

# A Test for Multivariate Structure

Fred W. Huffer  
Florida State University

Cheolyong Park  
Keimyung University

## Abstract

We present a test for detecting ‘multivariate structure’ in data sets. This procedure consists of transforming the data to remove the correlations, then discretizing the data and finally, studying the cell counts in the resulting contingency table. A formal test can be performed using the usual chi-squared test statistic. We give the limiting distribution of the chi-squared statistic and also present simulation results to examine the accuracy of this limiting distribution in finite samples. Several examples show that our procedure can detect a variety of different types of structure. Our examples include data with clustering, digitized speech data, and residuals from a fitted time series model. The chi-squared statistic can also be used as a test for multivariate normality.

*Key words and phrases:* Chi-squared statistic, data-dependent cells, testing for independence, testing for multivariate normality, clustering, time series residuals.

**Corresponding Author:** Fred W. Huffer, Dept. of Statistics, Florida State University, Tallahassee, Florida 32306-4330.

e-mail: [huffer@stat.fsu.edu](mailto:huffer@stat.fsu.edu), phone (850)644-6696, fax (850)644-5271

Cheolyong Park, Dept. of Statistics, Keimyung University, Taegu, 704-701, Korea.

e-mail: [cypark1@kmucc.keimyung.ac.kr](mailto:cypark1@kmucc.keimyung.ac.kr)

# 1 Introduction

Suppose we have multivariate data  $y_1, y_2, \dots, y_n$  consisting of  $n$  points in  $p$  dimensions. In this paper we propose a test statistic that can help in detecting the existence of structure in the data which may not be readily apparent or easily discovered by other means. Our statistic is easily and rapidly computed, and we envision its use as part of the initial phase of the exploratory analysis of raw data or the examination of residuals from fitted models.

We now briefly describe our general approach and the particular test statistic we are proposing. Given the data  $y_1, y_2, \dots, y_n$ , we first employ a linear transformation to remove the sample correlations between the  $p$  coordinates and standardize each coordinate to have mean zero and variance one. (This is often referred to as “sphering” the data.) After transforming the data, we test the hypothesis of independence of the coordinates by discretizing each coordinate and analyzing the resulting categorical data as a contingency table. More precisely, we discretize each of the  $p$  coordinates by using sample quantiles to “bin” or group the values of each coordinate into  $d$  groups of equal size. We then compute the cell counts in the resulting  $p$ -way contingency table. We compare the cell counts with those expected under independence and, if a formal test statistic is desired, we employ the usual chi-squared test of independence. If we find evidence of dependence in the contingency table, we take this as evidence of structure in the data set. The chi-squared statistic (denoted  $X^2$  below) can serve us as a rough overall measure of the amount of structure in the data.

The chi-squared statistic  $X^2$  is easily and rapidly computed, even for large data sets containing many variables, and could be used as part of the initial phase of the exploratory analysis of raw data or the examination of residuals from fitted models. That is, the statistic  $X^2$  could be used as part of a battery of techniques which are all “quick and easy” in the sense of requiring relatively little human and computer time. For example, during the initial examination of multivariate data, one might use  $X^2$  in addition to examining histograms for each of the variables, the sample correlation matrix, and bivariate scatter-plots for all pairs of the variables (and maybe other items as well). The hope is that our statistic  $X^2$  might reveal structure that is missed by these other techniques. If some structure is found, one might then go on to a second phase of analysis employing techniques which are more intensive in their use of human or computer time in order to better understand the nature of this structure. However, after the initial phase of the analysis, the user may decide the data appears to have a simple structure (it may look like a sample from a multivariate normal population or a population with independent variables) and that no further examination is needed.

The considerations above largely determine the form of our test statistic  $X^2$ . We spherize the data in part because we assume that the user of our statistic will also be examining the sample correlations between the variables and will thus be aware of any correlations that do exist. Binning each of the spherized variables by use of the sample quantiles into  $d$  groups of equal size essentially removes the structure or information contained in the marginal distributions. We do this because there are easy and well known techniques (histograms, Q-Q plots, etc.) for studying univariate marginal distributions. We assume that the user will be applying these techniques to study the marginal distributions of the original data, and we note that the same techniques can also be applied to study the marginals of the spherized data. Thus, by sphering and binning the data we hope to remove that part of the structure that the user is likely to be aware of already or can easily study by other means. The structure that remains in the resulting contingency table

is now more likely to correspond to structure in the original data that was previously unknown to the user and not readily apparent. We test for the existence of this structure by using the classical chi-squared test for independence in a contingency table.

In Section 4 we present a number of examples in which this procedure is used to detect structure in data. We examine data which consists of randomly located clusters, data arising from digitized speech, and data consisting of the output of a faulty random number generator. We also use our procedure to examine the residuals from a fitted time series model. The chi-squared statistic is able to detect a wide variety of different types of structure. It can often find structure in situations where it is not very apparent and could be easily missed by a data analyst.

As a general guide to judging the magnitude of  $X^2$  we use its limiting distribution (as the sample size  $n$  becomes large) when sampling from a multivariate normal population. This is given in our Theorem 1. After spherizing to remove the correlations, it seems reasonable to regard the multivariate normal distribution as having no remaining structure. Thus, the distribution of  $X^2$  for a multivariate normal population is a reasonable choice for a “reference” or “null” distribution. Another situation to consider is data sampled from a population which has independent coordinates. In this situation the sample correlations between the variables will be small and the spherizing transformation (typically) amounts to a small perturbation of the original coordinate system. Thus, the coordinates of the spherized data will be approximately independent and the contingency table of counts will usually reveal no evidence of dependence. So, the value of  $X^2$  will tend to be small in this situation. In fact, we show (via simulation, see Section 3.2) that the distribution of  $X^2$  is roughly the same for both multivariate normal populations and populations with independent coordinates. Throughout this paper, we shall regard both of these as “null” situations in which there is no structure in the data beyond that in the correlations and marginal distributions.

Our procedure is often very effective at signaling the existence of multivariate structure, but usually gives little information about the nature of that structure. When structure is found, one may need to employ other methods (e.g., projection pursuit, cluster analysis or dynamic graphical techniques) to discover the nature of this structure. We also note that our procedure cannot detect *all* types of multivariate structure, but only those kinds of structure which reveal themselves as some type of dependence between the coordinates of the spherized data. It is possible to construct examples of data sets which contain obvious structure which is not detected by our chi-squared statistic. Of course, since “structure” is such a vague and slippery concept, it seems unreasonable to expect any procedure to detect all possible types of structure.

It should be noted that we are not attempting to use the word “structure” as a precise, technical term. We call our procedure a “test for structure” mainly to emphasize the role we hope it will play in applications. If one wishes to regard our procedure as a test of a more formal statistical hypothesis, one can think of it as a test for dependence in the spherized coordinates. Our Theorem 1 then provides the appropriate adjustment to the null distribution to account for the fact that the test is carried out on the spherized data and not on the original data.

Since we know the approximate distribution of  $X^2$  when sampling from a multivariate normal population, the statistic  $X^2$  can also be used to test the hypothesis of multivariate normality. The resulting test is not an omnibus test. For example, the test has no sensitivity to non-normality in the marginal distributions of the spherized data. However, our chi-squared statistic is based on different principles than the existing procedures currently in use, and it is sensitive to different types of departures from multivariate normality. (See Sections 4.2 and 4.4.) Thus, our statistic

should be useful as part of a battery of tests for multivariate normality.

