

The Limiting Distribution of a Test for Multivariate Structure

Fred W. Huffer
Florida State University

Cheolyong Park
Keimyung University

Abstract

We define a chi-squared statistic for p -dimensional data as follows. First, we transform the data to remove the correlations between the p variables. Then we discretize each variable into groups of equal size and compute the cell counts in the resulting p -way contingency table. Our statistic is just the usual chi-squared statistic for testing independence in a contingency table. Because the cells have been chosen in a data-dependent manner, this statistic does not have the usual limiting distribution. We derive the limiting joint distribution of the cell counts and the limiting distribution of the chi-squared statistic when the data is sampled from a multivariate normal distribution. The chi-squared statistic is useful in detecting hidden structure in raw data or residuals. It can also be used as a test for multivariate normality.

AMS 1991 subject classifications: 62H15, 62E20, 62H20

Key words and phrases: contingency table, detecting structure, testing for independence, testing for multivariate normality.

Abbreviated Title: The Distribution of a Test for Structure

Address: (corresponding author) Fred W. Huffer, Dept. of Statistics, Florida State University, Tallahassee, Florida 32306-4330. (e-mail: 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)

1 Introduction and Summary of Results

Suppose we are given an $n \times p$ data matrix $Y = (y_{ij})$ whose rows y_1, y_2, \dots, y_n are a random sample from a p -variate distribution. The work in this paper was motivated by the search for a statistic which might help to detect the presence of hidden structure in the data Y . We wanted a statistic which is rapidly computed and is sensitive to types of structure which are difficult to detect by standard elementary techniques used during the initial examination of data (such as the sample correlation matrix, histograms, bivariate scatter-plots, etc.). If our statistic detects structure, this would then indicate to the statistician that the data warrants a closer look. In particular, the statistic might be used to predict whether it is worthwhile to carry out further analysis of the data using techniques which are more intensive in their use of human or computer time (e.g., dynamic graphics, cluster analysis, projection pursuit, etc.).

The statistic we propose to use for this general purpose is the chi-squared statistic described below. A discussion which motivates the particular form of this statistic is given in Section 1 of Huffer and Park (2000); Section 4 of the same paper presents a number of applications. The statistic may be used both as a diagnostic to test for structure and as a test for multivariate normality. In the present paper, our goal is to derive an appropriate ‘null’ (or ‘reference’) distribution for this test statistic. For this purpose we use the limiting distribution of the statistic when the data is sampled from a multivariate normal distribution. (The multivariate normal distribution serves to represent a situation which is lacking in structure.) Our method of choosing the data-dependent cells used in the construction of our chi-squared statistic leads to a particularly simple limiting distribution which we have found easy to use in applications.

In the remainder of this section we define the chi-squared statistic and state the

limiting distribution of this statistic when the data is sampled from a multivariate normal distribution. We also give the limiting joint distribution of the cell counts upon which the chi-squared statistic is based. We then briefly consider a closely related chi-squared statistic (the ‘modified’ statistic) and give the limiting distribution of this statistic. Section 2 presents an application of the chi-squared statistic. Section 3 contains the proofs of our results.

First, some brief remarks on notation. We will use I , e , and 0 to denote an identity matrix, a column vector of ones, and a column vector or matrix of zeros respectively. The dimensions will usually be clear from context, but will be specified by subscripts if necessary. Unless otherwise noted vectors will be column vectors, but for convenience they will be written in text as row vectors.

Our chi-squared statistic is computed by the following procedure. First, we apply a linear transformation to create a transformed data set Z in which the coordinates (columns) are uncorrelated and are standardized to have zero mean and unit variance. (This is often referred to as ‘sphering’ the data.) 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 $Z^t Z/n = I_p$. 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 been used before in the construction of tests for multivariate normality. See for example Moore and Stubblebine (1981) and Quiroz and Dudley (1991).

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 ΓD where Γ 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 $R = S^{-1/2}$. For our theory, the important thing is 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 Lemma 3.1.)

