15t^2 - 10t^3 - t^4}{3600}$</td>
</tr>
</tbody>
</table>

The power of a 0.05 level test using the $Q$ statistic is $P[Q > 5.99146]$

$$= P \left[ \left( \frac{U_1 - \bar{d}}{\sqrt{s_{11}}} \right)^2 + \left( \frac{U_2 - \bar{d}}{\sqrt{s_{22}}} \right)^2 - 2 \frac{s_{12}}{\sqrt{s_{11}} \sqrt{s_{22}}} \left( \frac{U_1 - \bar{d}}{\sqrt{s_{11}}} \right) \left( \frac{U_2 - \bar{d}}{\sqrt{s_{22}}} \right) + \left( 1 - \frac{s_{12}^2}{s_{11} s_{22}} \right) \right]$$

where $s_{ij} = \text{Cov}(U_i, U_j)$ under $H_0$ and $\left( \begin{array}{c} U_1 \\ U_2 \end{array} \right) \sim N \left( \begin{array}{c} E^*U_1 \\ E^*U_2 \end{array} \right) \left( \begin{array}{cc} s_{11} & s_{12}^* \\ s_{12}^* & s_{22}^* \end{array} \right)$

approximately. The only reasonable way to find this power is by simulation.

We can estimate the power of $Q$ for a given $t$ and $n$ by using 100,000 simulations. In particular, start with $Z_1, Z_2$ iid $N(0,1), let$

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} E^*U_1 + \sqrt{s_{11}}Z_1 \\ E^*U_2 + \sqrt{s_{22} - \frac{s_{12}^2}{s_{11}}}Z_2 + \frac{s_{12}Z_1}{\sqrt{s_{11}}} \end{pmatrix}$$

and estimate the power as the
proportion of times that Q is larger than 5.99146. To find the estimated relative efficiency, we find the n that gives the same power for the individual $U_1$ and for LR. It is easier to find the a.r.e. of an individual $U_i$ wrt the likelihood ratio test. The power of the LR with respect to the alternative where \( t = \frac{c}{\sqrt{n}} \) is $\beta_{ln}(n,c) = 1 + \Phi\left( \frac{-1.96\sqrt{3} + |c|}{\sqrt{3}} \right) - \Phi\left( \frac{1.96\sqrt{3} - |c|}{\sqrt{3}} \right) + \frac{|c|\sqrt{3}}{10\sqrt{2\pi}} c^2 e^{\frac{(6(|c|+1.96\sqrt{3})^2)}{6}} \left( 1 - e^{-\frac{(6(|c|+1.96\sqrt{3})^2)}{3}} \right) n^{-\frac{1}{2}} + O(n^{-2})$.

Under the alternative where \( t = \frac{c}{\sqrt{n}} \), the test based on $U_1$ has power

$\beta_i(n,c) = \Phi\left( \frac{-1.96\sqrt{s_{ii}} + \frac{1}{3} - E^*U_i}{\sqrt{s_{ii}}} \right) + 1 - \Phi\left( \frac{1.96\sqrt{s_{ii}} + \frac{1}{3} - E^*U_i}{\sqrt{s_{ii}}} \right)$

$= 3 - \Phi\left( 1.96 - \frac{5c}{4} \right) - \Phi\left( 1.96 + \frac{\sqrt{5}c}{4} \right) + \left\{ \frac{133\sqrt{5}c + 1300(1.96)}{640} \right\} c\phi\left( 1.96 + \frac{\sqrt{5}c}{4} \right)$

$- \frac{133\sqrt{5}c - 1300(1.96)}{640} c\phi\left( 1.96 - \frac{\sqrt{5}c}{4} \right) n^{-\frac{1}{2}} + O(n^{-1})$.

To find the a.r.e., $\rho_1$, of $U_1$ versus LR, first notice that $\rho_1$ satisfies

$\beta_1(n\rho_1, c\sqrt{\rho_1}) - \beta_{ln}(n,c) = 0$. So, using the first order approximations for the powers of the 2 statistics, $\rho_1$ satisfies

$1 - \Phi\left( -1.96 - \frac{|c|}{\sqrt{3}} \right) + \Phi\left( 1.96 - \frac{|c|}{\sqrt{3}} \right) = \Phi\left( 1.96 - \frac{c\sqrt{5}\rho_1}{4} \right) + \Phi\left( 1.96 + \frac{c\sqrt{5}\rho_1}{4} \right)$. So,

clearly the value of $\rho_1$ is $\frac{16}{15}$. But, for $n<173,500$ $U_1$ is better than (i.e. the relative efficiency is greater than 1) LR when $c>0$. The other triangle statistics ($U_2$ and $U_3$) both have the same a.r.e. as $U_1$. The chisquared triangle statistic (Q) is a compromise between the others as the graph in Figure 2.12 shows.

When $t<0$, the distribution is bimodal. Therefore, when both X's are in either of the modes, the X-X side of the triangle is almost always the shortest side.
whereas if the X's are in different modes, the X-X side is almost always the longest. It turns out that on average, there are just the right amount of triangles where the X-X side is the shortest. Hence, $U_2$ has very little power for $t>0$.

![Graph showing power as a function of t]

**Figure 2.12** Power of three versions of the Triangle test and LR (green=Q, yellow=$U_1$, red=$U_2$, blue=LR).

**Example 6. Monte Carlo Study of Tests for Multivariate Normality**

In order to further demonstrate the flexibility and power of the triangle test, in this section we compare the power of the statistic $Q$ to procedures designed specifically to test for multivariate normality. Consider the problem of testing whether the random sample $X_1, X_2, \ldots, X_n$ comes from a k-
dimensional normal distribution with unknown mean vector $\mu$ and
unknown covariance matrix $\Sigma$. First, transform the data into the scaled
residuals $Z_i = S^{-1/2}(X_i - \bar{X})$, where $\bar{X}$ and $S$ are the sample mean vector and
covariance matrix. If $n$ is very large, we can test whether $Z_1, Z_2, ..., Z_n$ come
from a standard multivariate normal distribution using the asymptotic
results of Theorem 2.1. If $n$ is not very large, we may use simulation to find
the null distribution of the triangle statistics. Start with a pseudorandom
sample from a standard multivariate normal distribution and calculate the
statistic. This will give an empirical distribution of the triangle statistic,
calculated from the scaled residuals, under the null hypothesis. The method
is valid because the distribution of $Z_1, Z_2, ..., Z_n$ does not depend on $\mu$ or $\Sigma$.

Table 2.12 gives the results of a small simulation study of the power of
various competing tests for bivariate normality with sample size 50 and
significance level 0.05 (using runun from IMSL). Table 2.13 gives a simulation
study of the power of the best tests for multivariate normality in five
dimensions. The alternatives chosen include all of those studied in Henze
and Zirkler (1990) - the first four columns are our replicates of their table 6.2-
as well as the angular/radial correlated alternatives from table of Quiroz and
Dudley (1991). The table gives results for those tests identified as "best" from
each class as discussed in Henze and Zirkler (1990) and Quiroz and Dudley
(1991). These include Mardia's tests of skewness (MS) and kurtoses (MK), the
Shapiro-Wilk statistic of Fattorini (FA), the empirical characteristic function
test of Henze and Zirkler (HZ which is $T_1$ in their notation), and the second
version of the chi-square test of Quiroz and Dudley (QD). Looking at the
table, we see that Fattorini's test does well for alternatives with independent

67
margins, and is the best choice against the alternatives with angle-radius correlation, while the Quiroz and Dudley statistic appears to have the most power against the spherically symmetric alternatives that were investigated. However, in overall comparisons, the statistic Q, using the max distance for computational simplicity, appears quite robust. It has the best average rank over the 36 alternatives in the table amongst the six tests. It does not have the lowest, or even the second lowest, power for any of the alternatives. Finally, comparing Q to each of the other five tests individually, we see that it beats each test for a majority of the 36 alternatives.
Table 2.12 Monte Carlo Study of Tests for Bivariate Normality.
Percentage of 1000 samples declared significant (α=0.05, n=50)

<table>
<thead>
<tr>
<th>Distributions</th>
<th>MS</th>
<th>MK</th>
<th>FA</th>
<th>HZ</th>
<th>QD</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternatives with independent margins1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential2</td>
<td>100</td>
<td>83</td>
<td>100</td>
<td>100</td>
<td>97</td>
<td>100</td>
</tr>
<tr>
<td>Lognormal2</td>
<td>98</td>
<td>68</td>
<td>98</td>
<td>96</td>
<td>84</td>
<td>92</td>
</tr>
<tr>
<td>Gamma (5,1)2</td>
<td>67</td>
<td>27</td>
<td>62</td>
<td>63</td>
<td>39</td>
<td>47</td>
</tr>
<tr>
<td>Chi-square (5)2</td>
<td>92</td>
<td>46</td>
<td>92</td>
<td>92</td>
<td>63</td>
<td>83</td>
</tr>
<tr>
<td>Chi-square (15)2</td>
<td>51</td>
<td>22</td>
<td>46</td>
<td>48</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>t (2)2</td>
<td>90</td>
<td>96</td>
<td>93</td>
<td>97</td>
<td>98</td>
<td>97</td>
</tr>
<tr>
<td>t (5)2</td>
<td>46</td>
<td>50</td>
<td>37</td>
<td>40</td>
<td>52</td>
<td>49</td>
</tr>
<tr>
<td>Logistic (0,1)2</td>
<td>26</td>
<td>28</td>
<td>18</td>
<td>19</td>
<td>31</td>
<td>23</td>
</tr>
<tr>
<td>Beta (1,1)2</td>
<td>0</td>
<td>91</td>
<td>76</td>
<td>49</td>
<td>84</td>
<td>91</td>
</tr>
<tr>
<td>Beta (1,2)2</td>
<td>26</td>
<td>22</td>
<td>84</td>
<td>77</td>
<td>39</td>
<td>80</td>
</tr>
<tr>
<td>Beta (2,2)2</td>
<td>0</td>
<td>43</td>
<td>13</td>
<td>10</td>
<td>34</td>
<td>37</td>
</tr>
<tr>
<td>N(0,1)⊗Exp</td>
<td>94</td>
<td>53</td>
<td>100</td>
<td>96</td>
<td>70</td>
<td>88</td>
</tr>
<tr>
<td>N(0,1)⊗Chi-square (5)</td>
<td>62</td>
<td>25</td>
<td>74</td>
<td>64</td>
<td>37</td>
<td>47</td>
</tr>
<tr>
<td>N(0,1)⊗t (5)</td>
<td>25</td>
<td>25</td>
<td>20</td>
<td>22</td>
<td>27</td>
<td>22</td>
</tr>
<tr>
<td>N(0,1)⊗Beta (1,1)</td>
<td>2</td>
<td>28</td>
<td>53</td>
<td>19</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td>N(0,1)⊗Beta (1,2)</td>
<td>12</td>
<td>9</td>
<td>62</td>
<td>39</td>
<td>13</td>
<td>37</td>
</tr>
<tr>
<td>mixture alternatives2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMIX (2,0,0)</td>
<td>2</td>
<td>18</td>
<td>18</td>
<td>11</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>NMIX (4,0,0)</td>
<td>2</td>
<td>49</td>
<td>100</td>
<td>95</td>
<td>79</td>
<td>65</td>
</tr>
<tr>
<td>NMIX (2,0,9,0)</td>
<td>82</td>
<td>17</td>
<td>66</td>
<td>84</td>
<td>23</td>
<td>61</td>
</tr>
<tr>
<td>NMIX (0.5,0,9,0)</td>
<td>32</td>
<td>24</td>
<td>29</td>
<td>35</td>
<td>31</td>
<td>47</td>
</tr>
<tr>
<td>NMIX (0.5,0.9,-0.9)</td>
<td>66</td>
<td>48</td>
<td>82</td>
<td>91</td>
<td>60</td>
<td>98</td>
</tr>
<tr>
<td>spherically symmetric alternatives3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSII2 (0)</td>
<td>0</td>
<td>99</td>
<td>52</td>
<td>57</td>
<td>93</td>
<td>83</td>
</tr>
<tr>
<td>PSII2 (1)</td>
<td>0</td>
<td>66</td>
<td>13</td>
<td>13</td>
<td>51</td>
<td>34</td>
</tr>
<tr>
<td>PSVII2 (2)</td>
<td>93</td>
<td>98</td>
<td>96</td>
<td>98</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td>PSVII2 (3)</td>
<td>64</td>
<td>73</td>
<td>57</td>
<td>68</td>
<td>79</td>
<td>77</td>
</tr>
<tr>
<td>PSVII2 (5)</td>
<td>31</td>
<td>35</td>
<td>25</td>
<td>27</td>
<td>36</td>
<td>31</td>
</tr>
<tr>
<td>SPH (Exp)</td>
<td>85</td>
<td>98</td>
<td>96</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>SPH (Gamma (5,1))</td>
<td>4</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>SPH (Beta (1,1))</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>SPH (Beta (1,2))</td>
<td>14</td>
<td>26</td>
<td>24</td>
<td>55</td>
<td>64</td>
<td>65</td>
</tr>
<tr>
<td>SPH (Beta (2,2))</td>
<td>0</td>
<td>38</td>
<td>7</td>
<td>6</td>
<td>22</td>
<td>16</td>
</tr>
</tbody>
</table>
Table 2.12 (continued)

alternatives with angular/radial correlation

<table>
<thead>
<tr>
<th>p</th>
<th>MS</th>
<th>MK</th>
<th>FA</th>
<th>HZ</th>
<th>QD</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>0.4</td>
<td>7</td>
<td>5</td>
<td>14</td>
<td>10</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>0.6</td>
<td>14</td>
<td>6</td>
<td>25</td>
<td>23</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>0.8</td>
<td>28</td>
<td>6</td>
<td>62</td>
<td>40</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>1.0</td>
<td>58</td>
<td>9</td>
<td>98</td>
<td>60</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>

average rank over the 36 alternatives 4.32 4.19 3.24 3.11 3.24 2.90

1The bivariate distribution is denoted D1 ⊗ D2 when the independent marginal distributions are D1 and D2. D^2 = D ⊗ D.

2NMIX (a,b,c) is a 50% mixture of the bivariate normals N((0,0), [1 b]

3PSII and PSVII denote the Pearson type II and type VII distributions. SPH(D) refers to a spherically symmetric distribution where the radius has the distribution D.