The remainder of the paper is organized as follows. In Section 2 we give a precise definition of the chi-squared statistic. We also give the limiting distribution of this statistic when our data is sampled from a multivariate normal population. In Section 3 we use simulation studies to examine the “null” behavior of the chi-squared statistic. That is, we study the distribution of the statistic in those situations (multivariate normal populations and populations with independent coordinates) we regard as being without structure. We find that the limiting distribution derived for a multivariate normal population offers a good general approximation to the null distribution of our statistic. In Section 4 we provide a number of examples to illustrate how our procedure can be used in practice.

## 2 The Chi-Squared Statistic

Suppose we have data  $y_1, y_2, \dots, y_n$  which are  $p \times 1$  vectors. Let the data matrix  $Y$  be the  $n \times p$  matrix whose  $i$ -th row is  $y_i$ .

To look for structure in  $Y$ , we employ the following procedure. First, we apply a linear transformation to “sphere” the data. This creates a transformed data set  $Z$  in which the coordinates (columns) are uncorrelated and have mean zero. More formally, the  $n \times p$  matrix  $Z = (z_{ij})$  of transformed data is defined by

$$Z = Q_e Y R(S), \tag{1}$$

where  $Q_e = I_n - ee^t/n$  and  $R(S)$  is a  $p \times p$  matrix chosen so that  $Z^t Z/n = I_p$ . Here we use  $I_n$  and  $I_p$  to denote identity matrices with the indicated dimensions, and  $e$  to denote a column vector of ones. We require the matrix  $R(S)$  to be a function of the sample covariance matrix  $S$  defined by  $S = n^{-1} Y^t Q_e Y$ . If we let  $z_i$  denote the  $i$ -th row of  $Z$ , we can write our transformation as  $z_i = R^t(y_i - \bar{y})$  for  $i = 1, \dots, n$ , where  $\bar{y}$  is the sample mean vector  $\bar{y} = n^{-1} Y^t e$ . Transformations of this type are frequently employed in statistics, and in particular, have often been used in the construction of tests for multivariate normality.

There are many possible choices for the function  $R = R(S)$ . Any choice satisfying  $R^t S R = I$  will give  $Z^t Z/n = I$ . A principal components transformation of the data  $Y$  corresponds to choosing a particular matrix  $R$  of the form  $\Gamma D$  where  $\Gamma$  is an orthogonal matrix and  $D$  is a diagonal matrix. A Gram-Schmidt transformation takes  $R$  to be upper triangular. Another commonly used transformation uses  $R = S^{-1/2}$ . In our work, it is important that the matrix  $R$  be chosen in a way which depends only on  $S$  and not directly on the raw data  $Y$ ; this is required for the validity of Theorem 1. Also, as a general rule, we recommend using transformations which are continuous as a function of  $S$  and satisfy  $R(D) = D^{-1/2}$  for any diagonal matrix  $D$ . (This point is discussed in Section 3.2.) The Gram-Schmidt transformation and  $R(S) = S^{-1/2}$  satisfy this rule, but the principal components transformation does not.

After obtaining the transformed data  $Z$ , we discretize each column of  $Z$  by dividing the values in each column into  $d$  groups (labeled  $1, 2, \dots, d$ ) of equal size  $n/d$ . If  $n$  is not divisible by  $d$ , the group sizes will not be exactly equal. This produces an  $n \times p$  matrix  $T = (t_{ij})$  whose entries  $t_{ij}$  are all integers in  $\{1, 2, \dots, d\}$ . A more precise definition of  $T$  is given by

$$t_{ij} = k, \quad \text{if } (k-1)n/d < r_{ij} \leq kn/d, \tag{2}$$

where  $r_{ij}$  is the rank of  $z_{ij}$  among the values  $z_{1j}, z_{2j}, \dots, z_{nj}$  in the  $j$ -th column.

We now form a contingency table from the  $n$  rows of the discretized matrix  $T$ . This contingency table contains  $d^p$  cells corresponding to the possible  $p$ -tuples of integers in  $\{1, 2, \dots, d\}$ . We have  $n$  observations distributed among these  $d^p$  cells. Under the null hypotheses that we consider, the expected number of observations in any given cell is approximately  $n/d^p$ . We use  $\pi = (\pi_1, \pi_2, \dots, \pi_p)$  with  $1 \leq \pi_i \leq d$  for all  $i$  to denote a particular cell in our table. For each cell  $\pi$ , the cell count  $U_\pi$  is given by

$$U_\pi = \sum_{i=1}^n I\{t_i = \pi\}, \quad (3)$$

where  $t_i$  is the  $i$ -th row of  $T$ . Some information about the structure in the data set  $Y$  can be gleaned from a direct examination of the distribution of the cell counts; see the examples in Section 4. As a summary measure for the amount of structure in the data (or for the degree of departure from multivariate normality), we use the chi-squared statistic  $X^2$  defined by

$$X^2 = \sum_{\pi} \frac{(U_\pi - n/d^p)^2}{n/d^p}. \quad (4)$$

This statistic can be rapidly computed even for very large data sets.

When sampling from a multivariate normal distribution, the limiting distribution of  $X^2$  (given below) is that of a weighted sum of independent chi-square random variables with appropriate degrees of freedom. In most applications, the number of cells  $d^p$  is fairly large. In this case, the limiting distribution is approximately normal and the  $z$ -score

$$z = \frac{X^2 - \mu_{x^2}}{\sigma_{x^2}} \quad (5)$$

can be used to give a simple test for structure. Here  $\mu_{x^2}$  and  $\sigma_{x^2}$  are the mean and the standard deviation of the limiting distribution of  $X^2$ .

The choice of  $d$  is somewhat arbitrary. In exploratory work we often try many different values of  $d$  since we do not know in advance what type of structure there might be in the data and on what scale this structure might be most easily observed. We generally prefer to have a fairly large number of cells and, at the same time, an average cell count  $n/d^p$  which is not too small. If we wish to use the limiting distribution of the chi-squared statistic for testing purposes, our simulation work seems to indicate that the usual guidelines apply: the limiting distribution is fairly accurate when  $n/d^p \geq 5$ . If the number of cells is sufficiently large, it is reasonably good even for  $n/d^p = 1$ . Since  $d^p$  grows rapidly with  $p$ , for high dimensional data sets we are often forced to use small values of  $d$  in order to avoid extremely small average cell counts.

The following theorem gives the limiting distribution of the chi-squared statistic  $X^2$  when the data  $Y$  is sampled from a multivariate normal population. A detailed proof of this result may be found in Huffer and Park (1999).

Let  $\phi$  and  $\Phi$  denote the density and cdf of the standard normal distribution. For  $i = 0, 1, \dots, d$ , we define  $\zeta_i = \Phi^{-1}(i/d)$ . Note that  $\zeta_0 = -\infty$  and  $\zeta_d = \infty$ . Now define

$$\psi_i = \phi(\zeta_{i-1}) - \phi(\zeta_i) \text{ for } 1 \leq i \leq d, \text{ and } c = \left( \sum_{i=1}^d \psi_i^2 \right)^2 \quad (6)$$

with the convention  $\phi(\pm\infty) = 0$ .