After obtaining the transformed data Z , we discretize (or ‘bin’) each column of Z . We choose an integer d . For each column, we use the sample quantiles to assign the values in that 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.) Now we form a contingency table from the discretized data. This contingency table contains d^p cells corresponding to the possible p -tuples of integers in $\{1, 2, \dots, d\}$. Let $\pi = (\pi_1, \pi_2, \dots, \pi_p)$ with $1 \leq \pi_i \leq d$ for all i denote a particular cell in our table, and $U_{n\pi}$ denote the number of observations in that cell (that is, the ‘cell count’). A more precise definition of the counts $U_{n\pi}$ is now given. Let $\hat{\zeta}_{i,j}$ be the (j/d) -th sample quantile of the values $z_{1i}, z_{2i}, \dots, z_{ni}$ in the i -th column of Z . We take $\hat{\zeta}_{i,0} = -\infty$ and $\hat{\zeta}_{i,d} = +\infty$. Now define

$$U_{n\pi} = \sum_{k=1}^n I(\hat{\zeta}_{i,\pi(i)-1} < z_{ki} \leq \hat{\zeta}_{i,\pi(i)} \text{ for } 1 \leq i \leq p). \quad (2)$$

Note that for vectors like π we shall use π_i and $\pi(i)$ interchangeably to avoid the use of double subscripts.

Our chi-squared statistic X^2 is defined by

$$X^2 = \sum_{\pi} \frac{(U_{n\pi} - n/d^p)^2}{n/d^p}. \quad (3)$$

This is just the usual Pearson chi-squared statistic for testing complete or total independence in a multi-way contingency table. Note that, in our situation, the ‘expected’ number of observations in each cell is taken simply to be n/d^p ; this is because of the discretization of each column into groups of *equal* size. Park (1992) studies a number of closely related chi-squared statistics which are arrived at by using different initial transformations and methods of discretization.

The choice of d in our procedure is somewhat arbitrary. We generally prefer to have as many cells as possible without allowing the average cell count n/d^p to be too small. If we wish to use the limiting distribution of the chi-squared statistic for testing purposes, our simulation work (reported in Huffer and Park (2000)) 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.

We now present the limiting distributions of the vector of cell counts $U_{n\pi}$ and of the chi-squared statistic X^2 when the data Y is sampled from a multivariate normal population.

First we introduce various matrices which are needed in the statements of our results. Let $U_n = (U_{n\pi})$ denote the $d^p \times 1$ vector of cell counts. For easy presentation of results, we assume the elements of U_n are arranged in such a way that the corresponding cell vectors $\pi = (\pi_1, \pi_2, \dots, \pi_p)$ are in a standard order; i.e. the first coordinate π_1 changes from 1 to d the fastest, the second coordinate π_2 changes the second fastest, and so on.

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

$i = 1, 2, \dots, d$, define

$$\psi_i = \phi(\zeta_{i-1}^0) - \phi(\zeta_i^0) \quad \text{and} \quad \omega_i = \zeta_{i-1}^0 \phi(\zeta_{i-1}^0) - \zeta_i^0 \phi(\zeta_i^0) \quad (4)$$

with the convention $\pm\infty \phi(\pm\infty) \equiv 0$. We define D_1 to be a $d^p \times p$ matrix the i -th column of which is d^{p-i} repetitions of the vector

$$\left(\underbrace{\psi_1, \dots, \psi_1}_{d^{i-1} \text{ times}}, \underbrace{\psi_2, \dots, \psi_2}_{d^{i-1} \text{ times}}, \dots, \underbrace{\psi_d, \dots, \psi_d}_{d^{i-1} \text{ times}} \right).$$

Also we define D_2 to be the same matrix as D_1 except that the ψ_i 's are replaced by ω_i 's. Let D_3 be the $d^p \times p(p-1)/2$ matrix obtained from D_1 in the following way: the $p(p-1)/2$ columns of D_3 are all the possible products of two distinct columns from D_1 .