4These distributions are given in Quiroz and Dudley (1991) and produce a correlation (increasing in p) between the angular and radial coordinates of the standardized data.
Table 2.13. Monte Carlo Study of Tests for Multivariate Normality in 5 Dimensions. Percentage of 1000 samples declared significant (α=0.05, n=20)

<table>
<thead>
<tr>
<th>Distributions</th>
<th>MS</th>
<th>MK</th>
<th>FA</th>
<th>HZ</th>
<th>OD</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>alternatives with independent margins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp(^5)</td>
<td>82</td>
<td>55</td>
<td>61</td>
<td>65</td>
<td>85</td>
<td>49</td>
</tr>
<tr>
<td>C(0,1)(^5)</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>76</td>
</tr>
<tr>
<td>G(0.5,1)(^5)</td>
<td>98</td>
<td>85</td>
<td>86</td>
<td>90</td>
<td>99</td>
<td>74</td>
</tr>
<tr>
<td>G(5,1)(^5)</td>
<td>20</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>LN(0,0.5)(^5)</td>
<td>63</td>
<td>42</td>
<td>47</td>
<td>49</td>
<td>57</td>
<td>22</td>
</tr>
<tr>
<td>$\chi^2_{5}$</td>
<td>40</td>
<td>21</td>
<td>30</td>
<td>31</td>
<td>37</td>
<td>17</td>
</tr>
<tr>
<td>$\chi^2_{5}$</td>
<td>15</td>
<td>9</td>
<td>14</td>
<td>15</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>t(_{2}) (^5)</td>
<td>86</td>
<td>85</td>
<td>80</td>
<td>81</td>
<td>73</td>
<td>36</td>
</tr>
<tr>
<td>t(_{5}) (^5)</td>
<td>34</td>
<td>24</td>
<td>28</td>
<td>29</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>L(0,1)(^5)</td>
<td>15</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>B(1,1)(^5)</td>
<td>0</td>
<td>28</td>
<td>1</td>
<td>1</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>B(1,2)(^5)</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>B(2,2)(^5)</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>N(0,1)(^4)⊗Exp</td>
<td>19</td>
<td>11</td>
<td>21</td>
<td>21</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>N(0,1)(^4)⊗$\chi^2_{5}$</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>N(0,1)(^4)⊗t(_5)</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

71
### Table 2.13 (continued)

(level $\alpha=0.05$, n=50)

<table>
<thead>
<tr>
<th>Distributions</th>
<th>MS</th>
<th>MK</th>
<th>FA</th>
<th>HZ</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>alternatives with independent margins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LN(0,0.5)$^5$</td>
<td>100</td>
<td>87</td>
<td>90</td>
<td>96</td>
<td>97</td>
</tr>
<tr>
<td>Exp$^5$</td>
<td>100</td>
<td>95</td>
<td>97</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>G(5,1)$^5$</td>
<td>70</td>
<td>35</td>
<td>45</td>
<td>54</td>
<td>100</td>
</tr>
<tr>
<td>$\chi^2_5$</td>
<td>96</td>
<td>61</td>
<td>73</td>
<td>84</td>
<td>86</td>
</tr>
<tr>
<td>$\chi^2_{15}$</td>
<td>49</td>
<td>22</td>
<td>31</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>$t_2^5$</td>
<td>99</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>$t_5^5$</td>
<td>64</td>
<td>68</td>
<td>55</td>
<td>56</td>
<td>72</td>
</tr>
<tr>
<td>B(1,1)$^5$</td>
<td>0</td>
<td>90</td>
<td>0</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>B(1,2)$^5$</td>
<td>10</td>
<td>20</td>
<td>5</td>
<td>14</td>
<td>34</td>
</tr>
<tr>
<td>B(2,2)$^5$</td>
<td>0</td>
<td>47</td>
<td>1</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>L(0,1)$^5$</td>
<td>35</td>
<td>36</td>
<td>27</td>
<td>27</td>
<td>39</td>
</tr>
<tr>
<td>N(0,1)$^4$$\otimes$Exp</td>
<td>62</td>
<td>26</td>
<td>56</td>
<td>72</td>
<td>37</td>
</tr>
<tr>
<td>N(0,1)$^4$$\otimes$$\chi^2_5$</td>
<td>29</td>
<td>13</td>
<td>29</td>
<td>34</td>
<td>23</td>
</tr>
<tr>
<td>N(0,1)$^4$$\otimes$t$_5$</td>
<td>18</td>
<td>15</td>
<td>19</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>N(0,1)$^4$$\otimes$B(1,1)</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>N(0,1)$^4$$\otimes$B(1,2)</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

| **spherically symmetric alternatives** |    |    |    |    |   |
| PSIi5 (0) | 0  | 100| 1  | 70 | 33|
| PSIi5 (1) | 0  | 96 | 1  | 31 | 58|
| PSIi5 (4) | 0  | 42 | 1  | 9  | 85|
| PSVII5 (5) | 88 | 94 | 78 | 77 | 100|
| PSVII5 (10) | 35 | 35 | 27 | 16 | 99|
| SPH ($\Gamma(5,1)$) | 66 | 75 | 48 | 53 | 100|
| SPH (Beta (1,1)) | 65 | 96 | 45 | 100| 100|
| SPH (Beta (1,2)) | 98 | 100| 94 | 100| 100|
| SPH (Beta (2,2)) | 26 | 51 | 12 | 64 | 80|
Example 7. Using the Projection Version of the Statistic as a Measure of Data Depth

The concept of data depth was introduced by Liu (1990). Depth can be described as a center-outward ranking of the data. There are four common measures of depth of a point \(x \in \mathbb{R}^p\) with respect to a probability measure \(F\) defined on \(\mathbb{R}^p\). Liu’s simplicial depth is the probability that \(x\) is a convex combination of \(p+1\) random points from \(F\) [Liu (1990)]. The Mahalanobis depth is the Mahalanobis distance of \(x\) from the center of \(F\) [Mahalanobis (1936)]. Oja’s depth is the average volume of a \(p\)-dimensional simplex formed from \(x\) and \(p\) random points from \(F\) [Oja (1983)]. Finally, Tukey’s depth is defined as \(\inf_{K} \{F(K); K \text{ is a closed half space containing } x\} \) [Tukey (1975)]. All of these except Oja’s depth require an algebraic structure on \(F\). The sample versions of these depths, which are used to define generalized medians, replace \(F\) by the empirical distribution function of the sample. The point which optimizes the sample version of depth can be viewed as a generalized median. As noted by Liu, all measures of depth also induce: i) generalizations of one-dimensional statistics based on the ordered sample; ii) multivariate measures of scale, skewness, and kurtosis; and iii) multivariate classification rules.

The so-called projection version of

\[
U_i = \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P[|Y - X_i| > |X_i - X_j|, |Y - X_i| > |X_i - X_j|] \text{ is}
\]

\[
U_i = \frac{1}{n} \sum_{i=1}^{n} P[|Y - X_i| > |X_i - Y'|, |Y - Y'| > |X_i - Y'|] \text{ where } Y \text{ and } Y' \text{ are independent}
\]
with the hypothesized distribution. It is often useful to examine the projection versions of U-statistics to derive results about their asymptotic behavior. In this case, we can interpret the kernel function of the projection version of one of the triangle statistics,

\[ h(X) = \mathbb{P}[|Y - Y'| > |X - Y|, |Y - Y'| > |X - Y'|], \]

as a measure of data depth. In fact, two new measures of depth are suggested by the idea of examining properties of random triangles. To be more precise, if \( x \) is a point and \( F \) is any distribution for which a distance between points can be defined, consider a triangle formed by \( x \) and two random points from \( F \). The first proposed measure of depth, \( D_1(x) \), is the probability that the leg joining the two random points is the longest leg in this triangle. The second measure of depth, \( D_2(x) \), is the probability that the angle between the legs joining \( x \) to the random points is the largest angle in the triangle. Notice that no algebraic structure on \( F \) is required. An example of a situation where this is useful is when the sample consists of trees. A method for defining distances between trees has been proposed by Critchlow, Pearl, and Qian (1996).

First, we will show that both proposed measures of depth defined by any norm \( \| \cdot \| \) on \( \mathbb{R}^p \) satisfy the first two of the properties of the measure of depth for \( p \)-dimensional vectors defined by Liu. The proofs of these theorems, which apply to both measures of depth, are supplied only for \( D_1 \) with the understanding that the proofs for \( D_2 \) follow similarly with the application of the Law of Cosines.
THEOREM 2.2. \( \sup_{x \in \mathcal{R}^p} D_i(x) \to 0 \) as \( M \to \infty \) for any \( F \) defined on \( \mathcal{R}^p \) where the supremum is taken over \( x \in \mathcal{R}^p \).

PROOF. Let \( X_1 \) and \( X_2 \) be independent random vectors with distribution \( F \). Notice that \( D_i(x) \leq P[X_1 - X_2 > X_1 - x, X_2 > x] \leq P[2\|X_1\| + \|X_2\| > M] \to 0. \)

THEOREM 2.3. If \( F \) is absolutely continuous on \( \mathcal{R}^p \), then \( D_i \) is continuous.

PROOF. Let \( X_1 \) and \( X_2 \) be independent random vectors with distribution \( F \) and let \( x \) and \( y \) be two fixed points in \( \mathcal{R}^p \). Now, \( \|D_i(x) - D_i(y)\| \)

\[
\leq 2P[\|X_1 - X_2\| > \|X_1 - x\|, \|X_1 - X_2\| > \|X_2 - x\|, \|X_1 - x\| < \|X_2 - y\|] \\
+ P[\|X_1 - X_2\| > \|X_1 - x\|, \|X_1 - X_2\| > \|X_2 - x\|, \|X_1 - x\| < \|X_1 - y\|, \|X_1 - X_2\| < \|X_2 - y\|] \\
+ 2P[\|X_1 - X_2\| > \|X_1 - x\|, \|X_1 - x\| < \|X_2 - x\|, \|X_1 - x\| > \|X_1 - y\|, \|X_1 - X_2\| > \|X_2 - y\|] \\
+ P[\|X_1 - X_2\| < \|X_1 - x\|, \|X_1 - x\| < \|X_2 - x\|, \|X_1 - x\| > \|X_1 - y\|, \|X_1 - X_2\| > \|X_2 - y\|] \\
\leq 3P[\|X_2 - y\| > \|X_1 - x\|, \|X_1 - x\| > \|X_2 - x\|] + 3P[\|X_2 - x\| > \|X_1 - x\|, \|X_1 - x\| > \|X_2 - y\|].
\]
The absolute continuity of \( F \) implies that both of these terms vanish as \( \|x - y\| \to 0 \) by conditioning on \( X_2 \).

The last property of Liu's depth only applies to distributions \( F \) which are angularly symmetric about a point \( b \). This means that if \( X \sim F \), then \( (X-b)/\|X-b\| \) has the same distribution as \( -(X-b)/\|X-b\| \) where \( \|\cdot\| \) is the Euclidean norm. It is unknown whether Oja's depth has this monotonicity property. However, Tukey's depth and the Mahalanobis depth do have this property.

Figure 2.13 shows a scatterplot for the depth of points of the form
(x, 0) with respect to the distribution of the bivariate random variable $Y'$ defined as follows: let $Y$ have a standard bivariate normal distribution, then define $Y' = \begin{cases} 16Y / \|Y\|^2 & \text{if } \|Y\| < 4 \text{ and } Y_1 > 0 \\ Y & \text{otherwise.} \end{cases}$

The depth was found by simulation using 10 replications and one million pairs of simulated points from the hypothesized distribution for each point. The Euclidean distance was used to define $D_1$ here. Not surprisingly, $D_2$ is almost perfectly correlated with $D_1$ when the Euclidean distance is used to define $D_1$. One immediately sees from Figure 2.13 that the maximum depth for each odf the ten replications does not occur at the center of angular symmetry (x=0). Thus, triangular depth does not have the monotonicity property that Liu's depth has for angularly symmetric distributions.

**Figure 2.13** Depth of points along the x-axis for an angularly symmetric distribution.
Chapter III

Two-sample Tests Based on Interpoint Distances

Introduction

Given two independent random samples of multivariate random vectors $X_1, X_2, ..., X_m \sim F$ and $Y_1, Y_2, ..., Y_n \sim G$, the two sample problem is to test the hypothesis $H_0: F = G$ versus the general alternative $F \neq G$.

When the data are categorical, chi-square tests for multinomial distributions provide a well developed framework [see, for example Bishop, Fienberg, and Holland (1975)]. The classical parametric test using Hotelling's $T^2$ statistic is appropriate when the two samples are assumed to be observed from multivariate normal populations with the same covariance matrix. The robustness of this test against departures from the equal covariance assumption was studied in Ito and Schull (1964) and robustness with respect to departures from the normality assumption was studied in Mardia (1975) and Everitt (1979). Nonparametric procedures for comparing the locations of two bivariate distributions can be found in Mardia (1967) and Fryer (1970) and
a nonparametric procedure for comparing the locations of two multivariate distributions is presented in Randles and Peters (1990).

Distributional results fail rapidly for the above techniques when the dimension of the raw data grows relative to the size of the sample. More success in the high-dimensional case is achieved by methods which rely on the interpoint distance relationships of the sample data. One such nonparametric test with broad-based application is discussed by Friedman and Rafsky (1981) who use minimal spanning trees to generalize the one-dimensional Wald-Wolfowitz runs test as well as the Kolmogorov-Smirnov test. Finally, a nonparametric test which is consistent against all alternatives was studied by Schilling (1986) and Henze (1988). This test statistic uses the proportion of nearest neighbors which are from the same sample. Tests based on nearest neighbors are intuitively appealing because they measure whether members from one sample are as close to members of the second sample as they are to themselves.

The tests proposed here use information about all of the interpoint distances, not just information from nearest neighbors. The motivation is in the theorem given in Maa et. al. (1996) which states that F=G if and only if H_{XX}=H_{YY}=H_{XY} where H_{XY} is the distribution function of the distance between a random variable with distribution F and an independent random variable with distribution G and H_{XX} and H_{YY} are defined similarly. For simplicity, we will refer to the objects which comprise the two samples as random vectors or random variables, but the theorem in Maa et. al. and the results here apply to samples of random objects of any type so long as a distance can be defined. For example, the objects themselves could be graphs or continuous curves
measured from some physical process. Also, the word "distance" will be used
even when the assumptions require only a non-negative bivariate function
\( \rho(\cdot, \cdot) \) defined for pairs of observations in the union of the supports of \( F \) and \( G \)
which is 0 iff the two arguments are identical.