**Theorem 1** *If  $y_1, y_2, \dots, y_n$  are i.i.d.  $N(\mu, \Sigma)$  with  $\Sigma$  nonsingular, then*

- (a) *The distribution of  $X^2$  does not depend on  $\mu$  or  $\Sigma$  (that is,  $X^2$  is ancillary), or on the choice of the transformation  $R(S)$ .*
- (b) *As  $n \rightarrow \infty$ , the distribution of  $X^2$  converges to that of  $W_1 + (1 - d^2c)W_2$  where  $W_1$  and  $W_2$  are independent chi-squared variates with degrees of freedom  $\nu_1 = d^p - 1 - p(d - 1) - p(p - 1)/2$  and  $\nu_2 = p(p - 1)/2$  respectively.*

Distributions like that in part (b) of our Theorem have been well known in the context of chi-squared tests since the work of Chernoff and Lehmann (1954). Our results are similar in character to those of Watson (1957) dealing with goodness-of-fit for the univariate normal distribution. However, we note that our statistic  $X^2$  does not have a precise univariate analog; when  $p = 1$  the statistic  $X^2$  is degenerate (constant with probability one).

The limiting distribution does not have a convenient closed form for either the density or the cdf, but it is still possible to obtain a great deal of information about this distribution. For example, it is routine to compute moments and cumulants of all orders. The mean and variance needed in (5) are given by

$$\mu_{x^2} = \nu_1 + (1 - d^2c)\nu_2 \quad \text{and} \quad \sigma_{x^2}^2 = 2\nu_1 + 2(1 - d^2c)^2\nu_2. \quad (7)$$

Weighted sums of chi-squared variates arise frequently in statistics and there has been much work on obtaining numerical approximations to their distributions. The cdf may be evaluated by numerical inversion of the characteristic function (Imhof (1961), Farebrother (1990)). There are also good approximations based on matching moments (Solomon and Stephens (1977)). Finally, we note that it is easy to simulate from the limiting distribution, so that many questions can be given quick approximate answers via simulations.

Park (1992) studies a number of closely related chi-squared statistics which are arrived at by using different initial transformations and methods of discretization than those in (1) and (2). He obtains results analogous to Theorem 1 for these statistics.

### 3 Simulations of Null Behavior

In this section, we present simulation results to illustrate the validity and accuracy (in finite samples) of the limiting distribution of  $X^2$ . We consider two situations: sampling (1) from a multivariate normal population, and (2) from populations with independent coordinates. We find that the limiting distribution in Theorem 1 gives a good approximation to the true distribution of  $X^2$  in both of these situations.

#### 3.1 Sampling from the Multivariate Normal Distribution

We have performed numerous simulations to study the distribution of  $X^2$  when sampling from a multivariate normal population. All gave very similar results. In this section we present the results of one such study. Part (a) of Theorem 1 states that, for given values of  $n$ ,  $d$  and  $p$ , the distribution of  $X^2$  is the same for all choices of  $\mu$ ,  $\Sigma$ , and method of transformation  $R(S)$ . For the simulations described below, we take  $\mu = 0$ ,  $\Sigma = I$ , and choose the Gram-Schmidt

transformation (taking  $R(S)$  to be upper triangular). Our simulated data matrices  $Y$  are thus simply matrices whose entries are i.i.d. standard normal random variables.

In these simulations we take the number of coordinates  $p$  to be four, and the number of categories  $d$  to be three. Thus, our  $X^2$  statistics are computed from a contingency table of counts which has  $3^4 = 81$  cells. In this situation, the limiting distribution given in Theorem 1 becomes  $\chi^2(66) + 0.3708\chi^2(6)$ . We shall consider three different samples sizes,  $n = 81, 405$  and  $810$ , which we refer to as small, moderate and large samples respectively. These sample sizes correspond to having an average of 1, 5 and 10 observations per cell respectively.

For each sample size  $n$ , we generated 500  $n \times 4$  matrices  $Y$  and computed the value of  $X^2$  for each of them. These 500 values were ordered and then plotted against the expected order statistics (see the remarks below) of a sample of size 500 from the limiting distribution. The resulting quantile-quantile plots are displayed in Figure 1. Each of our quantile-quantile plots displays the reference line having slope 1 and intercept 0. This represents the “ideal” case in which the empirical and theoretical distributions coincide. Examining the plots, we see that the limiting distribution is a good approximation in the moderate and large sample cases. The discreteness of the  $X^2$  statistic is apparent in the small sample case. Also, in this case the actual distribution is somewhat less dispersed than the limiting distribution. However, we feel that the limiting distribution fits well enough to serve as a useful rough approximation.

## Position of Figure 1

The “expected order statistics” (labeled as “theoretical quantiles”) we use in our plots are approximations obtained as follows: It is straightforward to generate random variates from any distribution expressible as a weighted sum of chi-squared variates. Thus, we simply generated 100 samples of size 500 from the limiting distribution and averaged the order statistics of these 100 samples to obtain estimates of the expected order statistics. We found that 100 samples give a reasonably accurate estimate.

Finally, we compare the sample moments of  $X^2$  in the small, moderate, and large sample cases to those from the limiting distribution in Table 1. The sample mean and standard deviation are

Table 1: Sample moments of  $X^2$  from the small, moderate, and large sample sizes and those from the limiting distribution

	small sample	moderate sample	large sample	limiting dist.
mean	69.06	68.07	67.59	68.22
s.d.	10.40	11.16	11.05	11.56

quite close to those from the limiting distribution except for a possibly under-estimated sample standard deviation for the small sample case. Thus, this table confirms the findings in the quantile-quantile plots.

### 3.2 Distributions with Independent Coordinates

When our data is sampled from a population which is *not* multivariate normal, the situation becomes complicated. We no longer have an invariance result like part (a) of Theorem 1, and the distribution of  $X^2$  will typically depend on both the parent population and on the particular choice of  $R(S)$ . We regard multivariate normal distributions and distributions with independent coordinates to be equally lacking in structure and would prefer that our  $X^2$  test not distinguish between these two situations. Our simulations indicate that, in fact, this is roughly the case. With an appropriate choice of the transformation  $R(S)$ , the limiting distribution of Theorem 1 continues to be approximately valid for distributions with independent coordinates. Another way to state this conclusion is the following: If we regard  $X^2$  as a statistic for testing the hypothesis of multivariate normality, it will give a test which has low power not only for alternatives close to the multivariate normal distribution, but also for alternatives for which the coordinates are close to being independent.

Before presenting our simulation results, we give an informal argument indicating why we expect that the limiting distribution of  $X^2$  under independent coordinates will *not* be radically different from the distribution under multivariate normality, at least when  $R(S)$  is appropriately chosen. Suppose the transformation  $R(S)$  is “smooth” as a function of  $S$ . Assume also that  $R(D) = D^{-1/2}$  for any diagonal matrix  $D$ . Both the Gram-Schmidt transformation and  $R(S) = S^{-1/2}$  satisfy these assumptions. Let the data  $y_1, y_2, \dots, y_n$  be i.i.d. from a  $p$ -variate distribution with independent coordinates. The covariance matrix  $\Sigma$  will then be diagonal. As the sample size  $n$  goes to infinity, we will have  $S \rightarrow \Sigma$  so that our assumptions on  $R$  ensure that  $R(S) \rightarrow R(\Sigma) = \Sigma^{-1/2}$ . It then seems reasonable that the limiting distribution of  $X^2$  will be not too different from that of the related chi-squared statistic  $\tilde{X}^2$  constructed using the fixed matrix  $\tilde{R} = \Sigma^{-1/2}$  in place of  $R(S)$  in equation (1). But the statistic  $\tilde{X}^2$  is essentially identical to the standard chi-squared test for independence in contingency tables and it is not hard to see it has the usual  $\chi^2(d^p - 1 - p(d - 1))$  limiting distribution. (See Park (1992) for a proof of this assertion.) When the number of cells  $d^p$  is sufficiently large, this distribution will be close to the limiting distribution given in Theorem 1. This gives us our desired conclusion.