Lastly, we define the $d^p \times d$ matrices E_1, E_2, \dots, E_p as follows: $E_i = (\boldsymbol{\varepsilon}_{i1}, \boldsymbol{\varepsilon}_{i2}, \dots, \boldsymbol{\varepsilon}_{id})$ with $\boldsymbol{\varepsilon}_{ij}$ a vector of length d^p formed by d^{p-i} repetitions of the vector

$$\left(\underbrace{0, 0, \dots, 0}_{(j-1) \times d^{i-1} \text{ times}}, \underbrace{1, 1, \dots, 1}_{d^{i-1} \text{ times}}, \underbrace{0, 0, \dots, 0}_{(d-j) \times d^{i-1} \text{ times}} \right).$$

It is frequently convenient to think of the elements of a $d^p \times 1$ vector as arranged in a p -way contingency table. If we do this, then E_i^t is simply the matrix which computes the marginal sums for the d categories on the i -th margin of the table.

Theorem 1.1 *If y_1, y_2, \dots, y_n are i.i.d. $N(\mu, \Sigma)$ where Σ is nonsingular, then*

$$\left(U_n - \frac{n}{d^p} e \right) / \sqrt{n/d^p} \xrightarrow{d} N(0, A) \quad \text{as } n \rightarrow \infty$$

where

$$A = I + ee^t(p-1)/d^p - \sum_{i=1}^p E_i E_i^t / d^{p-1} - D_3 D_3^t / d^{p-4}.$$

Corollary 1.1

$$X^2 \xrightarrow{d} W_1 + (1 - d^2c)W_2 \quad \text{as } n \rightarrow \infty$$

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, and $c = \left(\sum_{i=1}^d \psi_i^2\right)^2$ with ψ_i 's defined in (4).

When y_1, y_2, \dots, y_n are i.i.d. $N(\mu, \Sigma)$ and n is large, one expects the columns of the transformed data Z to have approximately a $N(0, 1)$ distribution so that the sample quantiles $\hat{\zeta}_{ij}$ appearing in (2) will be close to the population quantiles ζ_j^0 . If we replace sample quantiles by population quantiles in (2), we obtain the modified counts

$$\tilde{U}_{n\pi} = \sum_{k=1}^n I(\zeta_{\pi(i)-1}^0 < z_{ki} \leq \zeta_{\pi(i)}^0 \text{ for } 1 \leq i \leq p). \quad (5)$$

which lead to the modified chi-squared statistic

$$\tilde{X}^2 = \sum_{\pi} \frac{(\tilde{U}_{n\pi} - n/d^p)^2}{n/d^p}. \quad (6)$$

This statistic is quite similar to X^2 and can be used for some of the same purposes. Because \tilde{X}^2 does not use the sample quantiles, it can be computed much more rapidly than X^2 for very large data sets. Let $\tilde{U}_n = (\tilde{U}_{n\pi})$ denote the vector of modified cell counts. The limiting distributions of \tilde{U}_n and \tilde{X}^2 are given below.

Theorem 1.2 *If y_1, y_2, \dots, y_n are i.i.d. $N(\mu, \Sigma)$ where Σ is nonsingular, then*

$$\left(\tilde{U}_n - \frac{n}{d^p}e\right) / \sqrt{n/d^p} \xrightarrow{d} N(0, B) \quad \text{as } n \rightarrow \infty$$

where

$$B = I - ee^t/d^p - D_1D_1^t/d^{p-2} - D_2D_2^t/(2d^{p-2}) - D_3D_3^t/d^{p-4}.$$

and D_1, D_2 and D_3 are defined just prior to Theorem 1.1.

Corollary 1.2

$$\tilde{X}^2 \xrightarrow{d} W_1 + (1 - da)W_2 + (1 - db/2)W_3 + (1 - d^2c)W_4 \quad \text{as } n \rightarrow \infty$$

where W_1, W_2, W_3 and W_4 are independent chi-squared variates with degrees of freedom $\nu_1 = d^p - 1 - 2p - p(p-1)/2$, $\nu_2 = \nu_3 = p$, and $\nu_4 = p(p-1)/2$ respectively, and $a = \sum_{i=1}^d \psi_i^2$, $b = \sum_{i=1}^d \omega_i^2$, and $c = a^2$ with ψ_i and ω_i defined in (4).