Define the empirical counterpart of \( H_{XY} \), the cross-sample interpoint
distance distribution, by 
\[ \hat{H}_{XY}(t) = (mn)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{[0,t]}(\rho(X_i, Y_j)) \]
where the indicator function \( I_{S}(x) \) is defined to be 1 if \( x \in S \) and 0 otherwise. Also, define the
empirical counterparts of the two within-sample interpoint distance
distribution functions by 
\[ \hat{H}_{XX}(t) = \left( \frac{m}{2} \right)^{-1} \sum_{i=1}^{m} I_{[0,t]}(\rho(X_i, X_i)) \]
and 
\[ \hat{H}_{YY}(t) = \left( \frac{n}{2} \right)^{-1} \sum_{i=1}^{n} I_{[0,t]}(\rho(Y_i, Y_i)). \]
General results regarding the convergence of U-processes will be applied. Therefore, since \( \hat{H}_{XX}(t) \) and \( \hat{H}_{YY}(t) \) are U-
processes, but \( \hat{H}_{XY}(t) \) is not, we will need to define the related empirical process
\[ \hat{H}(t) = \left( \frac{m+n}{2} \right)^{-1} \left\{ \left( \frac{m}{2} \right) \hat{H}_{XX}(t) + \left( \frac{n}{2} \right) \hat{H}_{YY}(t) + mn \hat{H}_{XY}(t) \right\} \]
\[ = \left( \frac{m+n}{2} \right)^{-1} \sum_{i=1}^{m+n} I_{[0,t]}(\rho(W_i, W_j)) \]
where \( W_1, W_2, ..., W_{m+n} \) is the pooled sample. This is the empirical
counterpart of the pooled-sample interpoint distance distribution function,
\( H(t) \), and it is a U-process under the null hypothesis.

Two types of test statistics are examined. One type is based on an
attempt to measure the difference between the three empirical distance
distribution functions over their entire support. A second type of test statistic
is based on measuring the difference between the three empirical distribution

79
functions at a fixed number of points. In Section 3.1, we show that the first type of statistic is consistent versus all alternatives, but only the conditional distribution of these statistics is accessible. The second type of statistic will be introduced in Section 3.2. The statistics of this type are consistent against a wide variety of alternatives and their limiting nominal significance can be computed. Finding the limiting distribution of these statistics requires a theorem on the distribution of a vector of weighted symmetric statistics which have the form of weighted U-statistics. Since the distribution of this vector is needed under both the null and alternative hypotheses, the result must not require the identical distribution of the random variables on which the statistics are defined. The actual choice between the two types of statistics in an application will depend on whether it is desired to have i) consistency against all alternatives, ii) adaptability to a specific alternative, or iii) simplicity of application. In section 3.3, some example are given to illustrate i) the power of both types of statistics across a variety of departures from the null hypothesis, ii) the effect of using different distances, and iii) the choice of different evaluation points on the power of the unconditional test discussed in 3.2.

3.1. A Class of Consistent Conditional Test Statistics

In this section, we propose a class of statistics based on the difference between the three empirical distribution functions over their entire support and prove that each statistic of this type is consistent against all alternatives. The theorem in Maa et. al. tells us that the equality of $F$ and $G$ is equivalent to the equality of the three underlying interpoint distance distributions.
Halmos (1946) shows that if $F$ is known to lie within a suitably large class then, under the null hypothesis, $\hat{H}_{XX}(t), \hat{H}_{YY}(t),$ and $\hat{H}(t)$ have the smallest variance among all unbiased estimates of $H_{XX}(t), H_{YY}(t),$ and $H(t)$ for all $t$. Since $\hat{H}_{XX}(t), \hat{H}_{YY}(t),$ and $\hat{H}(t)$ are unbiased estimates of their population counterparts, they have the same expectation only under the null hypothesis. To construct a test, we need to form a statistic which gives a useful measure of how the observed empirical distribution functions differ from what is expected under the null hypothesis and we need a method of evaluating the significance of the observed value of the chosen statistic.

We measure the difference between the three empirical distribution functions at a fixed $t$ using

$$\ell(t) = \max\left\{H_{XX}(t) - \hat{H}_{XY}(t), H_{YY}(t) - \hat{H}_{YY}(t) \right\}/2.\right\}$$

The $1/2$ which multiplies the middle term in the definition of $\ell(t)$ is included so that each of the three components have (asymptotically) the same covariance at each $t$. Now, let $\mu$ be a probability measure defined on $[0, \infty)$ with support containing the support of $H$ and let $p$ be any positive real number. Define the statistics

$$Q_{p, \mu} = \left\{\int_0^\infty \ell(t)^p d\mu(t) - E_\mu\left\{\int_0^\infty \ell(t)^p d\mu(t)\right\}\right\}/\sqrt{\text{Var}_\mu\left\{\int_0^\infty \ell(t)^p d\mu(t)\right\}}$$

and

$$Q_* = \left\{\sup_t \ell(t) - E_\mu\left\{\sup_t \ell(t)\right\}\right\}/\sqrt{\text{Var}_\mu\left\{\sup_t \ell(t)\right\}}.$$

Intuitively, $Q$ is a measure of the size of the difference between the empirical interpoint distance curves which has been standardized under the null hypothesis (see Figure 3.1). This defines a family of test statistics for which the null hypothesis is rejected for large values of $Q$. The consistency of $Q$ for finite $p$ against all alternatives is shown by Theorem 3.1 below. Here, and
wherever limiting behavior is examined in this section, we will assume that the pooled sample size, \( N \), approaches infinity and that the individual sample sizes are functions of \( N \) which also approach infinity such that 
\[
\lim_{N \to \infty} m(N)/N = c
\]

**Figure 3.1** Q measures the size of the difference between the three empirical curves over the entire support. If \( p=1 \) and \( \mu \) is a uniform measure over a finite range, then \( Q_{p,\mu} \) is the shaded area shown here.

**THEOREM 3.1.** If \( \mu \) is a probability measure defined on \([0, \infty)\) whose support contains the support of \( H \), then \( Q_{p,\mu} \) is consistent against all alternatives for any \( p \in (0, \infty] \).

**PROOF.** The proof is given for finite \( p \) only. The \( p=\infty \) case follows by a similar argument. General results on the convergence of empirical processes [Silverman (1983); Schneemeier (1993)] show that 
\[
\sup_t \left| \hat{H}_{xx}(t) - H_{xx}(t) \right|
\]

82
\[ \sup_i \left| \hat{H}_{YY}(t) - H_{YY}(t) \right|, \text{ and } \sup_i \left| \hat{H}(t) - H(t) \right| \] converge almost surely to 0 under both the null and alternative hypothesis. This implies that \( \sup_i \left| \ell(t) - L(t) \right| \to 0 \) almost surely where

\[ L(t) = \max \left\{ \left| H_{XX}(t) - H_{XY}(t) \right|, \left| H_{XX}(t) - H_{YY}(t) \right|/2, \left| H_{XY}(t) - H_{YY}(t) \right| \right\}. \]

Moreover, \( \int_0^\infty \left\{ \ell(t) \right\}^p d\mu(t) - \int_0^\infty \left\{ L(t) \right\}^p d\mu(t) \leq \sup_i \left[ \left( \ell(t) \right)^p - \left( L(t) \right)^p \right] \), which implies that \( \int_0^\infty \left\{ \ell(t) \right\}^p d\mu(t) \to \int_0^\infty \left\{ L(t) \right\}^p d\mu(t) \) almost surely. Also, since \( \int_0^\infty \left\{ \ell(t) \right\}^p d\mu(t) \leq 1 \), the bounded convergence theorem implies that \( \text{Var}_\mu \left\{ \int_0^\infty \left\{ \ell(t) \right\}^p d\mu(t) \right\} \)

converges to 0 almost surely. The theorem follows from the fact that \( \int_0^\infty \left\{ L(t) \right\}^p d\mu(t) \) is 0 under the null hypothesis and is strictly positive otherwise.

The conclusions of the theorem also hold if \( \mu \) is replaced by an empirical measure \( \hat{\mu} \) derived from the data, provided \( \hat{\mu} \) converges in the sup-norm metric to some fixed \( \mu \) almost surely.

Unfortunately, the distribution of \( Q_{p,\mu} \) cannot be found unconditionally and it is therefore necessary to use either permutation or bootstrap methods to evaluate the significance of an observed statistic.

### 3.2. A Class of Unconditional Test Statistics

Again motivated by the theorem in Maas et. al., we propose a class of tests based on the processes

\[ Z_1(t) = \hat{H}_{XX}(t) - \hat{H}_{YY}(t) \text{ and} \]

\[ Z_2(t) = \hat{H}_{XX}(t) + \hat{H}_{YY}(t) - 2\hat{H}_{XY}(t) \]
\[
= \frac{n+m-1}{n} \hat{H}_{XX}(t) + \frac{n+m-1}{m} \hat{H}_{YY}(t) - \frac{(m+n)(m+n-1)}{mn} \hat{H}(t).
\]

By evaluating these processes at only a finite number of points, we lose the property of consistency, but we gain the ability to find the limiting distributions of the statistics without using conditional methods. We will first describe the heuristic motivation for using these processes. Each of the empirical distribution functions converges to a Gaussian process with unknown mean. But, the mean of any contrast in \( \hat{H}_{XX}(t) \), \( \hat{H}_{YY}(t) \), and \( \hat{H}(t) \) vanishes under the null hypothesis. Among all possible contrasts, it is desirable to find one which has the smallest variance (at each \( t \)) so that any deviation from the null hypothesis can be more easily detected. The contrast which defines \( Z_2 \) has a variance with quadratic order in the sample sizes while all other contrasts have a variance with a linear order in the sample sizes (at each \( t \)). The contrast which defines \( Z_1 \) is orthogonal to \( Z_2 \) and it therefore intuitively contains the information which is not contained in \( Z_2 \).

To find the distribution of our statistic based on \( Z_1 \) and \( Z_2 \), we first show that a vector which consists of the values of the two processes evaluated at any finite collection of points is asymptotically multivariate normal. Hoéfnding (1948) provided sufficient conditions for the asymptotic normality of a vector \( \left( \sum_{1 \leq i < j \leq n} h_i(x_i,x_j), \ldots, \sum_{1 \leq i < j \leq n} h_r(x_i,x_j) \right)^T \) where \( X_1, \ldots, X_n \) are assumed to be independent. Gregory (1977) found the limiting distribution of a U-statistic in the degenerate case. Shapiro and Hubert (1979) provided conditions so that a weighted U-statistic, \( \sum_{1 \leq i < j \leq n} \omega(i,j) h(X_i,X_j) \) has an asymptotic normal distribution and O'Neil and Redner (1993) discussed the limiting distribution.
of weighted U-statistics in the degenerate case where non-normal limits can occur. The following theorem gives conditions for the asymptotic normality of a vector of statistics which have the form of weighted U-statistics, but where the underlying sample is only assumed to be independent and not necessarily identically distributed.

**THEOREM 3.2.** Let $W_i \sim F_{iN}$; $i=1, 2, \ldots, N$ be independent random variables and let $U_{kN} = \sum_{1 \leq i < j \leq N} \omega_{kN}(i,j)h_k(W_{iN}, W_{jN})$; $k=1, \ldots, R$ where the functions $\omega_{kN}$ and $h_k$ are symmetric in their arguments. The vector $U_N=(U_{1N}, U_{2N}, \ldots, U_{RN})^T$ has the same asymptotic distribution as

$$V_N = \left( \sum_{i=1}^N \sum_{j=1}^N \omega_{IN}(i,j)h_i(W_{iN}, u) dF_{iN}(u), \ldots, \sum_{i=1}^N \sum_{j=1}^N \omega_{RN}(i,j)h_R(W_{iN}, u) dF_{iN}(u) \right)^T$$

provided that for every non-zero vector $\alpha=(\alpha_1, \alpha_2, \ldots, \alpha_R)^T$ the following two ratios converge to 0 as $N$ tends to infinity:

$$A_N = \frac{\sum_{i=1}^N \sum_{j=1}^N \text{Cov} \left\{ \sum_{k=1}^R \alpha_k \omega_{kN}(i,j)h_k(W_{iN}, W_{jN}), \sum_{t=1}^R \alpha_t \omega_{tN}(i,j)h_t(W_{iN}, W_{jN}) \right\}}{\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^R \text{Cov} \left\{ \sum_{t=1}^R \alpha_t \omega_{tN}(i,j)h_t(W_{iN}, W_{jN}), \sum_{k=1}^R \alpha_k \omega_{kN}(i,j)h_k(W_{iN}, W_{jN}) \right\}}$$

where the random vectors $W_{jn} \sim F_{jn}$ are independent of $W_{iN}$, and

$$B_N = \frac{\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^R \text{Var} \left\{ \sum_{t=1}^R \alpha_t \omega_{tN}(i,j)h_t(W_{iN}, W_{jN}) \right\}}{\sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^R \sum_{t=1}^R \text{Cov} \left\{ \sum_{t=1}^R \alpha_t \omega_{tN}(i,j)h_t(W_{iN}, W_{jN}), \sum_{k=1}^R \alpha_k \omega_{kN}(i,k)h_k(W_{iN}, W_{kN}) \right\}}$$

**PROOF.** Straightforward calculations show that $\text{Cov}(\alpha^T U_N, \alpha^T V_N) = A_N + C_N$, $\text{Var}(\alpha^T U_N) = B_N + C_N$, and $\text{Var}(\alpha^T V_N) = A_N + C_N$. Thus,

$$\frac{\text{Cov}(\alpha^T U_N, \alpha^T V_N)}{(\text{Var}(\alpha^T U_N) \cdot \text{Var}(\alpha^T V_N))} \to 1.$$
This implies that $\alpha^T U_N$ has the same asymptotic distribution as $\alpha^T V_N$. Since this holds for all $\alpha$, a theorem of Cramér and Wold (1936) implies that $V_N$ has the same asymptotic distribution as $U_N$.

Theorem 3.2 provides an easy way of checking if $U_N$ has the same asymptotic distribution as $V_N$. Since $\alpha^T V_N$ is the sum of $N$ independent random variables, it is relatively easy to determine if $V_N$ is asymptotically normal by applying Cramér and Wold's theorem again. For example, Corollary 3.1 below uses only Lyapunov's theorem [see Billingsley (1986)] to show that $\alpha^T V_N$ is asymptotically normal for a useful special case.

An important special case of Theorem 3.2 is when the sample can be partitioned into subsamples such that variables within a subsample have the same distribution and each $U_{kN}$ is a U-statistic defined on a union of subsamples. We will make use of this case when the entire sample is partitioned into two subsamples. Sufficient conditions for the asymptotic multivariate normality of a vector of U-statistics defined in this two-sample case are given in Corollary 3.1.