We now present our simulation results. In all of the simulations we now describe, we take  $p = 4$ ,  $d = 3$  and  $n = 405$ ; this is the “moderate sample” case used earlier. Each value of  $X^2$  is computed from a  $405 \times 4$  matrix whose entries are i.i.d. from a specified parent distribution. Four different parent distributions are used: the normal distribution, and three different log-normal distributions with increasing degrees of skewness. To be more precise, a log-normal random variate  $y$  is generated as  $y = e^X$  where  $X \sim N(0, \sigma^2)$  and  $\sigma^2$  takes on one of the three values 0.1, 0.5 or 1.0. These values of  $\sigma^2$  produce values of the standardized skewness  $\gamma_1 = E(y - \mu)^3 / (E(y - \mu)^2)^{3/2}$  equal to 1.01, 2.94 and 6.18 respectively, corresponding to moderate, large, and very large amounts of skewness. For convenience, we refer to the four parent distributions by number as 0, 1, 2, 3. We note that distribution 3 has both very large skewness and a very heavy right tail.

Figure 2 gives a number of boxplots each summarizing the distribution of 200 values of  $X^2$ ; the box gives the median and quartiles, and the whiskers indicate the 5% and 95% points of the distribution. The boxplots are divided into three groups according to the transformation  $R(S)$  used: Gram-Schmidt, symmetric ( $R(S) = S^{-1/2}$ ), or principal components. In each group, the four boxplots represent the distribution of  $X^2$  under the four parent distributions 0 – 3.

Distribution 0 is provided as a reference point. We see that, for the Gram-Schmidt and symmetric transformations, the introduction of moderate amounts of skewness (distribution 1) has little impact on the distribution of  $X^2$ . Even a large amount of skewness (distributions 2 and 3) has fairly modest effects. The situation is radically different for the principal components (PC) transformation. Here even a moderate amount of skewness produces a substantial change in the distribution of  $X^2$ . This is because the PC transformation does not satisfy the condition  $R(D) = D^{-1/2}$  mentioned above. In our simulation setting, if we let  $n \rightarrow \infty$ , the PC transformation produces a matrix  $R(S)$  which converges in distribution to a scalar multiple of a random orthogonal matrix;  $R(S)$  does *not* converge to the “correct” transformation.

**Position of Figure 2**

We have obtained similar results in other simulations using parent distributions different from the log-normal. For example, we have investigated the case where the entries in  $Y$  are i.i.d. uniform random variables. In this case, the distribution of  $X^2$  is virtually indistinguishable from the limiting distribution in Theorem 1 when we use the Gram-Schmidt transformation or  $R(S) = S^{-1/2}$ , but is radically different when we use the PC transformation. In conclusion, when using  $X^2$  as a test for multivariate structure, one should use either the Gram-Schmidt transformation or  $R(S) = S^{-1/2}$ . When this is done, the limiting distribution in Theorem 1 offers a reasonable guide for using  $X^2$ . The PC transformation should probably be avoided. (If your goal is the more narrow one of testing for multivariate normality, then there is no longer any reason to exclude the PC transformation.)

As a practical matter, when applying our  $X^2$  test to data  $Y$  having columns whose distributions are highly nonnormal, it is probably a good idea to first transform the columns to make them approximately normal. There are a couple of reasons for this. First, our procedure uses the sample covariance matrix  $S$  which can be highly variable for heavy-tailed distributions. Secondly, it seems likely that transforming the columns will make the null distribution of  $X^2$  closer to the distribution it would have for multivariate normal populations.

## 4 Examples

We now present examples to show how our procedure might be used in applications. Until now, our discussion has dealt exclusively with the chi-squared statistic  $X^2$ . In our examples, we give the value of  $X^2$ , but we also present additional information summarizing the observed distribution of the cell counts  $U_\pi$  defined in (3). To aid in interpreting this summary information, we introduce a simple Poisson approximation. We wish to explain and illustrate this new material on data *without* any structure (to observe the “null” behavior) before using it in examples with structure. For this reason, our first example will use data generated from a multivariate normal distribution.

For all the examples which follow, we shall use the Gram-Schmidt transformation as our choice for  $R(S)$ .

## 4.1 Sampling from the Multivariate Normal Distribution

In this example,  $Y$  is a  $1215 \times 5$  matrix composed of independent columns generated from the standard normal distribution. Our procedure leads to the output in Table 2. We have chosen to set  $d = 3$ ; this means we have divided the data space into  $d^p = 3^5 = 243$  cells. There are  $n = 1215$  observations, so that the average number of observations per cell is  $n/d^p = 5$ . The last

Table 2: Output from a normal distribution

```
*****
For d = 3,
The frequency distribution of the cell counts is:
      0    1    2    3    4    5    6    7    8
Observed 2.00 7.00 16.00 35.00 51.00 48.00 26.00 25.00 13.00
Expected 1.64 8.19 20.47 34.11 42.64 42.64 35.53 25.38 15.86

      9   10 11   12   13   14   15   16
Observed 13.00 4.00  1 2.00 0.00 0.00 0.00 0.00
Expected  8.81 4.41  2 0.83 0.32 0.11 0.04 0.01

The moments of the distribution of cell counts are:
      mean variance skewness kurtosis
Observed   5   4.8642  0.51786  0.19229
Expected   5   5.0000  0.44721  0.20000

Observed X^2 value =   236.4
Asymptotic mean and s.d. of X^2 =  225.71 21.14
z-score for X^2 =  0.51
*****
```

three lines of the output give the value of  $X^2$ , the mean and standard deviation of the limiting distribution in Theorem 1, and the  $z$ -score computed as in equation (5). The  $z$ -score of 0.51 would lead to our concluding that there is no structure in this data. This agrees with the known truth in this case.

The output in Table 2 also lists the “observed” frequency distribution: two cells are empty, seven cells contain exactly one observation, 16 cells contain exactly two observations, etc. Let  $N_k$  be the number of cells containing exactly  $k$  observations, that is,  $N_k = \sum_{\pi} I\{U_{\pi} = k\}$ . As a rough standard for comparison, the output gives an “expected” frequency distribution computed using a simple Poisson approximation:  $N_k$  is compared with  $E_k = d^p \lambda^k e^{-\lambda} / k!$  where  $\lambda = n/d^p$ . The output summarizes the observed distribution of cell counts by giving the sample moments: the mean, variance, standardized skewness ( $\mu_3/\sigma^3$ ), and standardized kurtosis ( $\mu_4/\sigma^4 - 3$ ). These are compared with the corresponding moments of the Poisson distribution with mean  $\lambda$  which are labeled the “expected” moments.