In connection with the results above, we note that, when sampling from a multivariate normal population, the statistics $U_n, X^2, \tilde{U}_n, \tilde{X}^2$ are all ancillary; their distributions do not depend on the values of μ or Σ . The distributions also do not depend on the choice of $R(S)$. These facts follow from Lemma 3.1 given later.

Distributions like those in Corollaries 1.1 and 1.2 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. The statistic \tilde{X}^2 may be regarded as a multivariate generalization of the procedures in Watson (1957). The statistic X^2 does not have a univariate analog; when $p = 1$ the statistic X^2 is degenerate (constant with probability one).

2 An Example

We give one brief example to illustrate the use of X^2 . Further examples may be found in Huffer and Park (2000). We shall apply the X^2 statistic to the so-called pollen data, a well known simulated data set devised for the Data Exposition at the 1986 Joint Statistical Meetings. The data consists of $n = 3848$ observations of dimension $p = 5$. It contains a variety of artificial features, the most well known being the word EUREKA embedded within it. However, the univariate and bivariate marginal distributions of

the data betray none of these features. If one were not alerted in advance that the data contains unusual structure, one would be unlikely to invest the effort needed to discover and elucidate this structure.

The statistic X^2 is intended as a tool to help us detect hidden or non-obvious structure. It does an admirable job on this data set. Using the Gram-Schmidt transformation, we computed X^2 for $d = 3, 4,$ and 5 , obtaining the values $X^2 = 484.3, 1534.9,$ and 4380.3 , respectively. In order to judge these values, we converted them to z -scores using the mean and variance of the limiting distribution in Corollary 1.1. (See the last two columns of Table 2.) The resulting z -scores of $12.24, 11.96,$ and 16.33 , respectively, give a definite warning to the data analyst that this data contains some form of structure and should be examined more closely.

In order to verify that the limiting distribution is a good approximation to the null distribution of X^2 in this situation, we carried out a simulation study. It is easy to generate random variables from the limiting distribution in Corollary 1.1. So, for $p = 5$ and each of the values $d = 3, 4,$ and 5 , we approximated the quantiles of the limiting distribution (corresponding to upper tail probabilities of $.75, .50, .25, .10, .05, .025, .01, .005,$ and $.001$) by the sample quantiles of a sample of size $1,000,000$. Then, for each of $d = 3, 4,$ and 5 , we generated $50,000$ matrices of dimension 3848×5 with i.i.d. $N(0, 1)$ entries, and computed the value of X^2 for each of these. This gave us samples of size $50,000$ from the true null distribution of X^2 . Table 1 records the proportion of these values which lie above each of the estimated quantiles of the limiting distribution. In Table 2, we compare the mean and variance of these samples of $50,000$ with the exact mean and variance of the limiting distribution in Corollary 1.1. These tables demonstrate that the limiting distribution is very close to the null distribution of X^2 in this example.

	.75	.50	.25	.10	.05	.025	.01	.005	.001
$d = 3$.74754	.49704	.24766	.10056	.05004	.02542	.01026	.00538	.00114
$d = 4$.75066	.49950	.25094	.09872	.04968	.02494	.01046	.00532	.00088
$d = 5$.74998	.50402	.25048	.10042	.05036	.02466	.01058	.00518	.00090

Table 1: Upper tail probabilities for the actual null distribution of X^2 (when $p = 5$ and $n = 3848$) evaluated at quantiles of the limiting distribution.

	actual		limiting	
	mean	variance	mean	variance
$d = 3$	225.69	446.50	225.71	446.75
$d = 4$	1000.49	1992.86	1000.59	1997.35
$d = 5$	3096.17	6158.11	3095.95	6188.76

Table 2: Comparison of the mean and variance of the actual null distribution of X^2 (when $p = 5$ and $n = 3848$) with those of the limiting distribution.