Before the corollary can be stated, we will need to introduce some notation. Let $I_a$ be the $a \times a$ identity matrix and let $0_a$ be an $a \times a$ matrix of zeroes. Given positive integers $r(1)$, $r(2)$, and $r(3)$ which sum to $R$, define the block diagonal matrices $D_1 = \text{diag}(I_{r(1)}, 0_{r(2)}, I_{r(3)})$, $D_2 = \text{diag}(0_{r(1)}, 0_{r(2)}, I_{r(3)})$, and $D_3 = \text{diag}(0_{r(1)}, I_{r(2)}, I_{r(3)})$. Recall that in the two-sample problem, we assume that the individual sample sizes, $m(N)$ and $n(N)$, approach infinity and

$$\lim_{N \to \infty} m(N)/N = c.$$
COROLLARY 3.1. Let \( W_1, ..., W_{m(N)} \sim F \) and \( W_{m(N)+1}, ..., W_N \sim G \) be independent random samples and let \( U_{1N}, ..., U_{RN} \) be weighted U-statistics which are defined as follows for \( k = 1, 2, ..., R \):

\[
    U_{k,N} = \sum_{i \leq j} \omega_{k,N}(i,j) h_k(W_{i,N}, W_{j,N})
\]

where

\[
    \omega_{k,N}(i,j) = \begin{cases} 
        I_{\{1,2, ..., m(N)\}}(i)I_{\{1,2, ..., m(N)\}}(j) & \text{if } k \leq r(1) \\
        I_{\{m(N)+1, m(N)+2, ..., N\}}(i)I_{\{m(N)+1, m(N)+2, ..., N\}}(j) & \text{if } r(1) < k \leq r(2) \\
        1 & \text{otherwise.}
    \end{cases}
\]

The vector of these U-statistics is asymptotically multivariate normal if

\[
    M = c \text{Var}\left\{ c D_1 v_F(X_1) + (1-c) D_2 v_G(X_1) \right\} + (1-c) \text{Var}\left\{ c D_2 v_F(Y_1) + (1-c) D_3 v_G(Y_1) \right\}
\]

is positive definite where \( v_\mu(W) = (\int h_1(W,u)du, ..., \int h_8(W,u)du) \).

PROOF We proceed by showing that both conditions in Theorem 3.1 are satisfied with \( F_{i,N} \).

To do this, notice that for all \( i, N, \delta, \) and non-zero \( \alpha \),

\[
    \left| \sum_{k=1}^{r} \alpha_k \sum_{j=1}^{N} \omega_{k,N}(i,j) \left( h_k(w,u) - \int h_k(v,u)du \right) \right| \leq \sum_{k \in \{1, r_1, ..., r_t \}} \alpha_k \left| \left( h_k(w,u) - \int h_k(v,u)du \right) \right| \\
    \leq N \max \left\{ \sum_{k \in \{1, r_1, ..., r_t \}} \alpha_k \left| \left( h_k(w,u) - \int h_k(v,u)du \right) \right| 
\]

Also, by expanding and collecting like terms, it may be verified that
$N^{-3}\text{Var}\{\alpha^T V_N\} \rightarrow \alpha^T M \alpha$. Therefore, Lyapunov's theorem implies that $\alpha^T V_N$ is asymptotically normal for every non-zero $\alpha$. Since the denominators of the two ratios are of order $N^3$ and the numerators are of order $N^2$ here, the vector of U-statistics is asymptotically normal.

We define a class of statistics based on the evaluation of the processes $Z_1$ and $Z_2$ at a fixed set of points. So, our application of Corollary 3.1 will require the specification of distinct points $t_1, t_2, \ldots, t_r$ in the case when $r(1)=r(2)=r(3)=r$. Corollary 3.1 is then applied to the vector of U-statistics defined by $h_i(X, Y) = h_{r,i}(X, Y) = h_{s,i}(X, Y) = I_{[0, x_i]}(\rho(X, Y)); i=1, 2, \ldots, r$.

Under the null hypothesis $F=G$, so that the matrix $M$ is always positive definite if $0<c<1$. Under an alternative, $M$ depends on both $F$ and $G$. However, we will only be using the corollary for contiguous alternatives where asymptotic normality is attained. This implies that the vector $Z=(Z_1(t_1), Z_1(t_2), \ldots, Z_1(t_r), Z_2(t_1), Z_2(t_2), \ldots, Z_2(t_r))$ is asymptotically multivariate normal since the components of $Z$ are linear combinations of the components of a vector which is asymptotically multivariate normal.

The elements of the covariance matrix of $Z$ can be found from the covariance matrix of the vector of empirical interpoint distance distributions evaluated at the same points. The elements of this covariance matrix are defined by:

$$\text{Cov}\left(\hat{H}_{XX}(s), \hat{H}_{XX}(t)\right) = \begin{pmatrix} m \end{pmatrix}^{-1} \text{Cov}\left(I_{[0,s]}(\rho(X_1, X_2)), I_{[0,t]}(\rho(X_1, X_2))\right)$$

$$+ 2(m-2) \begin{pmatrix} m \end{pmatrix}^{-1} \text{Cov}\left(I_{[0,s]}(\rho(X_1, X_2)), I_{[0,t]}(\rho(X_1, X_3))\right),$$

$$\text{Cov}\left(\hat{H}_{XX}(s), \hat{H}_{XY}(t)\right) = 2m^{-1} \text{Cov}\left(I_{[0,s]}(\rho(X_1, X_2)), I_{[0,t]}(\rho(X_1, Y_1))\right),$$

88
\[
\text{Cov}\left(\hat{H}_{XY}(s), \hat{H}_{XY}(t)\right) = (mn)^{-1}\text{Cov}\left(I_{[0,s]}(\rho(X_t,Y_t)), I_{[0,t]}(\rho(X_t,Y_t))\right)
\]

\[
+ (n-1)(mn)^{-1}\text{Cov}\left(I_{[0,s]}(\rho(X_1,Y_1)), I_{[0,t]}(\rho(X_1,Y_2))\right)
\]

\[
+ (m-1)(mn)^{-1}\text{Cov}\left(I_{[0,s]}(\rho(X_2,Y_1)), I_{[0,t]}(\rho(X_2,Y_1))\right),
\]

and the remaining non-zero elements of the covariance matrix can be derived from the three listed by interchanging $X$'s and $Y$'s. Under the null hypothesis, the mean of $Z$ is the zero vector and the parameters that determine the covariance matrix are all identical, e.g.

\[
\text{Cov}\left(I_{[0,s]}(\rho(X_1,X_2)), I_{[0,t]}(\rho(X_1,X_1))\right) = \text{Cov}\left(I_{[0,s]}(\rho(X_1,X_2)), I_{[0,t]}(\rho(X_1,Y_1))\right).
\]

There are two different strategies for choosing the points $t_1, t_2, ..., t_r$ at which to evaluate the empirical processes. The first method is to choose the points before the data is collected. In this case, we recommend using the points $t_i = \hat{H}^{-1}(i/(r+1))$; $i=1, 2, ..., r$ and the choice of $r$ would depend on the sample sizes. The second method would be to use part of the data to try to estimate the optimal point or points. Fortunately, if $T_1, T_2, ..., T_r$ are statistics which are consistent estimators of $t_1, t_2, ..., t_r$, then the vector $Z$ evaluated at the points $T_1, T_2, ..., T_r$ remains asymptotically normal.

Both of these recommendations for choosing the evaluation points result in consistent tests. The first method results in a consistent test provided the points asymptotically cover the support of $H$, i.e. for all $\varepsilon>0$ and $t$ in the support of $H$, there is an $N_0$ such that for all $N>N_0$ there is a $t_i$ with $|t-t_i|<\varepsilon$. The recommended sequence of evaluation points satisfies this condition. The second method results in a consistent test provided the estimated evaluation points converge to fixed points where there is a difference in the theoretical distribution functions. With either method, it is also possible to evaluate the two processes at different sets of points. For
example, it is possible to use the vector
\[ Z = \left( Z_1(\hat{H}^{-1}(1/2)), Z_2(\hat{H}^{-1}(1/4)), Z_2(\hat{H}^{-1}(3/4)) \right)^T. \]

Since \( Z \) is asymptotically normal, it is straightforward to find a one-dimensional statistic for testing whether \( Z \) has mean 0. In particular, rejecting for large values of the statistic
\[
S = Z^T \hat{\Sigma}_Z^{-1} Z
\]
where \( \hat{\Sigma}_Z \) is the estimated covariance matrix of \( Z \), is the uniformly most powerful invariant test for the hypothesis \( EZ = (0, 0, ..., 0)^T \) versus the general alternative. Most importantly, there is a small number of parameters which define \( \Sigma_Z \) and this number is independent of the dimension of the data. For example, there are only two unknown parameters in the covariance matrix of \( (Z_1(t), Z_2(t))^T \) and we will see in Section 4 that the test based on the statistic defined in (3.1) with this vector performs very well.

### 3.3. Monte Carlo Study

Friedman and Rafsky (1981) and Schilling (1986) both provided Monte Carlo investigations of their tests when the two samples are multivariate normal with the second sample differing either in location or scale from the first. We have supplemented those Monte Carlo studies with additional alternatives in making comparisons with the tests described in Sections 3.2 and 3.3. The alternatives are chosen to represent broad classes of departures from the null hypothesis.

In Table 3.1, the first sample is drawn from a multivariate normal distribution with dimension \( d \), mean \((0, 0, ..., 0)^T \) and covariance matrix \( I_d \).
The second sample is drawn from a population with a different distribution for each row in the table. These distributions are defined as follows:

i) Standard normal. The same distribution as the first population.

ii) Location shift. Multivariate normal distribution with dimension $d$, mean $(\Delta_d, 0, ..., 0)^T$ and covariance matrix $I_d$ where $\Delta_2=0.5$, $\Delta_5=0.75$, and $\Delta_{10}=1.0$.

iii) Difference in scale. Multivariate normal distribution with dimension $d$, mean $(0, 0, ..., 0)^T$ and covariance matrix $\sigma_d^2 I_d$ where $\sigma_2=1.2$, $\sigma_5=1.2$, and $\sigma_{10}=1.1$.

iv) Mixture. All coordinates are independent. The first coordinate has distribution $0.5\Phi\left((1 + \Delta_d^2)^{-\frac{1}{2}} (y - \Delta_d)\right) + 0.5\Phi\left((1 + \Delta_d^2)^{-\frac{1}{2}} (y + \Delta_d)\right)$ where $\Phi$ is the distribution function of a standard normal random variable and $\Delta_2=1$, $\Delta_5=1$, and $\Delta_{10}=2$. The remaining $d-1$ coordinates are standard normal random variables.

v) Correlated normals. Multivariate normal distribution with dimension $d$, mean $(0, 0, ..., 0)^T$ and covariance matrix with the $(i, j)^{th}$ entry given by $\rho_{d^{i-j}}$ where $\rho_2=0.5$, $\rho_5=0.3$, and $\rho_{10}=0.2$.

vi) Tukey. Independent coordinates, each with a Tukey(5.2) distribution [see, for example, Joiner and Rosenblatt (1971)], rescaled to have the same first four moments as a standard normal random variable.

In Table 3.2, the first sample is drawn from a distribution which is uniform on the $d$-dimensional hypercube. The distributions from which the second population is drawn are defined as follows:

i) Uniform. The same distribution as the first population.

ii) Beta. Independent coordinates, each having a Beta($\alpha_d$, 1) where $\alpha_2=0.5$, $\alpha_5=0.7$, and $\alpha_{10}=0.8$. 

91
iii) Correlated uniforms. Let $\xi_1$ and $\xi_2$ be iid Unif(0, 1). Conditioned on $\xi_1$ and $\xi_2$, the coordinates $Y_1, \ldots, Y_{k(d)}$, $Y_{k(d)+1}, \ldots, Y_d$ are independent with distributions: $Y_j \sim$ Unif(0,1) if $j > k(d)$, $Y_j \sim$ Unif(0, $\xi_1$) if $j \leq k(d)$ and $\xi_2 < \xi_1$, and $Y_j \sim$ Unif($\xi_1$, 1) if $j \leq k(d)$ and $\xi_1 < \xi_2$ where $k(2)=2$, $k(5)=3$, and $k(10)=5$.

iv) Checkerboard. The d-dimensional hypercube is partitioned into $k(d)^d$ equally spaced congruent hypercubes, where $k(2) = 3$, $k(5) = 5$, and $k(10) = 10$. The sample distribution is uniform on alternating hypercubes, with no probability in between.

v) Neyman clustering. Fifty clusters of average size two are chosen where the location of each cluster is uniform on the d-dimensional hypercube (see Cuzak and Edwards, 1990).

To estimate the significance of the consistent statistic $Q_{2,1}$ defined in Section 2, we compare the observed value with a random sample from the permutation distribution. In particular, a sample of $m = 100$ values is taken without replacement from the pooled sample $W_1, \ldots, W_{200}$. These values are given the "X" label and the statistic is calculated. This process is then repeated 1000 times giving 1000 observations of the test statistic under the null permutation distribution.

We also investigate the power of the unconditional statistic $S_{1,1}$, $S_{1,2}$, and $S_{1,1}^{'}$, defined by (3.1). $S_{1,1}$ and $S_{1,2}$ use Euclidean distances with $Z = \left[Z_1\left(\hat{H}^{-1}(1/2)\right), Z_2\left(\hat{H}^{-1}(1/2)\right)\right]^T$ and $Z = \left[Z_1\left(\hat{H}^{-1}(1/2)\right), Z_2\left(\hat{H}^{-1}(1/4)\right), Z_2\left(\hat{H}^{-1}(3/4)\right)\right]^T$ respectively. $S_{1,1}^{'}$ uses a "distance" which combines the Euclidean distance between two points with the angle which is formed from the vectors connecting the sample mean to the two points. The final distance is the sum of the squares of the two components after rescaling each component by the
sample standard deviation. For all three statistics, the estimated covariance matrix of \( Z \) is calculated using estimates of the following parameters:

\[
\text{Var}\{Z_i(\hat{H}^{-1}(1/2))\} = 2\text{Var}\left\{\hat{H}_{xx}(\hat{H}^{-1}(1/2))\right\} - 2\text{Var}\{\hat{H}_{xx}(\hat{H}^{-1}(1/2))\} \\
= (m(m-1))^{-1} + 4(m-2) \binom{m}{2}^{-1} \text{Cov}\{I_{[0,\hat{H}^{-1}(1/2)]}(\rho(X_1, X_2)), I_{[0,\hat{H}^{-1}(1/2)]}(\rho(X_1, X_3))\},
\]

\[
\text{Cov}\{Z_2(t_1), Z_2(t_2)\} \\
= 2\text{Cov}\{\hat{H}_{xx}(t_1), \hat{H}_{xx}(t_2)\} + 4\text{Var}\{\hat{H}_{xy}(t_1), \hat{H}_{xy}(t_2)\} - 8\text{Cov}\{\hat{H}_{xx}(t_1), \hat{H}_{xy}(t_2)\} \\
= (8m - 4)m^{-2}(m-1)^{-1}\{t_1(1-t_2) - 2\text{Cov}\{I_{[0,t_1]}(\rho(X_1, X_2)), I_{[0,t_2]}(\rho(X_1, X_3))\}\} \text{ for } t_1 \leq t_2 \\
\text{and } \text{Cov}\{Z_i(t_1), Z_2(t_2)\} = 0.
\]

The significance of \( S \) is then found using its asymptotic chi-square approximation.

In both tables, Friedman and Rafsky's generalization of the runs test is denoted by FR and Schilling's unweighted test based on 3 nearest neighbors is denoted by SC. We also include the likelihood ratio test (LRT) where all parameters are estimated and the populations are assumed to be normal. This test is the asymptotically most powerful unbiased invariant test when both populations are normal [see Anderson (1958)]. All of the tests were conducted at significance level 0.05, with sample sizes \( m=n=100 \), and the power was estimated by the percent of rejections from 1000 replications.
Table 3.1 Power of the normal theory test and various broad based nonparametric tests when the first population is standard normal.