The Poisson approximation is based on the following rationale. In most of the applications of our methods, the number of cells  $d^p$  is quite large. The cells are (at least approximately) equally

likely, that is, an observation (row of  $Y$ ) has an approximate probability  $1/d^p$  of belonging to any given cell. Moreover, the  $n$  observations are roughly independent with regard to their cell membership. (They are not exactly independent, the initial transformation (1) and the method of discretization (2) impose some dependence.) We have a large number  $n$  of observations, each with a small probability  $1/d^p$  of belonging to any given cell  $\pi$ . Thus, we expect the number of observations  $U_\pi$  belonging to cell  $\pi$  to have approximately a Poisson distribution with a mean of  $\lambda = n/d^p$ . The values  $U_\pi$  should behave roughly like a random sample of size  $d^p$  from a Poisson distribution with mean  $\lambda$ . This implies that the number of cells  $N_k$  containing exactly  $k$  observations should have approximately a binomial distribution with mean  $E_k = d^p P_k$  and variance  $V_k = d^p P_k(1 - P_k)$  where  $P_k = \lambda^k e^{-\lambda}/k!$ . Our output lists the observed values  $N_k$  and the expected values  $E_k$ .

The observed frequency distribution of the cell counts in Table 2 is close to the expected frequency distribution. Similarly, the observed moments are close to the expected moments. This has been our general experience; the Poisson approximation fairly accurately describes the distribution of the cell counts when there is no structure in the data and the number of cells  $d^p$  is large. See the technical report by Huffer and Park (1999) for a simulation study of the Poisson approximation.

When we use our procedure on real data, a large value of  $X^2$  indicates there is structure in the data, but tells us nothing about the type of structure. Comparing the frequency distribution of the observed cell counts with the “expected” distribution gives us some information concerning the type of structure. In particular, we can see whether the large  $X^2$  is due to just a few cells with very large counts (perhaps due to a single clump in the data), or whether it reflects a more global change in the frequency distribution (suggesting a more extended form of structure).

In very large samples, the limiting distribution of  $X^2$  may no longer be useful for testing; it will often detect structure which is statistically significant, but too small to be of practical importance. In this situation, it may be useful to rescale the  $X^2$  statistic so that its magnitude is a meaningful measure of the degree of structure in the data. The “observed” variance of the cell counts, which equals  $(n/d^{2p}) \times X^2$ , is a useful rescaling. For very large samples, an informal comparison of the “observed” and “expected” variance of the cell counts may be preferable to a formal test based on the limiting distribution of  $X^2$ . (The Poisson approximation for the “expected” variance of the cell counts is simply  $\lambda = n/d^p$ .)

## 4.2 An Example with Randomly Located Clusters

We now consider a data set consisting of many randomly located clusters in dimension  $p = 5$ . There are  $n = 405$  observations made up of 135 clusters of size 3. The cluster centers (denoted  $\mu_1, \mu_2, \dots, \mu_{135}$ ) are independently generated from  $N(0, I_5)$ . The members of cluster  $i$  are generated from  $N(\mu_i, \sigma^2 I_5)$ , with  $\sigma = 0.25$ . In bivariate scatter-plots, there is no obvious structure in the data set. However, our method clearly signals the existence of structure.

Applying our method with  $d = 3$  leads to the output in Table 3. The chi-squared statistic is highly significant with a  $z$ -score of 6.13. This large value is caused by the larger than expected number of cells with  $U_\pi = 0$  and  $U_\pi \geq 6$ .

We have experimented with many variants of this example, using different dimensions  $p$ , numbers of clusters, cluster sizes, and cluster dispersions  $\sigma$ . Our procedure does very well at detecting this type of structure. In this example, we chose  $\sigma = .25$  because, with this value,

Table 3: Output from randomly located clusters

```

*****
For d = 3,
The frequency distribution of the cell counts is:
      0      1      2      3      4      5      6      7      8      9
Observed 63.0 73.00 43.00 33.00 18.00 4.00 8.00 1.00 0.00 0.00
Expected 45.9 76.49 63.75 35.41 14.76 4.92 1.37 0.33 0.07 0.01

The moments of the distribution of cell counts are:
      mean variance skewness kurtosis
Observed 1.66667  2.43621  1.02379  0.64896
Expected 1.66667  1.66667  0.77460  0.60000

Observed X^2 value =   355.2
Asymptotic mean and s.d. of X^2 =  225.71 21.14
z-score for X^2 =   6.13
*****

```

there is no structure visible in the bivariate scatter-plots. As one tries smaller and smaller values of  $\sigma$ , the chi-squared statistic becomes more and more sensitive, that is, the values  $X^2$  become progressively larger. However, for sufficiently small values of  $\sigma$  (say, for  $\sigma \leq .10$ ) the clustering becomes fairly evident in the bivariate scatter-plots, so that a procedure such as ours is less necessary.

This example can also be viewed in the narrower context of testing for multivariate normality. The data in this example have univariate marginals which are roughly normal and no apparent structure in the bivariate scatter-plots, so one might suspect the data comes from a multivariate normal population and wish to test this hypothesis. There are many statistics in the literature for testing multivariate normality. These are reviewed by Gnanadesikan (1977, pp. 161–175), Mardia (1980), and Koziol (1986). Romeu and Ozturk (1993) studied the performance of a number of statistics. Three statistics which performed well in their study were the skewness and kurtosis tests of Mardia (1970, 1980) and the  $Q_n$  test (with Cholesky-implementation) of Ozturk and Romeu (1992). We have conducted simulations comparing our statistic  $X^2$  with these three tests on data similar to that in this example. In our simulations,  $X^2$  and Mardia's skewness test did well, and Mardia's kurtosis test and  $Q_n$  did badly. For the exact situation in this example ( $\sigma = .25$ ,  $p = 5$ ), the test based on  $X^2$  is more powerful than Mardia's skewness test. As we vary  $\sigma$ , we find that  $X^2$  is more powerful than Mardia's skewness test for  $\sigma \leq .30$  (very roughly). In our simulations, it seems that something like this holds in general;  $X^2$  is better for 'small'  $\sigma$ , and Mardia's skewness test is better for 'large'  $\sigma$ . However, the cutoff between 'large' and 'small'  $\sigma$  varies with the dimension  $p$ , the number of clusters, and the cluster sizes. In our simulations, the values of  $X^2$  and Mardia's skewness test were essentially uncorrelated. This lends empirical support to the notion that the two statistics are looking at different aspects of the data.

### 4.3 An Example using Speech Data

The data matrix in this example is  $3393 \times 10$ . The data was obtained by sampling from a much larger matrix of digitized speech data consisting of 10 dimensional ‘lpc’ vectors. The lpc vectors in this sample correspond to ‘unvoiced’ sounds.

For our purposes, the exact nature of the lpc vectors is unimportant, but we can give some rough idea of what they are. In digitizing speech, the intensity of speech sounds is recorded at regular intervals of time (say, 10,000 times per second) and the resulting measurements are viewed as a time series. The time series is then broken down into small chunks (sub-series), each representing a fraction of a second of speech. An autoregressive process of order 10 is fit to the data in each chunk. The lpc vector is a one-to-one function of the vector of estimated autoregressive coefficients. Using only the lpc vector, one can fairly accurately reproduce the sound in the chunk. Thus, the sequence of lpc vectors allows us to compress the speech data.

The collection and analysis of this data set was motivated by an attempt to further compress the speech data by quantizing the space of lpc vectors. By “quantizing” we mean breaking down the 10 dimensional space of the lpc vectors into disjoint regions. Then, when recording speech data, we throw away the lpc vectors and record only which regions they lie in. If the regions are chosen appropriately, the remaining information will suffice to approximately reconstruct the speech.