3 Proofs

We first give an overview of the proof of Theorem 1.1 and Corollary 1.1, and comment on various aspects of the proof. The proof consists of the following parts. We begin with two lemmas: an “invariance” lemma (Lemma 3.1) using an argument from Eaton (1983), and an “approximation” lemma (Lemma 3.2) using ideas and methods from the large sample theory of general chi-squared statistics in Moore and Spruill (1975) and Pollard (1979). A projection argument then completes the proof of Theorem 1.1. Finally, obtaining the eigenvalues of the covariance matrix A in Theorem 1.1 leads us to our Corollary 1.1.

The invariance lemma states that the distribution of the spherized (standardized) data Z does not depend on μ or Σ or on the choice of the transformation $R(S)$. Since the cell counts U_n and the chi-squared statistic X^2 are functions of Z , the distributions

of these quantities are invariant in the same sense. Similarly, the limiting distributions of U_n and X^2 do not depend on μ or Σ or $R(S)$. This greatly simplifies our proof by allowing us to assume without loss of generality that $\mu = 0$ and $\Sigma = I$. Also, we can assume that $R(S)$ is upper triangular with positive diagonal elements, that is, we can restrict ourselves to using the Gram-Schmidt transformation. Setting $\mu = 0$ and $\Sigma = I$ simplifies the notation and calculations considerably, but is perhaps not absolutely necessary. However, the restriction to the Gram-Schmidt transformation (or some other well-behaved transformation) is essential. The Gram-Schmidt transformation $R(S)$ is a continuous and differentiable transformation satisfying $R(I) = I$. (In contrast, the principal components transformation $R(S)$ is not even continuous at $S = I$.) This guarantees that $R(S) = I + O_p(n^{-1/2})$. We need this for the proof of Lemma 3.2 and also in the projection argument which completes the proof of Theorem 1.1. The weaker statement $R(S) = I + o_p(1)$ suffices for the proof of Lemma 3.2, but the stronger rate of convergence is definitely required for the projection argument.

The invariance lemma is a new result. One special case, the Gram-Schmidt transformation, was proved by Eaton (1983), and another special case (involving $R(S) = S^{-1/2}$) was conjectured, but not proved, by Quiroz and Dudley (1991, p. 544). Since the “spherizing” of data is a commonly used procedure, the invariance lemma is likely to have other applications in statistics.

The approximation lemma (Lemma 3.2) uses ideas from the theory of empirical processes (surveyed in Chapter 2.1 of van der Vaart and Wellner (1996)) to approximate the cell counts U_n , for the random cells defined in terms of the spherized data Z , by the cell counts for the limiting cells to which these random cells converge. (Precise definitions of the random cells and the limiting cells are given later.) Lemma 3.2 is a specialized variant of results well known in the literature. We included it only because

we could not find a result in the literature that exactly suited our purposes, and because the available results were usually stated in much greater generality than we needed.

We now indicate why we need the projection argument which concludes the proof of Theorem 1.1. Let $P_n = (P_{n\pi})$ denote the vector of estimated cell probabilities for the random cells defined in terms of the spherized data Z . (The cell probabilities $P_{n\pi}$ are computed under the $N(\mu, \Sigma)$ distribution with μ and Σ estimated by their maximum likelihood estimates \bar{y} and S . The quantities $P_{n\pi}$ are defined more precisely later.) Applying Lemma 3.2 and doing some calculation leads fairly directly to results for the limiting distribution of $n^{-1/2}(U_n - nP_n)$. From these results we could then obtain the limiting distribution of the chi-squared statistic defined by $\sum_{\pi}(U_{n\pi} - nP_{n\pi})^2/nP_{n\pi}$. (The theory in Pollard (1979) applies to test statistics of this form.) But our statistic X^2 is not exactly in this form; it has $1/d^p$ in place of $P_{n\pi}$. It is clear that we can replace $P_{n\pi}$ by any quantity which differs from it by an amount which is $o_p(n^{-1/2})$, since this changes the chi-squared statistic by an amount which is $o_p(1)$ and thus does not affect the asymptotics. Unfortunately, in our situation $P_{n\pi} - 1/d^p = O_p(n^{-1/2})$ so that we cannot employ this simple fix. The projection argument gets us around this difficulty; it gives us a correct way to replace $P_{n\pi}$ by $1/d^p$ and obtain the limiting distribution of $(U_n - (n/d^p)e)$. Something like this projection argument seems to be absolutely needed because the theorems available in the literature lead to results for $(U_n - nP_n)$ instead of $(U_n - (n/d^p)e)$ as we require.