% of 1000 samples declared significant ($\alpha=0.05$, $m=n=100$)

<table>
<thead>
<tr>
<th>Dimension=2</th>
<th>LRT</th>
<th>FR</th>
<th>SC</th>
<th>$S_{1,1}$</th>
<th>$S_{1,2}$</th>
<th>$S^*_{1,1}$</th>
<th>$Q_{2,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard normal</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>location shift</td>
<td>77</td>
<td>43</td>
<td>29</td>
<td>67</td>
<td>78</td>
<td>60</td>
<td>29</td>
</tr>
<tr>
<td>difference in scale</td>
<td>45</td>
<td>15</td>
<td>13</td>
<td>58</td>
<td>50</td>
<td>42</td>
<td>70</td>
</tr>
<tr>
<td>mixture</td>
<td>90</td>
<td>8</td>
<td>47</td>
<td>39</td>
<td>57</td>
<td>71</td>
<td>12</td>
</tr>
<tr>
<td>correlated normals</td>
<td>76</td>
<td>30</td>
<td>24</td>
<td>19</td>
<td>20</td>
<td>31</td>
<td>9</td>
</tr>
<tr>
<td>Tukey</td>
<td>5</td>
<td>33</td>
<td>25</td>
<td>16</td>
<td>14</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td><strong>Average Rank</strong></td>
<td>3.6</td>
<td>3.9</td>
<td>3.2</td>
<td>2.8</td>
<td>2.9</td>
<td>4.6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension=5</th>
<th>LRT</th>
<th>FR</th>
<th>SC</th>
<th>$S_{1,1}$</th>
<th>$S_{1,2}$</th>
<th>$S^*_{1,1}$</th>
<th>$Q_{2,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard normal</td>
<td>5</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>location shift</td>
<td>90</td>
<td>70</td>
<td>58</td>
<td>93</td>
<td>97</td>
<td>71</td>
<td>43</td>
</tr>
<tr>
<td>difference in scale</td>
<td>63</td>
<td>28</td>
<td>20</td>
<td>93</td>
<td>90</td>
<td>92</td>
<td>97</td>
</tr>
<tr>
<td>mixture</td>
<td>100</td>
<td>6</td>
<td>100</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>correlated normals</td>
<td>69</td>
<td>26</td>
<td>20</td>
<td>10</td>
<td>13</td>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>Tukey</td>
<td>5</td>
<td>15</td>
<td>18</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td><strong>Average Rank</strong></td>
<td>3.6</td>
<td>3.4</td>
<td>3.5</td>
<td>2.9</td>
<td>3.1</td>
<td>4.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension=10</th>
<th>LRT</th>
<th>FR</th>
<th>SC</th>
<th>$S_{1,1}$</th>
<th>$S_{1,2}$</th>
<th>$S^*_{1,1}$</th>
<th>$Q_{2,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard normal</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>location shift</td>
<td>92</td>
<td>88</td>
<td>78</td>
<td>99</td>
<td>100</td>
<td>77</td>
<td>60</td>
</tr>
<tr>
<td>difference in scale</td>
<td>18</td>
<td>11</td>
<td>8</td>
<td>72</td>
<td>67</td>
<td>68</td>
<td>84</td>
</tr>
<tr>
<td>mixture</td>
<td>100</td>
<td>4</td>
<td>100</td>
<td>44</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>correlated normals</td>
<td>41</td>
<td>16</td>
<td>12</td>
<td>7</td>
<td>6</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>Tukey</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td><strong>Average Rank</strong></td>
<td>3.2</td>
<td>3.5</td>
<td>3.2</td>
<td>3.2</td>
<td>3.9</td>
<td>4.0</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 Power of the normal theory test and various broad based nonparametric tests when the first population is uniform on the d-dimensional hypercube.

% of 1000 samples declared significant ($\alpha=0.05$, m=n=100)

<table>
<thead>
<tr>
<th>Dimension=2</th>
<th>LRT</th>
<th>FR</th>
<th>SC</th>
<th>$S_{1,1}$</th>
<th>$S_{1,2}$</th>
<th>$S_{2,1}^{*}$</th>
<th>$Q_{2,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Beta</td>
<td>100</td>
<td>98</td>
<td>90</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>correlated uniforms</td>
<td>84</td>
<td>53</td>
<td>80</td>
<td>38</td>
<td>53</td>
<td>57</td>
<td>29</td>
</tr>
<tr>
<td>checkerboard</td>
<td>2</td>
<td>8</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>6</td>
</tr>
<tr>
<td>Neyman clustering</td>
<td>21</td>
<td>86</td>
<td>90</td>
<td>26</td>
<td>36</td>
<td>27</td>
<td>30</td>
</tr>
<tr>
<td>Average Rank</td>
<td>3.6</td>
<td>3.5</td>
<td>3.9</td>
<td>2.8</td>
<td>2.9</td>
<td>4.4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension=5</th>
<th>LRT</th>
<th>FR</th>
<th>SC</th>
<th>$S_{1,1}$</th>
<th>$S_{1,2}$</th>
<th>$S_{2,1}^{*}$</th>
<th>$Q_{2,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Beta</td>
<td>0</td>
<td>72</td>
<td>56</td>
<td>97</td>
<td>98</td>
<td>89</td>
<td>95</td>
</tr>
<tr>
<td>correlated uniforms</td>
<td>0</td>
<td>72</td>
<td>57</td>
<td>24</td>
<td>59</td>
<td>65</td>
<td>25</td>
</tr>
<tr>
<td>checkerboard</td>
<td>0</td>
<td>7</td>
<td>6</td>
<td>92</td>
<td>95</td>
<td>74</td>
<td>6</td>
</tr>
<tr>
<td>Neyman clustering</td>
<td>0</td>
<td>90</td>
<td>100</td>
<td>31</td>
<td>48</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Average Rank</td>
<td>3</td>
<td>4.1</td>
<td>4</td>
<td>2</td>
<td>3.5</td>
<td>4.4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension=10</th>
<th>LRT</th>
<th>FR</th>
<th>SC</th>
<th>$S_{1,1}$</th>
<th>$S_{1,2}$</th>
<th>$S_{2,1}^{*}$</th>
<th>$Q_{2,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Second population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uniform</td>
<td>0</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Beta</td>
<td>0</td>
<td>45</td>
<td>32</td>
<td>84</td>
<td>86</td>
<td>59</td>
<td>80</td>
</tr>
<tr>
<td>correlated uniforms</td>
<td>0</td>
<td>96</td>
<td>87</td>
<td>37</td>
<td>83</td>
<td>93</td>
<td>60</td>
</tr>
<tr>
<td>checkerboard</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>52</td>
<td>61</td>
<td>35</td>
<td>5</td>
</tr>
<tr>
<td>Neyman clustering</td>
<td>0</td>
<td>88</td>
<td>100</td>
<td>37</td>
<td>53</td>
<td>36</td>
<td>47</td>
</tr>
<tr>
<td>Average Rank</td>
<td>3.1</td>
<td>4</td>
<td>3.8</td>
<td>2.2</td>
<td>3.8</td>
<td>4.1</td>
<td></td>
</tr>
</tbody>
</table>
The likelihood ratio test is the best test to use when the two population are normal or close to normal. However, the test is not robust when the data are far away from normality. In spite of this, there are several cases within Table 3.1, where the populations are normal or close to normal, where the proposed nonparametric tests are more powerful than LRT. This is explained by the fact that the calculation of LRT in five dimension, say, requires the estimation of 5 means and 15 covariance parameters for each sample separately and for the pooled sample. All of the distance based tests do not require the estimation of any additional parameters in higher dimensions.

In order to facilitate the comparison of the nonparametric tests, the average rank was calculated in each table for each dimension. $S_{1,2}$ has the lowest average rank uniformly. $U_2$ has the highest average rank in all cases. The proposed unconditional tests seem to perform as well or better than SC or FR in all cases except the Neyman clustering alternative. A test which measures the differences in the empirical interpoint distance distributions for small distances (take $Z = (Z_1(\hat{H}^{-1}(0.05)), Z_2(\hat{H}^{-1}(0.05)))^T$ in equation (3.5), for example) performs almost equally as well as SC and FR for the clustering alternatives, but not as well as $S_{1,1}$ in the table for the other alternatives.

The consistent test $Q_\infty$ of Section 2 was also investigated, but its power was found to be nearly identical to $Q_{2,\tilde{H}}$ in every case.

### 3.4 Asymptotic Power and Efficiency of the Unconditional Tests

We now find the asymptotic relative efficiency of $S_1$ as defined above versus the likelihood ratio test for bivariate normal populations when the difference is in scale. Let $X_1, ..., X_m \sim F_n$ and $Y_1, ..., Y_m \sim G_n$ be two independent
random samples of bivariate normal random vectors with identical means and covariance matrices given by \( \text{Var}(X_1) = I_2 \) and \( \text{Var}(Y_1) = (1 + a \cdot n^{-1})^2 I_2 \) for the \( n^{th} \) alternative. For each \( n \), \( H_{XX}(t) = 1 - \exp \left\{ -t/4 \right\} \), \( H_{XY}(t) = 1 - \exp \left\{ -t \left( 1 + a n^{-1/2} \right)^2 / 4 \right\} \). Also, recall that \( \mathbb{E} \left\{ Z_1 \left( \hat{H}^{-1}(1/2) \right) \right\} \rightarrow H_{XX} \left( H_{XX}^{-1}(1/2) \right) - H_{YY} \left( H_{XX}^{-1}(1/2) \right) \), \( \mathbb{E} \left\{ Z_2 \left( \hat{H}^{-1}(1/2) \right) \right\} \rightarrow H_{XX} \left( H_{XX}^{-1}(1/2) \right) + H_{YY} \left( H_{XX}^{-1}(1/2) \right) - 2 H_{XY} \left( H_{XX}^{-1}(1/2) \right) \), \( \text{mVar} \left\{ Z_1 \left( \hat{H}^{-1}(1/2) \right) \right\} \rightarrow 8 \text{Cov} \left( I_{[0,H_{XX}^{-1}(1/2)]} (\rho(X_1, X_2)), I_{[0,H_{XX}^{-1}(1/2)]} (\rho(X_1, X_3)) \right) \), and \( \text{m}^2 \text{Var} \left\{ Z_2 \left( \hat{H}^{-1}(1/2) \right) \right\} \rightarrow 2 - 16 \text{Cov} \left( I_{[0,H_{XX}^{-1}(1/2)]} (\rho(X_1, X_2)), I_{[0,H_{XX}^{-1}(1/2)]} (\rho(X_1, X_3)) \right) \).

The power of a 0.05 level test based on \( S_1 \) is given by

\[
\Phi \left\{ -z_{0.975} - \frac{1/2 - H_{YY} \left( H_{XX}^{-1}(1/2) \right)}{\sqrt{\text{Var} \left\{ Z_1 \left( H_{XX}^{-1}(1/2) \right) \right\}}} \right\} + 1 - \Phi \left\{ z_{0.975} - \frac{1/2 - H_{YY} \left( H_{XX}^{-1}(1/2) \right)}{\sqrt{\text{Var} \left\{ Z_1 \left( H_{XX}^{-1}(1/2) \right) \right\}}} \right\} + O(n^{-1})
\]

where \( z_{0.975} \) is the 0.975 quantile of a standard normal random variable. The likelihood ratio test described above is asymptotically equivalent to the most powerful test of whether the variance of each coordinate of the vectors in the second sample is 1. Consequently, the power of the likelihood ratio test when \( m = n \) is given by

\[
\Phi \left\{ -z_{0.975} - 2a + \frac{-a^2 + 2a + az_{0.975}}{\sqrt{n}} \right\} + 1 - \Phi \left\{ z_{0.975} - 2a + \frac{-a^2 + 2a - az_{0.975}}{\sqrt{n}} \right\} + O(n^{-1}).
\]

Therefore, the asymptotic relative efficiency can be computed to be 0.444.

The asymptotic relative efficiency of the statistic \( S(t) \) defined in section 3.3 with \( Z = \left( Z_1(\hat{H}^{-1}(t)), Z_2(\hat{H}^{-1}(t)) \right)^T \) can be calculated for every dimension. The graphs of the asymptotic relative efficiency for \( 0 < t < 1 \) in dimensions 2, 5, and 10 are shown in Figure 3.2. It can be seen from these graphs that the optimal location of \( t \) is close to 0.5, but the asymptotic relative efficiency for \( t \) near the
optimal \( t \) remains close to the optimal value. Hence, the recommendation of choosing \( t=0.5 \) is not far from optimal in this case.

Instead of choosing \( t=0.5 \) heuristically, we may instead use part of the data to choose the number and the location of the evaluation points, as discussed previously. An alternate approach is to use a random number generator to choose the location of the points. For simplicity, assume that we will use \( r \) points and we will only evaluate the process \( Z_2 \), i.e. we will use the statistic defined in section 5.2 with

\[
Z = \begin{pmatrix}
Z_2(\hat{H}^{-1}(t_1)), & Z_2(\hat{H}^{-1}(t_2)), & \ldots, & Z_2(\hat{H}^{-1}(t_r))
\end{pmatrix}^T
\]

where \( t_1, t_2, \ldots, t_r \) are iid Unif(0,1). If the measure (with respect to \( H \)) of the set, \( M \), where \( Z_2 \) does not vanish is \( \epsilon>0 \), say, then the asymptotic power of the random evaluation point statistic is

\[
P[\text{any } t_i \text{ is in } M] + 0.05P[\text{all } t_i \text{ are in } M^c] = \left(1 - (1 - \epsilon)^r\right) + 0.05(1 - \epsilon)^r.
\]

Since this rapidly approaches 1, the random evaluation point test is asymptotically consistent (in \( r \)).
Figure 3.2 Asymptotic power of the Unconditional Test versus LRT for dimensions 2, 5, and 10
Figure 3.2 (continued)
Chapter IV

Goodness-of-fit Tests and Two-sample Tests for Very High Dimensional Data

Introduction

The two-sample tests in Chapter III use generalizations to dependent samples of nonparametric tests to test for the equality of the three interpoint distance distributions. Under mild conditions on the underlying high-dimensional distributions, the interpoint distances are approximately normal. Hence, a goodness-of-fit test or a two-sample test can be performed by simply testing for the equality of the means and variances of the interpoint distance distributions.