The method one employs to quantize the space of lpc vectors depends on whether or not the distribution of lpc vectors has structure and on the nature of this structure. A procedure such as ours is a helpful first step in this investigation.

Examination of a histogram reveals that the first coordinate of the lpc vectors is highly skewed to the left. Also, the bivariate scatter-plots of the first coordinate versus the other coordinates reveal some definite nonlinear patterns. There is obvious structure involving the first coordinate, so we shall omit this variable and see if any structure exists in the remaining nine. Inspecting a matrix of scatter-plots for variables two through ten reveals no obvious structure or pattern. However, applying our procedure to this  $3393 \times 9$  matrix leads to the output in Table 4.

Examining the frequency distribution, we see there is one cell containing 29 observations and several cells containing more than 20. If there is no structure in this data set, we do not expect to see any cells with more than 20. (The expected frequency is less than .01.) Also there are 78 cells containing 0, 1, or 2 observations, which is much more than expected. In other words, the observed frequency distribution is more dispersed than the expected frequency distribution. This gives strong evidence for the presence of structure in this data. This conjecture is supported by the  $z$ -score and the ratio of the observed variance to the expected variance.

After discovering structure in variables two through ten, we would like to determine the nature of this structure. Since the data we are examining is high-dimensional, this is not an easy task. Our methods can help us by suggesting subsets of the variables on which to focus our attention. In the course of computing the output in Table 4, we calculated and stored the ranks  $r_{ij}$  (see (2)) for the spherized data obtained from variables two through ten. Using this  $3393 \times 9$  matrix of ranks, we computed the  $X^2$  statistic and corresponding  $z$ -score for each of the 501 subsets of two or more columns from this matrix and each of the values  $d = 2, 3, 4, 6$ , producing 2,004  $z$ -scores in all. (This took about 2 minutes on our computer system.) There were a number of low-dimensional subsets of the variables with large  $z$ -scores, a few of which are listed below. These particular subsets are good candidates for more detailed study. Let  $Z_i$  denote the  $i$ -th

Table 4: Output from speech data

```

*****
For d = 2,
The frequency distribution of the cell counts is:
      0      1      2      3      4      5      6      7      8
Observed 9.00 26.00 43.00 57.00 47.00 51.00 47.00 48.00 40.00
Expected 0.68  4.49 14.89 32.89 54.48 72.21 79.76 75.51 62.55

      9      10     11     12     13     14     15     16     17     18
Observed 41.00 28.00 21.00 13.00 10.00 7.00 2.00 5.00 2.00 4.00
Expected 46.06 30.52 18.39 10.15  5.18 2.45 1.08 0.45 0.17 0.06

      19     20 21 22 23 24 25 26 27 28 29
Observed 2.00 1.00 1  1  1  4  0  0  0  0  1
Expected 0.02 0.01 0  0  0  0  0  0  0  0  0

The moments of the distribution of cell counts are:
      mean variance skewness kurtosis
Observed 6.62695 19.09716  1.36521  3.10011
Expected 6.62695  6.62695  0.38846  0.15090

Observed X^2 value =  1475.45
Asymptotic mean and s.d. of X^2 = 487.41, 30.94
z-score for X^2 = 31.93
*****

```

column of the spherized data. The  $X^2$  statistic for  $(Z_1, Z_3, Z_4, Z_5)$  with  $d = 3$  had a  $z$ -score of 49.3. The  $X^2$  statistics for the subsets  $(Z_4, Z_5)$  and  $(Z_3, Z_8)$  with  $d = 4$  had  $z$ -scores of 35.3 and 25.9 respectively. Finding pairs of variables with very large  $z$ -scores like this was somewhat surprising to us, since there was little apparent structure in the matrix of scatter-plots. (Perhaps the small and rather crowded plots which one gets when creating a scatter-plot matrix for a large high-dimensional data set should not be relied upon except to reveal very gross features of the data.)

After locating plausible subsets of the variables, we can then bring other techniques to bear to investigate these subsets. For example, dynamic graphical methods (such as “spinning” and “brushing”) are now widely available in commercial software and provide very natural and intuitive ways to investigate the structure in relatively low-dimensional data sets. Applying these methods to the subsets found above, we find there is a great deal of structure in the data which is associated with the omitted first variable. For example, the non-normality in the scatter-plot of  $Z_4$  versus  $Z_5$  can be largely explained by viewing it as a superposition of two separate groups of points corresponding to the cases with high and low values of the first variable.

## 4.4 Examining a Faulty Random Number Generator

We will now show that our procedure can detect the structure in simulated data which results from the use of a faulty random number generator. RANDU is a linear congruential generator which generates a sequence of integers  $\{V_i\}$  according to the rule  $V_{i+1} = 65539 V_i \bmod 2^{31}$ . Taking  $U_i = V_i/2^{31}$  produces a sequence  $\{U_i\}$  of pseudo-random uniform variates. RANDU has a major defect: Marsaglia (1968) showed that the triples  $(U_i, U_{i+1}, U_{i+2})$  produced by RANDU lie on 15 parallel hyperplanes. Given a sequence of pseudo-random uniform variates  $\{U_i\}$ , the Box-Muller method produces a sequence  $\{Z_i\}$  of pseudo-random normal variates by using the transformation  $(Z_i, Z_{i+1}) = (-2 \log U_i)^{1/2}(\cos(2\pi U_{i+1}), \sin(2\pi U_{i+1}))$  for odd values of  $i$ . We shall use RANDU in combination with the Box-Muller method to generate normal variates. The Box-Muller transformation is highly non-linear and will deform the hyperplane structure produced by RANDU into something very peculiar. We will see if our method can detect this structure; there is no apparent structure in bivariate plots of  $(Z_i, Z_{i+1})$ .

We applied our method to a  $50625 \times 4$  data matrix  $Y$  in which each row consisted of four consecutive normal variates produced by the procedure described above. Setting  $d = 15$ , we obtained the output given in Table 5. The distribution of the cell counts and the  $z$ -score of 46.14 for the statistic  $X^2$  give very clear evidence of structure. There is nothing very special about the choice of  $p = 4$  and  $d = 15$  employed in this example; many other choices also lead to the same conclusion that structure exists. However, in this situation, we do need to take a fairly large value of  $d$  in order to detect structure. It is interesting to note that none of the three tests of multivariate normality (Mardia's skewness and kurtosis tests and Ozturk and Romeu's  $Q_n$ ) mentioned in Section 4.2 detect anything unusual in this data.

As a check, we repeated the analysis of this example replacing the flawed RANDU generator with the default uniform generator used in S-Plus. We generated four different  $50625 \times 4$  matrices  $Y$  which led to  $X^2$  statistics with the fairly modest  $z$ -scores of 2.29, 0.75, 0.51,  $-0.25$ .

Our statistic  $X^2$  is primarily intended for situations where the population mean vector  $\mu$  and covariance matrix  $\Sigma$  are unknown. There are many situations, such as the testing of random numbers in this example, which involve a null hypothesis where  $\mu$  or  $\Sigma$  (or both) are known *a priori*. In these situations, one may wish to use an initial linear transformation different from that given in (1). Park (1992) considers some alternative transformations and gives the limiting distributions of the resulting modified  $X^2$  statistics.