The proof of Corollary 1.1 is essentially an eigenvalue decomposition of the covariance matrix A . The details of this calculation and the various calculations which go into the proof of Theorem 1.1 are mostly omitted, but these calculations actually form the backbone of the paper. For very general classes of cells, the theory of empirical processes allows us to show that the vector of (appropriately standardized) cell counts

converges to a limiting normal distribution, and then standard statistical theory tells us that the limiting distribution of a quadratic form involving these cell counts will be that of a linear combination of chi-squared variates with the coefficients obtained as the eigenvalues of a certain matrix. However, none of this theory guarantees that the answers we get will be tractable and usable. It is interesting and surprising that all the calculations work out so that in the end we get a very simple answer, the limiting distribution in our Corollary 1.1. The fact that this limiting distribution is simple and usable is one important contribution of our paper.

To emphasize this point, imagine a situation (involving, say, a different class of cells, or a somewhat different test statistic, or sampling the data from a different family of distributions) in which we can prove an asymptotic normality result like that in Theorem 1.1 and obtain an explicit formula for the limiting covariance matrix A . Now suppose the formula for A is not as simple as that in our Theorem 1.1, so that we are unable to obtain closed form expressions for the eigenvalues of A . In this situation we can still obtain a version of Corollary 1.1, but the coefficients of the chi-squared variates would now have to be obtained by numerically computing the eigenvalues of A . This would render the results much less useful because the dimension of the matrix A is $d^p \times d^p$, and, in many examples in which we have used our procedure, the number of cells d^p is quite large. In Huffer and Park (2000) we give a number of examples in which we apply our statistic X^2 . In Example 4.3 (involving speech data) the number of cells was $2^9 = 512$, and in Example 4.4 (which examined a faulty random number generator) the number of cells was $15^4 = 50,625$. For very high-dimensional matrices, the numerical computation of eigenvalues is difficult and time-consuming. If the formula for A were awkward, even merely computing the mean and variance of the limiting distribution would be cumbersome when d^p is sufficiently large.

We now begin the proofs.

It is notationally convenient to prove our “invariance” lemma in the setting of a multivariate regression model. Let

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim N(0, I \otimes \Sigma)$, that is the rows of $\boldsymbol{\epsilon}$ are i.i.d. $N(0, \Sigma)$. The matrices in (1) have the following dimensions: Y is $n \times p$, X is $n \times q$, $\boldsymbol{\beta}$ is $q \times p$, and $\boldsymbol{\epsilon}$ is $n \times p$. We assume that X has full rank. We define the projection matrix $Q = I - X(X^tX)^{-1}X^t$. Then the “standardized” data $Z = Z(Y)$ is the $n \times p$ matrix given by

$$Z(Y) = QYR(S), \tag{2}$$

where $S = S(Y) = n^{-1}Y^tQY$ and $R = R(S)$ satisfies $R^tSR = I$ so that $n^{-1}Z^tZ = I$. (Note: In Theorem 1.1 we are interested in the special case where $q = 1$ and $X = e$.)