4.1 Motivation and Discussion

For random vectors $X_d$ and $Y_d$, the $L^p$ distance is defined as

$$
\|X_d - Y_d\|_p = \left\{ \sum_{i=1}^{d} |X_{d,i} - Y_{d,i}|^p \right\}^{\frac{1}{p}}.
$$

Under many circumstances, when $X_d$ and $Y_d$ are iid, the distribution of $\left\{\|X_d - Y_d\|_p\right\}^p$ is asymptotically normal as $d$ goes to infinity by an appropriate Central Limit Theorem. In those circumstances
where \( \left\{ \| X_d - Y_d \|_p \right\}^p \) is asymptotically normal, it is also true that
\[
f\left( \left\{ \| X_d - Y_d \|_p \right\}^p \right)
\] is asymptotically normal provided \( f(x) \) is differentiable at
\[
E\left\{ \| X_d - Y_d \|_p \right\}^p
\] and it is often desirable to choose a transformation
\[
f\left( \left\{ \| X_d - Y_d \|_p \right\}^p \right)
\] which is most rapidly approximated by a normal random variable. For example, when \( X_d \) and \( Y_d \) are iid and their coordinates are iid with \( E|X_{i1} - Y_{i1}|^p = \mu \), \( \text{Var}|X_{i1} - Y_{i1}|^p = \sigma^2 \), and \( E\left( |X_{i1} - Y_{i1}|^p - E|X_{i1} - Y_{i1}|^p \right)^3 = \theta \), then the skewness of \( \left\{ \| X_d - Y_d \|_p \right\} \) is
\[
\frac{1}{\sqrt{\text{Var}\left( \sum_{i=1}^d |X_{i1} - Y_{i1}|^p \right)^{\frac{3}{2}}}} E\left( \sum_{i=1}^d |X_{i1} - Y_{i1}|^p - E\sum_{i=1}^d |X_{i1} - Y_{i1}|^p \right)^3 = \frac{d\theta}{\left( d\mu^2 \right)^{\frac{3}{2}}} = \frac{\theta}{\sigma^3}.
\]
Since the skewness vanishes rather slowly in this case, it is desirable to search for a transformation \( f\left( \left\{ \| X_d - Y_d \|_p \right\}^p \right) \) such that the skewness is 0. Using the Taylor series for \( f(W) \) about \( EW \) with \( W = W_d = \frac{1}{d} \left\{ \| X_d - Y_d \|_p \right\}^p \), we find
\[
f(W) = f(EW) + f'(EW)(W - EW) + \frac{1}{2} f''(EW)(W - EW)^2 + \cdots.
\]
Hence, \( Ef(W) = f(\mu) + \frac{1}{2} f'(\mu) \sigma^2 d^{-1} + o(d^{-2}) \). So,
\[
\text{Var}(f(W)) = E[f'(W)]^2 - \left[ f(W) \right]^2 = (f'(\mu))^2 \sigma^2 d^{-1} + o(d^{-2}), \text{ and}
\]
\[
E(f(W) - Ef(W))^3 = E[f^3(W)] - 3\text{Var}(f(W))Ef(W) - (Ef(W))^3
\]
\[
= \left( 3\sigma^4 (f'(\mu))^2 f''(\mu) + \theta (f'(\mu))^3 \right) d^{-2} + o(d^{-3}).
\]
Hence, the skewness of \( f(W) \) is
\[
\frac{\left( 3\sigma^4 (f'(\mu))^2 f''(\mu) + \theta (f'(\mu))^3 \right) d^{-2}}{\left( (f'(\mu))^2 \sigma^2 d^{-1} \right)^{\frac{3}{2}}} + o\left( d^{-\frac{3}{2}} \right) = \left( \frac{3\sigma f''(\mu)}{f'(\mu)} + \frac{\theta}{\sigma^3} \right) d^{-\frac{1}{2}} + o\left( d^{-\frac{3}{2}} \right).
\]
In summary, the optimal transformation (for making the skewness vanish) would satisfy \( f''(\mu) = -\frac{\theta f'(\mu)}{3\sigma^4} \). The optimal power transformation, \( f(W) = W^j \), would satisfy \( j(j-1)\mu^{-2} = -\frac{\theta |\mu|^{-1}}{3\sigma^4} \), or \( j = 1 - \frac{\theta \mu}{3\sigma^4} \). When \( p = 2 \) and the coordinates are iid Unif(0,1) random variables, this optimal power is \( j = \frac{589}{1029} \approx \frac{1}{2} \) and when the coordinates are iid normal random variables, the optimal power transformation is \( j = \frac{1}{3} \).

An important particular case involves vectors of iid Uniform(0,1) random variables. Let \( X_d \) and \( Y_d \) be independent vectors of \( d \) independent coordinates which are Uniform(0,1) and define \( Z_d = \sum_{k=1}^{d} (X_{d_k} - Y_{d_k})^2 \). The distribution function for \( Z_1 = (X_d - Y_d)^2 \) is \( F_1(z) = \begin{cases} 2\sqrt{z} - z & \text{if } 0 \leq z \leq 1 \\ 0 & \text{elsewhere} \end{cases} \) and the density is \( f_1(z) = \begin{cases} \frac{1}{\sqrt{z}} - 1 & \text{if } 0 \leq z \leq 1 \\ 0 & \text{elsewhere} \end{cases} \). Hence, it may be verified that \( EZ_1 = \frac{1}{6} \), \( EZ_1^2 = \frac{1}{15} \), \( EZ_1^3 = \frac{1}{28} \) and in general \( EZ_1^n = \frac{1}{(n+1)(2n+1)} \). Thus, \( EZ_d = \frac{d}{6} \) and the skewness is

\[
\frac{1}{\sqrt{\frac{7d}{180}}} \left[ Z_d - \frac{d}{6} \right]^{\frac{3}{2}} = \frac{1}{\sqrt{\frac{7d}{180}}} \left[ \frac{11}{945} \right] = \frac{88\sqrt{5}}{\sqrt{16807d}}
\]

For small values of \( d \), it is possible to find the exact distribution of \( Z_d \). For example, the density of \( Z_2 \) is

\[
f_2(z) = \begin{cases} \pi + z - 4\sqrt{z} & \text{if } 0 < z < 1 \\ -2 + 4\sqrt{z} - 1 - z + 2\sin^{-1}\left( \frac{1}{\sqrt{z}} \right) - 2\sin^{-1}\left( \frac{\sqrt{z} - 1}{\sqrt{z}} \right) & \text{if } 1 < z < 2 \\ 0 & \text{elsewhere} \end{cases}
\]
However, the distribution of $Z_2$ cannot be expressed with elementary functions. The density of $Z_2 = \sum_{k=1}^{3} (X_{3k} - Y_{3k})^2$ is

$$f_3(z) = \begin{cases} 
-\frac{3}{2} + 4z \frac{3}{2} - 3\pi z + 2\pi \sqrt{z} & \text{if } 0 < z < 1 \\
\frac{1}{2} \int f_1(z-t) f_2(t) dt & \text{if } 1 < z < 2 \\
0 & \text{if } 2 < z < 3 \\
\end{cases}$$

and again, the distribution cannot be expressed with elementary functions.

A second important case is when $X_d$ is multivariate normal. Here, it is not always the case that the coordinates are independent and therefore we will discuss sufficient conditions for asymptotic normality of the interpoint distances. Let $X_d$ be multivariate normal with mean vector $\mu_d$ and variance matrix $\Sigma_d$. Under the hypothesis $X_d = Y_d$ we know that $\sum_{k=1}^{d} (X_{dk} - Y_{dk})^2$ has the distribution of $\sum_{k=1}^{d} 2\lambda_{dk} \chi_{dk}^2$ where $\lambda_{d1}, ..., \lambda_{dd}$ are the eigenvalues of $\Sigma_d$ and $\chi_{d1}^2, ..., \chi_{dd}^2$ are iid chi-square random variables with 1 degree of freedom. The following facts will be useful in finding the moments of the interpoint distance distribution in this case:

i) $E \sum_{k=1}^{d} (X_{dk} - Y_{dk})^2 = 2 \sum_{k=1}^{d} \lambda_{dk} = 2 \text{tr}[\Sigma_d]$

ii) $\text{Var} \left( \sum_{k=1}^{d} (X_{dk} - Y_{dk})^2 \right) = 8 \sum_{k=1}^{d} \lambda_{dk}^2 = 8 \text{tr}[\Sigma_d^2]$ using the fact that $\Sigma^2 v = \lambda^2 v$

provided $\Sigma v = \lambda v$.

iii) The moment generating function of $2\lambda_{dk} \chi_{dk}^2$ is

$$M_{dk}(t) = E e^{t\lambda_{dk} \chi_{dk}^2} = \sqrt{\frac{1}{1 - 4\lambda_{dk} t}}$$

so the moment generating function of

$$\sum_{k=1}^{d} 2\lambda_{dk} \chi_{dk}^2$$

is $M_d(t) = \prod_{k=1}^{d} M_{dk}(t) = \sqrt{\frac{1}{\prod_{k=1}^{d} (1 - 4\lambda_{dk} t)}} = \sqrt{\frac{1}{(4t)^d f_d \left( \frac{1}{4t} \right)}}$
where \( f_d(x) \) is the characteristic polynomial of \( \Sigma_d \).

Now, if \( 2\lambda_{di(d)} < \varepsilon \sum_{j=1}^{d} \lambda_{dj}^2 \) for all sufficiently large \( d \), then we only need to consider the right tail. The density in the right tail of the centered random variables are dominated by the density of \( 2\lambda_{di(d)} \chi_{di(d)}^2 - 2\lambda_{d(d)} \) where \( \lambda_{di(d)} \) is the largest eigenvalue of \( \Sigma_d \). So, the condition that needs to be satisfied in order to apply Lindeberg's central limit theorem is

\[
\lim_{d \to \infty} \frac{4d\lambda_{di(d)}^2}{\sum_{j=1}^{d} 8\lambda_{dj}^2} E \left[ \left( \chi^2 - 1 \right)^2 \left| \chi^2 - 1 > \frac{\varepsilon}{2\lambda_{di(d)}} \sqrt{\sum_{j=1}^{d} 8\lambda_{dj}^2} \right. \right] = 0
\]

where \( \chi^2 \) is a chi-square random variable with 1 d.f. Now,

\[
g(t) = E \left[ \left( \chi^2 - 1 \right)^2 \left| \chi^2 - 1 > t \right. \right] = 2 - \sqrt{\frac{2}{\pi}} \left\{ \sqrt{2\pi} \left( 2\Phi\left( \sqrt{t} \right) - 1 \right) - e^{-\frac{1}{2}t} \sqrt{t(1+t)} \right\}
\]

and this converges to 0 at a rate faster than any rational function of \( t \), indeed

\[
g(t)^{\sqrt{t^{-3}e^t}} = \sqrt{\frac{2}{\pi}} + \sqrt{\frac{2}{\pi}} t^{-1} + 2\sqrt{\frac{2}{\pi}} t^{-2} + o(t^{-3}).
\]

So, for the theorem to apply, we need

\[
\lim_{d \to \infty} \frac{\varepsilon}{\sqrt{2\lambda_{di(d)}}} \frac{4d\lambda_{di(d)}^2}{\sum_{j=1}^{d} 8\lambda_{dj}^2} = 0.
\]

This is a sufficient condition, but by no means necessary. For small \( d \), the normal approximation will be accurate provided there are a few large eigenvalues of roughly the same size since the condition derived above intuitively will hold if the largest eigenvalue does not dominate the sum of the squares of the eigenvalues.

An example where \( X_d \) is multivariate normal and the interpoint distance is asymptotically normal is when the coordinates are consecutive observations from an autoregressive process defined by \( X_{d1} = \varepsilon_{d1} \) and

\[
X_{di} = \rho X_{di-1} + \sqrt{1-\rho^2} \varepsilon_{di} \text{ for } i = 2, 3, \ldots, d \text{ where } \varepsilon_{di} \text{ are iid N}(0,1). \]

The variance
matrix in this case is $\Sigma_d = \begin{bmatrix} 1 & \rho & \ldots & \rho^{d-1} \\ \rho & 1 & \ddots & \\ \vdots & \ddots & \ddots & \rho \\ \rho^{d-1} & \ldots & \rho & 1 \end{bmatrix}$. In this case, since it is not possible to find a formula for the largest eigenvalue of $\Sigma_d$ as a function of $d$, it is easier to appeal to a result from the literature on time series. For Normal iid errors Mann (1943) shows that if $Z_1, Z_2, \ldots Z_d$ are consecutive observations from a stationary autoregressive process, then $\sum_{i=1}^{d} Z_i^2$ is asymptotically normal. A similar result for more general errors is shown in Walker (1954). To prove the result that we need, take $Z_i = X_i - Y_i$. The mean and variance of the interpoint distance distribution are $E\sum_{k=1}^{d} 2\lambda_{dk} X_{dk}^2 = \sum_{k=1}^{d} 2\lambda_{dk} = 2\text{tr}[\Sigma_d] = 2d$ and

$$\text{Var}\left[\sum_{k=1}^{d} 2\lambda_{dk} X_{dk}^2 \right] = \sum_{k=1}^{d} \text{Var}[2\lambda_{dk}] = 8\text{tr}[\Sigma_d^2] = 8\left\{ d + \sum_{k=2}^{d} 2(d+1-k)\rho^{2(k-1)} \right\}$$

$$= 8\frac{2\rho^{2d+2} - d\rho^4 - 2\rho^2 + d}{(1-\rho^2)^2}.$$ 

Goodness-of-fit Test

A goodness-of-fit test statistic can be performed by testing for the equality of sample estimates of the first two sample moments of the two unknown interpoint distance distributions with the corresponding moments of the known hypothesized interpoint distance distribution. With high dimensional data, as described above, under most circumstances all three of these distributions are normal. The equality of the first two moments is always a necessary condition for the equality of distributions, but the asymptotic normality of the distributions tells us that it is also sufficient.
Let $x_1, x_2, ..., x_n$ be the observed sample of d-dimensional vectors and let $h: \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, \infty)$ be a function which satisfies $h(X, Y) = 0$ iff $X = Y$. In order to perform the goodness-of-fit test, first estimate $\theta_{XX}^{(1)}$ and $\theta_{XX}^{(2)}$, the first and second moments of the distribution of $h(X_1, X_2)$ where $X_1$ and $X_2$ are independent random variables with distribution $F$. This can be done with the estimates $\hat{\theta}_{XX}^{(1)} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} h(x_i, x_j)$ and $\hat{\theta}_{XX}^{(2)} = \frac{1}{\binom{n}{2}^2} \sum_{1 \leq i < j \leq n} h(x_i, x_j)^2$. Next, estimate $\theta_{XY}^{(1)}$ and $\theta_{XY}^{(2)}$, the first and second moments of the distribution of $h(X, Y)$ where $X$ is a random variable with distribution $F$ and $Y$ is an independent random variable with distribution $G$. This can be done with the estimates $\hat{\theta}_{XY}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} h(x_i, Y_i)$ and $\hat{\theta}_{XY}^{(2)} = \frac{1}{n} \sum_{i=1}^{n} h(x_i, Y_i)^2$ where $Y_i; i=1, 2, ..., n$ are independent random variables with distribution $G$ which are also independent of the observed sample. In specific problems, it may be desired to use other estimates of the moments of the interpoint distance distributions. For instance, if $F$ is the distribution function of a multivariate normal random variable with mean $\mu$ and variance matrix $\Sigma$ and $G$ is the distribution function of a multivariate normal random variable with mean $\mu_0$ and variance matrix $\Sigma_0$ and $h(X, Y)$ is the square of the Euclidean distance between $X$ and $Y$, then we saw in the previous section that $\theta_{XX}^{(1)} = 2 \text{tr} [\Sigma]$ and $\theta_{XX}^{(2)} = 8 \text{tr} [\Sigma^2]$ and since the sample covariance matrix, $\hat{\Sigma}$, is a complete and sufficient statistic for $\Sigma$, it is clearly better to use the moment estimates $\hat{\theta}_{XX}^{(1)} = 2 \text{tr} [\hat{\Sigma}]$ and $\hat{\theta}_{XX}^{(2)} = 8 \text{tr} [\hat{\Sigma}^2]$. In this discussion, we will not restrict $F$ to be inside a parametric family and therefore, without the benefit of sufficient statistics, we will use the natural estimates described previously. The vector
of estimates of the moments for the two unknown interpoint distances based on all of the sample interpoint distances is a 4-dimensional U-statistic and is therefore asymptotically jointly normal.