We have included the present example to illustrate the variety of structures which can be detected by our approach. When used as a test for random number generators, our statistic  $X^2$  closely resembles a well known test that Knuth (1981, Section 3.3) refers to as the "Serial test". There are also other procedures for testing random number generators which exploit the same underlying "cell count" idea used in the  $X^2$  statistic; see Marsaglia and Zaman (1993).

## 4.5 An Example using Geyser Data

In this example, we look at some data concerning the eruptions of the Old Faithful geyser in Yellowstone National Park, Wyoming. Two time series have been recorded: the *waiting time* between eruptions and the *duration* of the eruptions. (This data is available in S-Plus.) We shall examine the series of durations. The data were collected continuously from August 1st until August 15th, 1985. There are a total of 299 observations. The times are measured in minutes

Table 5: Application to Faulty Random Number Generator

```

*****
For d = 15,
The frequency distribution of the cell counts is:
      0      1      2      3      4      5      6
Observed 21204.0 15907.0 8359.00 3420.00 1189 378.0 112.00
Expected 18623.9 18623.9 9311.95 3103.98 776 155.2 25.87

      7      8      9     10 11 12
Observed 38.0 9.00 8.00 0.00 0 1
Expected 3.7 0.46 0.05 0.01 0 0

The moments of the distribution of cell counts are:
      mean variance skewness kurtosis
Observed 1 1.28857 1.37112 2.53929
Expected 1 1.00000 1.00000 1.00000

Observed X^2 value = 65234
Asymptotic mean and s.d. of X^2 = 50562.29, 318
z-score for X^2 = 46.14
*****

```

(see Azzalini and Bowman (1990) for further details).

We shall attempt to model the durations as an autoregressive process. S-plus includes a procedure (named “ar”) which uses the Akaike information criterion to automatically select the order of an autoregressive process. This procedure selects an AR(2) model for the durations. We shall now use our approach to examine the residuals obtained by fitting this AR(2) model.

In order to apply our method, we must choose a dimension  $p$  and convert the time series of residuals into  $p$ -variate data. We do so by dividing the residuals into disjoint subseries (or blocks) of  $p$  consecutive residuals and then taking each subseries as an observation  $y_i$ . Thus, if we start with  $m$  residuals, we create an  $(m/p) \times p$  data matrix  $Y$  as the input for our method. For the durations data, we have 297 residuals. We divided these residuals into subseries of 3 consecutive residuals leading to a  $99 \times 3$  data matrix  $Y$ .

Commonly used residual diagnostics display no obvious problems with the fit of the AR(2) model. The residual autocorrelation and partial autocorrelation functions are fairly well behaved, and the residuals are approximately normally distributed. However, our method points strongly to the existence of structure in the residuals. Our method with  $d = 3$  leads to the output in Table 6. The chi-squared statistic has a very large  $z$ -score, and there are many more cells with  $U_\pi = 0$  and  $U_\pi \geq 8$  than one would expect from the Poisson approximation.

The results of our method warn us to go back and study the residuals in greater detail. Close examination of the time series plot of the residuals suggests that the series does not have constant variance; the series has “bursts” of greater variability. This is confirmed by examining

Table 6: Output from geyser data

```

*****
For d = 3,
The frequency distribution of the cell counts is:
      0    1    2    3    4    5    6    7    8    9    10   11   12
Observed 5.00 3.00 2.00 4.00 6.0 0.00 2.00 0.00 3.00 1.00 0.00 1.00 0.00
Expected 0.69 2.53 4.64 5.67 5.2 3.81 2.33 1.22 0.56 0.23 0.08 0.03 0.01

The moments of the distribution of cell counts are:
      mean variance skewness kurtosis
Observed 3.66667  9.11111  0.67606 -0.37210
Expected 3.66667  3.66667  0.52223  0.27273

Observed X^2 value =   67.09
Asymptotic mean and s.d. of X^2 =  18.11 5.9
z-score for X^2 =   8.3
*****

```

the “lag 1” plot of the residuals (a scatter-plot of  $e_t$  versus  $e_{t-1}$ ) which shows a strong tendency for the variance of  $e_t$  to increase with the value of  $e_{t-1}$ . We think this accounts for most of the structure detected by the chi-squared test. There is another odd feature of the data which may also be producing some of the structure: At night the durations were recorded only as being short, medium or long with these possibilities represented by the values 2, 3 and 4 respectively.

In this example, our method clearly detects the existence of structure that routine diagnostic methods have missed or only hinted at. Other techniques, in this case lagged plots, are then used to investigate the nature of this structure. For another example in which our method is applied to time series residuals see the technical report by Huffer and Park (1999).

In the example above, we have used the limiting distribution given in Theorem 1 to evaluate the significance of  $X^2$ . Simulation studies indicate that this distribution is still approximately valid. For example, if we generate many series of length 299 from the fitted AR(2) model for the duration series, fit an AR(2) model to each of these generated series, and then analyze the residuals exactly as we did above, obtaining a value of  $X^2$  for each series, we find that these values of  $X^2$  follow the limiting distribution fairly well. That is, the limiting distribution is approximately the distribution of  $X^2$  when the chosen model is correct.

The technical report by Huffer and Park (1999) contains additional details and discussion on many points. It may be obtained at the web address given below:

<http://www.stat.fsu.edu/~huffer/>

## References

- Azzalini, A., and Bowman, A.W. (1990) A Look at Some data on the Old Faithful Geyser, **Applied Statistics – Journal of the Royal Statistical Society Series C**, 39, pp. 357–365.
- Chernoff, H., and Lehmann, E.L. (1954) The Use of Maximum Likelihood Estimates in  $\chi^2$  Tests for Goodness of Fit, **Annals of Mathematical Statistics**, 25, pp. 579–586.
- Farebrother, R.W. (1990) The Distribution of a Quadratic Form in Normal Variables, **Applied Statistics – Journal of the Royal Statistical Society Series C**, 39, pp. 294–309.
- Gnanadesikan, R. (1977) **Methods for Statistical Data Analysis of Multivariate Observations**, (Wiley, New York).
- Huffer, F.W., and Park, C. (1999) The Limiting Distribution of a Test for Multivariate Structure, (submitted for publication).
- Huffer, F.W., and Park, C. (1999) A Test for Multivariate Structure, Technical Report, Dept. of Statistics, Florida State University.
- Imhof, J.P. (1961) Computing the Distribution of Quadratic Forms in Normal Variables, **Biometrika**, 48, pp. 419–426.
- Knuth, D.E. (1981) **The Art of Computer Programming, Volume 2/Seminumerical Algorithms** (second edition). (Addison-Wesley, Reading, Mass)
- Koziol, J.A. (1986) Assessing Multivariate Normality: a Compendium, **Communications in Statistics – Theory and Methods**, 15, pp. 2763–2783.
- Mardia, K.V. (1970) Measures of Multivariate Skewness and Kurtosis with Applications, **Biometrika**, 57, pp. 519–530.
- Mardia, K.V. (1980) Tests of Univariate and Multivariate Normality, in: P.R. Krishnaiah (Ed.), **Handbook of Statistics**, Vol. 1 pp. 279–320 (North-Holland, Amsterdam).
- Marsaglia, G. (1968) Random Numbers Fall Mainly in the Planes, **Proceedings of the National Academy of Sciences**, 60, pp. 25–28.
- Marsaglia, G., and Zaman, A. (1993) Monkey Tests for Random Number Generators, **Computers & Mathematics with Applications**, 26(9), pp. 1–10.
- Ozturk, A., and Romeu, J.L. (1992) A New Method for Assessing Multivariate Normality with Graphical Applications, **Communications in Statistics – Simulation and Computation**, 21, pp. 15–34.
- Park, C. (1992) A Preliminary Test for Structure, (Ph.D. Dissertation, Dept. of Statistics, Florida State University, Tallahassee, FL, .
- Romeu, J.L., and Ozturk, A. (1993) A Comparative Study of Goodness-of-Fit Tests for Multivariate Normality, **Journal of Multivariate Analysis**, 46, pp. 309–334.
- Solomon, H., and Stephens, M.A. (1977) Distribution of a Sum of Weighted Chi-Square Variables,

**Journal of the American Statistical Association**, 72, pp. 881–885.

Watson, G.S. (1957) The  $\chi^2$  Goodness-of-fit Test for Normal Distributions, **Biometrika**, 44, pp. 336–348.