Lemma 3.1 *Under the multivariate regression model described above, the standardized data $Z(Y)$ is ancillary and the distribution of $Z(Y)$ is the same for all possible choices of $R(S)$.*

Proof: We shall use results and arguments from Example 7.19 (page 292) of Eaton (1983). In this example Eaton shows that $Z(Y)$ is ancillary in the special case where $R(S)$ is an upper triangular matrix with positive diagonal elements. We shall extend his argument to handle the more general matrices $R(S)$. The matrix $Z(Y)$ is an element of the space

$$\mathcal{F} = \{z : z^tz = nI_p \text{ and } Qz = z\}.$$

Let \mathcal{O}_n be the set of $n \times n$ orthogonal matrices. The group

$$H = \{\Gamma : \Gamma \in \mathcal{O}_n, \Gamma Q = Q\Gamma\}$$

is compact and acts transitively on \mathcal{F} under the group action

$$z \rightarrow \Gamma z, \quad z \in \mathcal{F}, \quad \Gamma \in H.$$

Thus, there is a unique H -invariant probability measure ν on \mathcal{F} .

To prove our lemma it suffices to show that $\mathcal{L}(Z(Y)) = \nu$, or equivalently, $\mathcal{L}(\Gamma Z(Y)) = \mathcal{L}(Z(Y))$ for all $\Gamma \in H$. First, we note that Z can be expressed as a function of QY alone. Since $QY \sim N(0, Q \otimes \Sigma)$, this implies that the distribution of $Z(Y)$ is free of the regression parameters β . This allows us to assume without loss of generality that $Y \sim N(0, I \otimes \Sigma)$. In this case $\mathcal{L}(\Gamma Y) = \mathcal{L}(Y)$ for all $\Gamma \in \mathcal{O}_n$. Also, we know that $S(\Gamma Y) = S(Y)$ for $\Gamma \in H$. Therefore, for $\Gamma \in H$ we have

$$\begin{aligned} \mathcal{L}(\Gamma Z(Y)) &= \mathcal{L}(\Gamma QY R(S(Y))) = \mathcal{L}(Q\Gamma Y R(S(\Gamma Y))) \\ &= \mathcal{L}(QY R(S(Y))) = \mathcal{L}(Z(Y)), \end{aligned}$$

which completes the proof. □

Now we return to the case $X = e$ of interest in Theorem 1.1. Lemma 3.1 allows us to simplify our problem in two ways. First, without loss of generality, we may restrict ourselves to using Gram-Schmidt transformations; we take the matrix R to be upper triangular with positive diagonal elements. This restriction ensures that $R(S)$ is a continuous and differentiable function of the positive definite matrix S satisfying $R(I) = I$. Let $\theta = (\mu, \Sigma)$ denote the parameters of a multivariate normal population. Because Z is ancillary, we may assume without loss of generality that Y is sampled from a population with parameters $\theta_0 = (0, I)$. These two restrictions will remain in force throughout the remainder of this section.

The cell count $U_{n\pi}$ is the number of observations y_i lying in a certain region $\Lambda_{n\pi}$ of R^p . We now introduce some notation to describe this region. Let ζ be a $p \times (d - 1)$

matrix with entries ζ_{ij} satisfying $\zeta_{i1} \leq \zeta_{i2} \leq \dots \leq \zeta_{i(d-1)}$ for all i . For a given ζ and vector of integers $\tau = (\tau_1, \tau_2, \dots, \tau_p)$ satisfying $0 \leq \tau_i \leq d$ for all i , we define the $p \times 1$ vector ζ_τ by $\zeta_\tau = (\zeta_{1\tau(1)}, \zeta_{2\tau(2)}, \dots, \zeta_{p\tau(p)})$ where we take $\zeta_{i0} = -\infty$ and $\zeta_{id} = \infty$. For vectors $a = (a_i)$ and $b = (b_i)$, we say that $a \prec b$ if $a_i < b_i$ for all i , and $a \preceq b$ if $a_i \leq b_i$ for all i . Now for given $\theta = (\mu, \Sigma)$ and ζ , for each cell π we define the region

$$\Lambda_\pi(\theta, \zeta) = \left\{ y \in R^p : \zeta_{\pi-e} \prec R^t(\Sigma)(y - \mu) \preceq \zeta_\pi \right\}. \quad (3)$$