4.2 Examples

Example 1. Random Number Generator Testing

As an application, we will use this goodness-of-fit test with $h(X, Y)$ being the square of the Euclidean distance between $X$ and $Y$ to test several pseudo-random number generators. The many tests for independence of successive observations from a pseudo-random process include Savir's test [Savir 1983, Marsaglia 1985], the runs test, the poker test, and the coupon collector's test [Dagnupar (1988), Knuth (1968)].

For our proposed test, a sample size of $n$ vectors of length $d$ consisting of consecutive outcomes of the pseudo-random number generator are observed. $G$ is the distribution of a random vector which is uniform on the $d$-dimensional hypercube. Marsaglia [1968] showed that the observed vectors actually lie on a finite number of equally spaced parallel hyperplanes of dimension less than $d$. Using the comments in the proceeding paragraphs, it may be verified that under $H_0$,

$$\hat{\Theta}^{(1)}_{XX} = \frac{1}{n \choose 2} \sum_{1 \leq i < j \leq n} \sum_{k=1}^{d} (x_{ik} - x_{jk})^2,$$

$$\hat{\Theta}^{(2)}_{XX} = \frac{1}{n \choose 2} \sum_{1 \leq i < j \leq n} \left\{ \sum_{k=1}^{d} (x_{ik} - x_{jk})^2 \right\}^2.$$
\[
\hat{\theta}_{(2)}^{(2)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \left( x_{ij} - \bar{Y}_{ij} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \left( x_{ij} - x_{ij} + \frac{1}{3} \right) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \left( x_{ij}^2 - x_{ij} + \frac{1}{3} \right),
\]

We have shown that \( \hat{\theta} = \begin{pmatrix} \hat{\theta}_{(1)}^{(1)} \\ \hat{\theta}_{(2)}^{(1)} \\ \hat{\theta}_{(1)}^{(2)} \\ \hat{\theta}_{(2)}^{(2)} \end{pmatrix} \) is asymptotically multivariate normal with

\[
\begin{pmatrix} d/6 \\ 7d + 5d^2 \\ 180 \frac{d}{6} \\ 7d + 5d^2 \\ 180 \end{pmatrix}
\]

mean \( \theta = \begin{pmatrix} \frac{d}{6} \\ \frac{180}{d} \\ \frac{6}{7d + 5d^2} \\ \frac{180}{180} \end{pmatrix} \). The elements of the covariance matrix are found as

follows:

\[
\text{Var}(\hat{\theta}_{(1)}^{(1)}) = \frac{d \text{Var}\{(X_{11} - X_{21})^2\} + 2(n - 2)d \text{Cov}\{(X_{11} - X_{21})^2, (X_{11} - X_{31})^2\}}{n \choose 2}
\]

\[
= \frac{d}{180} + 2(n - 2)d \frac{1}{180} = d(3 + 2n) \frac{n}{90n(n - 1)}.
\]
\[
\begin{align*}
\left(\frac{n}{2}\right)\text{Var}\left(\hat{\theta}_{XX}^{(2)}\right) &= \text{Var}\left(\sum_{i=1}^{d}(X_{xi} - X_{x_2})^2\right) + 2(n-2)\text{Cov}\left(\sum_{i=1}^{d}(X_{xi} - X_{x_2})^2, \left(\sum_{i=1}^{d}(X_{xi} - X_{x_2})^2\right)\right) \\
&= \left\{ d \text{Var}\left((X_{x_1} - X_{x_2})^4\right) + 2d(d-1)\text{Var}\left((X_{x_1} - X_{x_2})^2(X_{x_2} - X_{x_2})^2\right) \right\} \\
&\quad + 4d(d-1)\text{Cov}\left((X_{x_1} - X_{x_2})^4, (X_{x_1} - X_{x_2})^2(X_{x_2} - X_{x_2})^2\right) \\
&\quad + 24 \binom{d}{3}\text{Cov}\left((X_{x_1} - X_{x_2})^2(X_{x_2} - X_{x_2})^2, (X_{x_1} - X_{x_2})^2(X_{x_3} - X_{x_3})^2\right) \\
&\quad + 2(n-2)d\text{Cov}\left((X_{x_1} - X_{x_2})^4, (X_{x_1} - X_{x_2})^4\right) \\
&\quad + 2(n-2)8 \binom{d}{2}\text{Cov}\left((X_{x_1} - X_{x_2})^4, (X_{x_1} - X_{x_3})^2(X_{x_2} - X_{x_2})^2\right) \\
&\quad + 2(n-2)4 \binom{d}{2}\text{Cov}\left((X_{x_1} - X_{x_2})^2(X_{x_2} - X_{x_2})^2, (X_{x_1} - X_{x_3})^2(X_{x_2} - X_{x_2})^2\right) \\
&\quad + 2(n-2)24 \binom{d}{3}\text{Cov}\left((X_{x_1} - X_{x_2})^2(X_{x_2} - X_{x_2})^2, (X_{x_1} - X_{x_3})^2(X_{x_3} - X_{x_3})^2\right) \\
&= \left\{ \frac{d}{225} + 2d(d-1)\frac{119}{32400} + 4d(d-1)\frac{31}{7560} + 4d(d-1)(d-2)\frac{7}{6480} \right\} \\
&\quad + 2(n-2)\left\{ d\frac{1}{1350} + 8 \binom{d}{2}\frac{1}{1512} + 4 \binom{d}{2}\frac{11}{32400} + 24 \binom{d}{3}\frac{1}{6480} \right\} \\
&= d\frac{-45 + 555d + 210d^2 + 174n + 334nd + 140nd^2}{113400},
\end{align*}
\]

\[
\left(\frac{n}{2}\right)\text{Cov}\left(\hat{\theta}_{XX}^{(1)}, \hat{\theta}_{XX}^{(2)}\right) = \text{Cov}\left(\sum_{i=1}^{d}(X_{xi} - X_{x_2})^2, \left(\sum_{i=1}^{d}(X_{xi} - X_{x_2})^2\right)\right) \\
+ 2(n-2)\text{Cov}\left(\sum_{i=1}^{d}(X_{xi} - X_{x_2})^2, \left(\sum_{i=1}^{d}(X_{xi} - X_{x_3})^2\right)\right) \\
= d\text{Cov}\left((X_{x_1} - X_{x_2})^2, (X_{x_1} - X_{x_2})^4\right) \\
+ 2d(d-1)\text{Cov}\left((X_{x_1} - X_{x_2})^2, (X_{x_2} - X_{x_2})^2\right) \\
+ 2(n-2)d\text{Cov}\left((X_{x_1} - X_{x_2})^2, (X_{x_1} - X_{x_3})^4\right) \\
+ 2(n-2)2d(d-1)\text{Cov}\left((X_{x_1} - X_{x_2})^2, (X_{x_2} - X_{x_2})^2\right)
\]

110
\[
\begin{align*}
&= d \left\{ \frac{31}{1260} + 2(d - 1) \frac{7}{1080} + 2(n - 2) \frac{1}{252} + 2(n - 2)2(d - 1) \frac{1}{1080} \right\} \\
&= \frac{12 + 21d + 16n + 14nd}{3780},
\end{align*}
\]

\[
\text{Var}\left[\hat{\theta}_{XY}^{(1)}\right] = \frac{nd\text{Var}\{X_{i1}^2 - X_{i1}\}}{n^2} = \frac{d}{180n},
\]

\[
\text{Cov}\left[\hat{\theta}_{XY}^{(1)}, \hat{\theta}_{XY}^{(1)}\right] = -\frac{2d\text{Cov}\{(X_{i1} - X_{i2})^2, X_{i1}^2 - X_{i1}\}}{n} = \frac{d}{90n},
\]

\[
\text{Cov}\left[\hat{\theta}_{XX}, \hat{\theta}_{XY}^{(1)}\right] = \frac{2d\text{Cov}\{(X_{i1} - X_{i2})^4, X_{i1}^2 - X_{i1}\} + \frac{d}{2}\text{Cov}\{(X_{i1} - X_{i2})^2(X_{i2} - X_{i2})^2, X_{i1}^2 - X_{i1}\}}{n}
\]

\[
= \frac{\frac{d}{756} + \frac{d(d - 1)}{540}}{n} = \frac{7d^2 + d}{1890n},
\]

\[
\text{Var}\left[\hat{\theta}_{XY}^{(2)}\right] = \frac{nd\text{Var}\{X_{i1}^4 - 2X_{i1}^3 + 2X_{i1}^2 - X_{i1}\}}{n^2}
\]

\[
= 4n\left(\frac{d}{2}\text{Var}\left\{X_{i1}^2 - X_{i1} + \frac{1}{3}\right\}\left\{X_{i2}^2 - X_{i2} + \frac{1}{3}\right\}\right)
\]

\[
+ \frac{4n(d - 1)d\text{Cov}\{X_{i1}^4 - 2X_{i1}^3 + 2X_{i1}^2 - X_{i1}, X_{i1}^2 - X_{i1} + \frac{1}{3}\}\left\{X_{i2}^2 - X_{i2} + \frac{1}{3}\right\}\}}{n^2}
\]

\[
+ \frac{n24\left(\frac{d}{3}\right)\text{Cov}\{X_{i1}^2 - X_{i1} + \frac{1}{3}\}\left\{X_{i2}^2 - X_{i2} + \frac{1}{3}\right\}\left\{X_{i1}^2 - X_{i1} + \frac{1}{3}\right\}\left\{X_{i3}^2 - X_{i3} + \frac{1}{3}\right\}\}}{n^2}
\]

\[
= \frac{\frac{d}{350} + \frac{11}{2400} + \frac{4d(d - 1)}{1512} + \frac{1}{6480}}{n} = \frac{87d + 167d^2 + 70d^2}{113400n},
\]

111
\[ \text{Cov}\{\hat{\theta}_{xx}^{(2)}, \hat{\theta}_{xy}^{(2)}\} = \frac{2 \binom{n}{2} d \text{Cov}\{(X_{11} - X_{21})^4, X_{11}^4 - 2X_{11}^3 + 2X_{11}^2 - X_{11}\}}{n \binom{n}{2}} \]

\[ = \frac{2d \frac{1}{252} + 4d(d - 1) \frac{1}{540}}{n} = \frac{8 + 7d}{1890n}, \]

\[ \text{Cov}\{\hat{\theta}_{xx}^{(2)}, \hat{\theta}_{xy}^{(2)}\} = \frac{2 \binom{n}{2} d \text{Cov}\{(X_{11} - X_{21})^4, X_{11}^4 - 2X_{11}^3 + 2X_{11}^2 - X_{11}\}}{n \binom{n}{2}} \]

\[ + \frac{\binom{n}{2} 4d(d - 1) \text{Cov}\{(X_{11} - X_{21})^4, (X_{12} - X_{22})^2, X_{11}^4 - 2X_{11}^3 + 2X_{11}^2 - X_{11}\}}{n \binom{n}{2}} \]

\[ + \frac{\binom{n}{2} 2d(d - 1) \text{Cov}\{(X_{11} - X_{21})^4, (X_{11}^2 - X_{11} + \frac{1}{3})(X_{12}^2 - X_{12} + \frac{1}{3})\}}{n \binom{n}{2}} \]

\[ + \frac{16 \binom{n}{2} \binom{d}{2} \text{Cov}\{(X_{11} - X_{21})^4, (X_{12} - X_{22})^2, (X_{11}^2 - X_{11} + \frac{1}{3})(X_{12}^2 - X_{12} + \frac{1}{3})\}}{n \binom{n}{2}} \]

\[ + \frac{\binom{n}{2} 24 \binom{d}{3} \text{Cov}\{(X_{11} - X_{21})^4, (X_{12} - X_{22})^2, (X_{11}^2 - X_{11} + \frac{1}{3})(X_{13}^2 - X_{13} + \frac{1}{3})\}}{n \binom{n}{2}} \]
\[
2d \frac{1}{350} + 4d(d-1) \frac{1}{1512} + 4d(d-1) \frac{1}{1512} + 16 \left( \frac{d}{2} \right) \frac{11}{32400} + 24 \left( \frac{d}{3} \right) \frac{1}{6480} \\
= d \frac{87 + 167d + 70d^2}{56700n},
\]
\[
\text{Cov}\{\hat{\theta}^{(1)}_{XY}, \hat{\theta}^{(2)}_{XY}\} = \frac{n \text{Cov}\{X_{11}^2 - X_{11}, X_{11}^4 - 2X_{11}^3 + 2X_{11}^2 - X_{11}\}}{n^2}
\]
\[
+ 2 \frac{n(d-1) \text{Cov}\{X_{11}^2 - X_{11}, X_{11}^2 - X_{11} + \frac{1}{3}\} \left(X_{12}^2 - X_{12} + \frac{1}{3}\right)}{n^2},
\]
\[
= d \frac{1}{252} + 2d(d-1) \frac{1}{1080} = d \frac{77 + 7d}{3780n}.
\]

The likelihood ratio test statistic \( \left[ \hat{\theta} - \theta \right] \psi^{-1} \left[ \hat{\theta} - \theta \right] \) where \( \psi \) is the covariance matrix of \( \hat{\theta} \) under \( H_0 \) would be the best statistic if no information was available on the alternative.

It may be verified that under the null hypothesis, the estimates of the first two moments of the unknown interpoint distance distributions here are asymptotically perfectly correlated. Therefore, just one of these statistics is sufficient as a test statistic. Since it is the easiest to calculate, we will use the statistic which is the estimate of the mean of the X-to-Y interpoint distance distribution. After standardizing by the mean and standard deviation of the statistic under the null hypothesis, denote this statistic by \( T \).