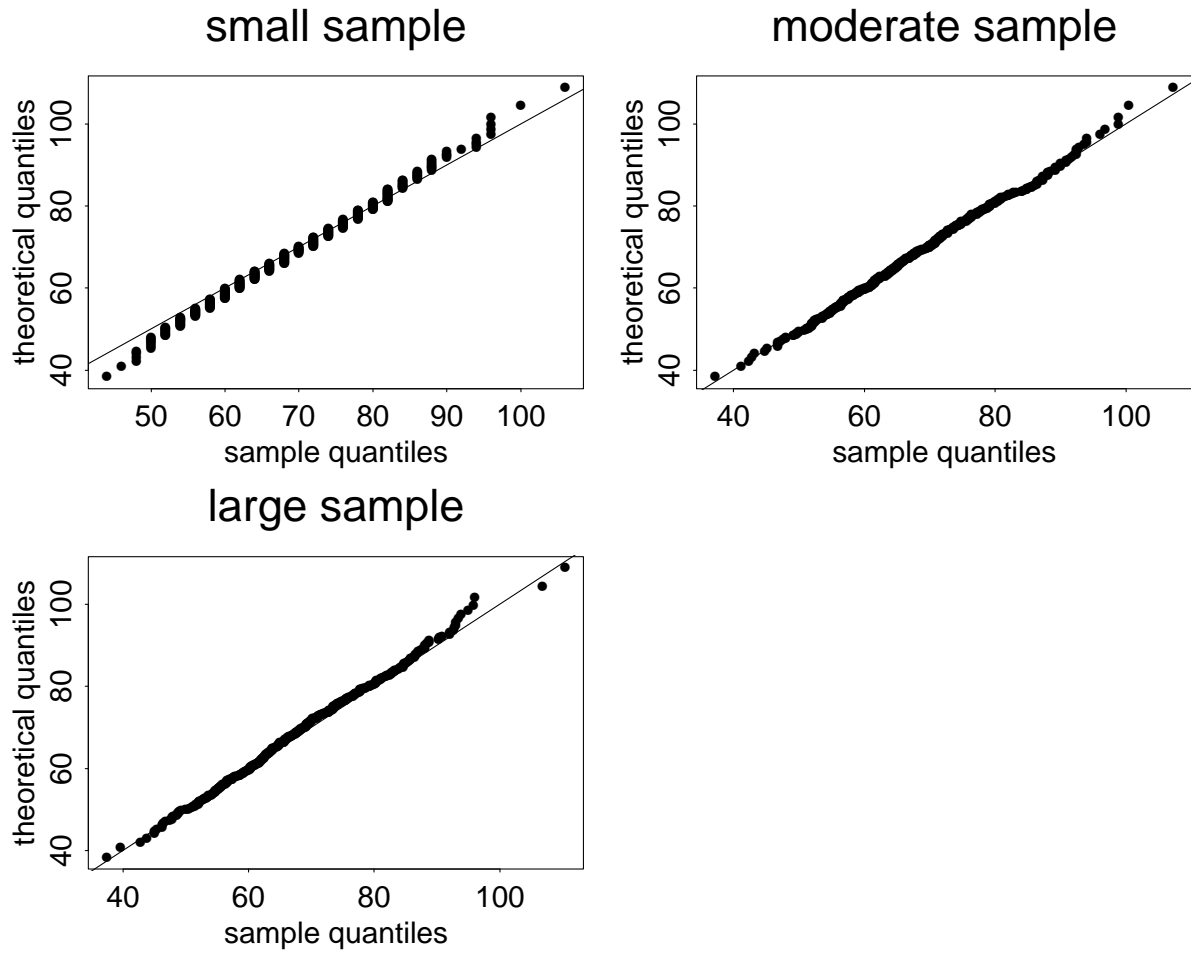


Figure 1: Quantile-quantile plots for small, moderate, and large sample cases

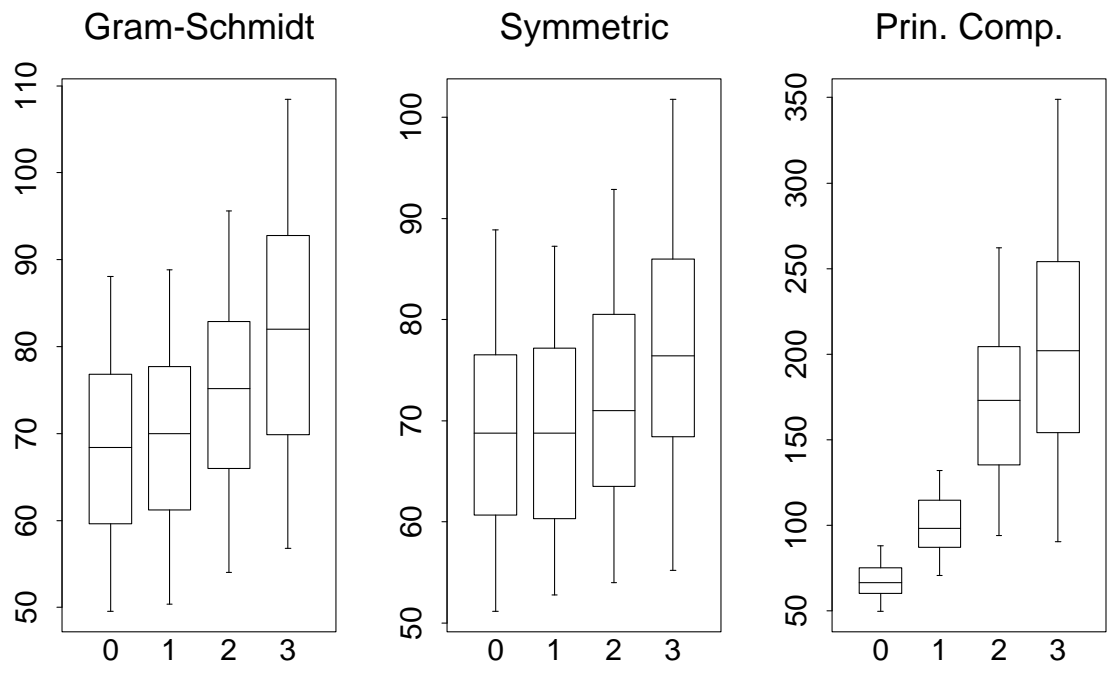


Figure 2: Boxplots depicting the effects of increasing skewness on the distribution of  $X^2$ .