Figure 4.1 shows the sample estimates of the first four moments of \( T \) using samples of size \( n=5 \) million of vectors of dimension \( d=1, 2, ..., 20 \) from IBM's RANDU and L'Ecuyer's random number generators. These estimates were based on 1000 observations of \( T \). Figure 4.2 shows the sample moments from 1000 (simulated) normal random variables replicated 100 times. One immediately sees that the kurtosis of \( T \) for RANDU is too far away from what
is expected under the null hypothesis for higher dimensional vectors. By the
definition of $T$, this could only happen if the high dimensional vectors are
not independent. The moments of $T$ for vectors generated by L'Ecuyer's
random number generator do not show any deviation from what would be
expected under the null hypothesis.
Figure 4.1 Sample moments of T for data from IBM's RANDU and L'Ecuyer's Pseudorandom number generators.
L’Ecuyer’s RNG

Moments of T

Dimension

- Mean of $T$ (0 under the Null Hypothesis)
- St. Dev. of $T$ (1 under the Null Hypothesis)
- Skewness of $T$ (0 under the Null Hypothesis)
- Kurtosis of $T$ (0 under the Null Hypothesis)

Figure 4.1 (continued)
Figure 4.2 Boxplots of Moments of 1000 Independent Standard Normal Random Variables

Example 2. The College Football Pool Data

Do the faculty and staff of the Statistics Department at The Ohio State University have different strategies of selecting winners of college football games than the graduates students? Every college football season, about half a dozen graduate students and half a dozen faculty and staff participate in a pool where every participant picks the winners of 21 football games each week.

The data examined here consists of the choices of seven graduate students and 7 faculty/staff over five weeks. This produces two samples
denoted by $X_1, \ldots, X_7$, the 105 dimensional vectors which contain the picks of the seven nongraduate students for five weeks of the season and $Y_1, \ldots, Y_7$, the vectors containing the picks of the graduate students for these five weeks. Figure 4.3 shows the picks for one week.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>OK</td>
<td>AL</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>by</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>ch</td>
<td>nw</td>
<td>OK</td>
<td>tn</td>
<td>KU</td>
<td>fl</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>PU</td>
<td>oh</td>
<td>nw</td>
<td>tx</td>
<td>AL</td>
<td>KU</td>
<td>fl</td>
<td>UM</td>
<td>wa</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>1</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>OK</td>
<td>tn</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>wa</td>
<td>LA</td>
<td>by</td>
<td>WV</td>
<td>af</td>
</tr>
<tr>
<td>1</td>
<td>MS</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>OK</td>
<td>AL</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>wa</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>1</td>
<td>IL</td>
<td>IA</td>
<td>PU</td>
<td>oh</td>
<td>nw</td>
<td>tx</td>
<td>AL</td>
<td>IS</td>
<td>fl</td>
<td>UM</td>
<td>SF</td>
<td>AZ</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>1</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>OK</td>
<td>tn</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>wa</td>
<td>LA</td>
<td>by</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>PU</td>
<td>WI</td>
<td>MN</td>
<td>OK</td>
<td>AL</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>NV</td>
</tr>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>tx</td>
<td>AL</td>
<td>KU</td>
<td>fl</td>
<td>UM</td>
<td>wa</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>NV</td>
</tr>
<tr>
<td>1</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>WI</td>
<td>nw</td>
<td>tx</td>
<td>AL</td>
<td>KU</td>
<td>fl</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>OK</td>
<td>tn</td>
<td>KU</td>
<td>fl</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>1</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>WI</td>
<td>nw</td>
<td>OK</td>
<td>AL</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>1</td>
<td>MS</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>OK</td>
<td>AL</td>
<td>KU</td>
<td>AU</td>
<td>RG</td>
<td>wa</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
<tr>
<td>0</td>
<td>IL</td>
<td>IA</td>
<td>ps</td>
<td>oh</td>
<td>nw</td>
<td>tx</td>
<td>TN</td>
<td>KU</td>
<td>AU</td>
<td>UM</td>
<td>SF</td>
<td>LA</td>
<td>AS</td>
<td>BC</td>
<td>af</td>
</tr>
</tbody>
</table>

**Figure 4.3** Data for 15 games from one week of the college football pool. Column 1 indicates graduate/nongraduate status (0=nongraduate student, 1=graduate student)

Figure 4.4 shows the normal probability plots for the three types of interpoint distances. These graphs show demonstrate that even for this situation, where the coordinates are binomial and far from independent, the interpoint distances can be close to normal.
Figure 4.4 Normal probability plots of the three types of observed interpoint distances
Figure 4.4 (continued)

The summary statistics in Table 4.1 indicate that there is a difference in the means and variances of the three interpoint distance distributions.

Table 4.1 Summary statistics for the three types of observed interpoint distances

<table>
<thead>
<tr>
<th>Distance</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nongrad-Nongrad</td>
<td>5.047</td>
<td>0.649</td>
</tr>
<tr>
<td>Nongrad-Grad</td>
<td>5.265</td>
<td>0.372</td>
</tr>
<tr>
<td>Grad-Grad</td>
<td>5.513</td>
<td>0.188</td>
</tr>
</tbody>
</table>

This suggests that there is a difference in the strategies of the two groups of people. The conditional significance of the Cramer-von Mises motivated test (the statistic denoted by $Q_2$ in Chapter 3) is 0.08. This means that among all of
the \(\binom{14}{7}\) different ways of partitioning this group of 14 people into two groups of size 7, only eight percent of these produce a value of this test statistic which is more significant than this one.

Example 3. Are Successive Gaps Between Primes Independent?

At age 15, Gauss conjectured that the number of primes less than \(n\), say \(\pi(n)\), satisfies \(\pi(n)\log(n)n^{-1} \to 1\). This is now called the Prime Number Theorem after being proven by Hadamard and de la Vallée Poussin.

Heuristic arguments for conjectures about primes numbers can be made by pretending that every integer \(n\) is prime independently with probability \(\frac{1}{\log(n)}\). For example, Polignac's conjecture states that for every positive integer \(k\), there are an infinite number of consecutive primes which differ by \(2k\). A special case \((k=1)\) of this conjecture is the Twin Prime Conjecture. Let \(A_n\) be the event that \(2nk+1\) and \(2nk+1+2k\) are consecutive "random prime numbers". These events are independent and the series \(\sum_{n=1}^{\infty} P(A_n)\) diverges. So, the Borel-Cantelli lemma implies \(A_n\) occurs infinitely often almost surely. This is a convincing argument that \(A_n\) occurs infinitely often for the real sequence of prime numbers provided the prime numbers have the properties of a random prime sequence.

The essential property of the random prime sequence is that the standardized gaps between consecutive primes, \(X_1 = \frac{P_2 - P_1}{\log(p_1)}\), ..., \(X_n = \frac{P_{n+1} - P_n}{\log(p_n)}\) are i.i.d. \(\text{Exp}(1)\). Figure 4.5 is a histogram of the 50,000 standardized gaps for the primes following \(10^{50}\). This is consistent with a histogram of marginally
exponential random variables. In order to test whether the gaps are independent, we form d-dimensional vectors which consist of d consecutive standardized gaps and test whether these 50,000/d vectors come from a distribution $G$ which has i.i.d. coordinates with the marginal distribution of the standardized gaps. Although the standardized gaps are very close to being marginally exponential, the edge effects are significant enough that we must rely on the permutation distribution to effectively test for independence. Table 4.2 shows the conditional p-values for $d=2, 3, 4, 5, 6,$ and 7. These are conditional p-values because the significance was obtained by conditioning on the observed standardized gaps and permuting their order to recalculate the statistic a sufficient number of times to estimate the p-value. This procedure legitimately tests for independence among the gaps because the marginal distribution of the gaps is defined to be identical to the marginal distribution under the null hypothesis. This empirical evidence suggests that there may be some form of dependence in the gaps when considered five or seven at a time.
Figure 4.5 Histogram of the standardized gaps between the 50,000 primes following $10^{50}$
Table 4.2  Conditional p-value for the goodness-of-fit test for the hypothesis that d-dimensional vectors of consecutive gaps between primes consist of i.i.d. coordinates.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Conditional P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.86</td>
</tr>
<tr>
<td>3</td>
<td>0.93</td>
</tr>
<tr>
<td>4</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
</tr>
<tr>
<td>7</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Chapter V

Conclusion

Tests of fit for high dimensional data using interpoint distances are intuitively appealing and have been shown to be broadly applicable, powerful, versatile, and have desirable theoretical properties. There are many future research possibilities building on the research described here. In the remainder of this chapter, two of these are outlined.

Researchers often need to determine the effect of changing the values of input variables on the outcome of a stochastic process. When the cost of physically performing all of the desired combinations of input values is expensive, a computer-simulated model of the process is often used. It is necessary to estimate values of parameters used by the algorithm to realistically model the data obtained from experiments. Methods based on the likelihood function work well, but the likelihood function is not always computable. In these cases, there may be a large number of parameters that need to be estimated and a problem-specific search strategy must be used [Maa, Pearl, Bartoszyński, and Horn (1993)].

The proposed optimization algorithm borrows some ideas from the literature on simulation-based parameter estimation [Pakes and Pollard (1989) and Lee and Ingram (1991)]. The directions of random local searches in the
parameter space for the minimizing value are guided by the degree of success of recent searches and a multiple regression fit of the recently investigated portion of the criterion's response surface. To guard against entrapment at a local minimum, the algorithm searches the global parameter space at random times. Unlike simulated annealing, where the criterion function can be evaluated exactly, the algorithm must take into account that the observed value of the criterion is itself based on simulation and thus subject to variability. This difficulty is handled through a cross-validation based procedure.

In addition to simulation-based parameter estimation, another application for a parameter estimation procedure is in the estimation of model parameters when the errors are assumed to be from some parametric family of distributions. In the linear model it is assumed that the response can be modeled as a function of a set of predictor variables, \( Y = \beta^T X + \epsilon \) where \( \epsilon \) has some specified structure. Normally, the relationship between \( X \) and \( Y \) is of more concern, and choosing a model which is faithful to the assumed error structure is of secondary importance to making the observed responses as close as possible to the predicted responses. For example, least squares is generally a reasonable technique even when diagnostics show that the errors are not normal. However, there are instances where choosing a model which is faithful to the distributional structure of the errors is of primary importance. In such cases, we propose to choose the parameters which minimize a goodness-of-fit statistic applied to the residuals, \( Y_i - \beta^T X_i \). Other estimation procedures tend to overfit or incorporate the noise into the estimate of \( \beta \). This type of parameter selection would be most effective in an
ideal situation where the measurement error is the only source of error in the model, this error can be accurately modeled, and the observations are sparse in the sample space.

There are two major goals in this area of the proposal. The first goal is to determine under what conditions the proposed estimates provide consistent estimators of the true parameters. The second goal is to design a method of checking the validity of the assumption that the true model lies within the specified parametric family. The triangle test can be used to provide a measure of fit of the observed data to simulated data using the estimated parameters. The significance level of the goodness-of-fit test can be calculated by using the permutation distribution, but it must be adjusted because of the estimation procedure involved, possibly using the intersection-union principle [Berger (1982)].

Many procedures used in these two areas of statistical methodology are not effective for high-dimensional data. In both problems, there are i.i.d. random vectors \( \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim F \) and \( \mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \sim G \) which are not observed directly. We observe random vectors \( \mathbf{Z}_i = \delta_i \mathbf{X}_i + (1-\delta_i) \mathbf{Y}_i \) where the group membership is denoted by \( \delta_1, \delta_2, \ldots, \delta_n \). It is assumed that there is an unknown transformation \( \theta \), which is in a parametric family of transformations, such that \( \theta(\mathbf{Y}_1) \) is equal in distribution to \( \mathbf{X}_1 \). A simple example would be when the two populations differ only in location.

The classification problem assumes that the first \( n-1 \) values of \( \delta_i \) are known and the goal is to design a rule which will determine \( \delta_n \). The most common approach is Fisher's linear discriminant analysis [Fisher (1936)] which is derived from the assumption that both populations are normal.
Methods using neural networks and regression trees also appear in the literature [Ripley (1994)]. For a given value of $\delta_n$, let $S_1 = \{i: \delta_i = 1\}$. A two-sample statistic is calculated from the elements in $\{Z_i: i \in S_1\}$ and the elements in $\{\theta(Z_i): i \notin S_1\}$. The parameter estimation procedure is then used to search for a transformation and a value of $\delta_n$ which minimizes the two-sample criterion.

The goodness-of-fit procedure described in Chapter 2 will be used to check the model assumptions. It must be modified because, unlike the situation where the distribution is completely specified by the parameters, here it is only assumed that the parameter defines the transformation that relates $G$ to $F$. To test if the family actually contains a transformation that carries $G$ to $F$, we again use the triangle statistic with two vertices from the first sample and the third vertex from the second sample after the transformation has been applied. The permutation or bootstrap distribution of this statistic can be used to provide a measure of significance.

For the dissection problem, all values of $\delta_i$ are unknown and the goal is to determine all $\delta$. As in the classification problem, the solution is completed by searching for a transformation and dissection which minimizes the two-sample criterion.

Cross-validation can be used in the estimation procedure for both problems to protect against overfitting and to provide a measure of the efficacy of the classification rule. The proposed research would include finding the asymptotic efficiency of the procedures described. This is defined by comparing the expected number of correct classifications using the
proposed procedure with the corresponding number using the optimal rule assuming both populations are known.

In summary, methods based on interpoint distance distributions have many applications for statistical problems. Since the testing of distributional assumptions is at the heart of nearly all statistical procedures, the proposed areas of future research are just two among many possible applications of the previous work. Other possible applications not described here include multiple comparison procedures, surface metrology, tests for independence in contingency tables, and image analysis.
REFERENCES


Interpoint Distance Methods for the Analysis of High Dimensional Data

By

John Lawrence, Ph. D.
The Ohio State University, 1996
Professor Dennis K. Pearl, Advisor

Suppose $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ are independent high-dimensional random vectors from populations with distribution $F$ and $G$ respectively. Many approaches for comparing two high-dimensional distributions are increasingly less effective as the dimension increases. In this dissertation, new approaches based on interpoint distances are surveyed. If $\rho$ is any appropriately defined distance, then the $\binom{m}{2}$ "X-to-X interpoint distances" are defined by $\rho(X_i, X_j)$ where $1 \leq i < j \leq m$. Similarly, one can define the X-to-Y interpoint distances and the Y-to-Y interpoint distances. Although some of the information from the original samples may be lost, theoretical properties and Monte Carlo studies show that tests based upon these interpoint distances can be effective in reducing the dimensionality of the problem. Moreover, the comparison of the two distributions can be made even easier with minor assumptions on the two distributions since the interpoint distances are approximately normal when the dimension is very
large. A consistent nonparametric two-sample test statistic and a goodness-of-fit test statistic are discussed. Finally, numerous examples are presented to illustrate the discussion.