Let $\hat{\zeta}_n = (\hat{\zeta}_{ij})$ be a matrix whose entries are sample quantiles of the transformed data Z . More precisely, $\hat{\zeta}_{ij}$ is the (j/d) -th sample quantile of the i -th coordinate $z_{1i}, z_{2i}, \dots, z_{ni}$. It is easy to show (see (4) below) that $\hat{\zeta}_n \rightarrow \zeta_0$ in probability where $\zeta_0 = (\zeta_{ij}^0)$ is the matrix of population quantiles with entries $\zeta_{ij}^0 = \zeta_j^0 = \Phi^{-1}(j/d)$. Let θ_n be the maximum likelihood estimate of $\theta = (\mu, \Sigma)$ based on y_1, \dots, y_n , that is, $\theta_n = (\bar{y}, S)$. We know that $\theta_n \rightarrow \theta_0$ in probability. We may now define the vector of regions $\Lambda_n = (\Lambda_{n\pi})$ by $\Lambda_{n\pi} = \Lambda_\pi(\theta_n, \hat{\zeta}_n)$ and the limiting regions $\Lambda_0 = (\Lambda_{0\pi})$ by $\Lambda_{0\pi} = \Lambda_\pi(\theta_0, \zeta_0)$. The regions in Λ_n form a partition of R^p .

For an arbitrary region $\Gamma \subset R^p$, we define $U_n(\Gamma) = \sum_{i=1}^n I\{y_i \in \Gamma\}$ and define $P(\Gamma, \theta)$ to be the probability assigned to Γ by the normal distribution with parameter θ . We also let $D(\Gamma)$ be $\partial P(\Gamma, \theta)/\partial \theta$ evaluated at $\theta = \theta_0$. For a vector of regions $\Gamma = (\Gamma_i)$ we use the obvious vector analogs of the above definitions so that $U_n(\Gamma) = (U_n(\Gamma_i))$, $P(\Gamma, \theta)$ is a vector of probabilities, and $D(\Gamma)$ is a matrix of partial derivatives. Our vector of cell counts U_n may now be written as $U_n(\Lambda_n)$. Finally, we define the process $V_n(\Gamma) = n^{-1/2}(U_n(\Gamma) - nP(\Gamma, \theta_0))$. Note that the vector of estimated cell probabilities $P(\Lambda_n, \theta_n)$ was referred to as P_n in the discussion at the beginning of this section.

Using this notation, we can now state the following approximation lemma.

Lemma 3.2

$$n^{-1/2} (U_n(\Lambda_n) - nP(\Lambda_n, \theta_n)) = V_n(\Lambda_0) - D(\Lambda_0)\sqrt{n}(\theta_n - \theta_0) + o_p(1).$$

This lemma is similar to the approximation given by Moore and Spruill (1975) in their Theorem 4.1. Moore and Spruill were limited to using rectangular cells with sides parallel to the coordinate axes. In our setting this would correspond to requiring $R(S)$ to be a diagonal matrix. This limitation was removed by the discussion in Sections 3 and 5 of Pollard (1979) which used the results of Dudley (1978). In our very specialized situation it is easier to give a self-contained argument than to use the general theorems and conditions given by Moore and Spruill (1975) and Pollard (1979). This argument is now given.

Proof: We shall first list four facts we need in our proof. Our first two facts are consequences of Dudley's Central Limit Theorem for Empirical Measures; see Dudley (1978) or Giné and Zinn (1984). We rely, in particular, on the discussion just before Theorem 2 in Pollard (1979). The class of half-spaces \mathcal{C} is a Donsker class for $P(\cdot, \theta_0)$. This implies that $\sup\{|V_n(C)| : C \in \mathcal{C}\} = O_p(1)$ which gives a uniform $O_p(n^{-1/2})$ rate of convergence for the empirical cdf's of all linear combinations $\{a^t y_1, a^t y_2, \dots, a^t y_n\}$. We also know that $\theta_n = \theta_0 + O_p(n^{-1/2})$ which implies $R(S) = I + O_p(n^{-1/2})$. Combining these two facts we obtain

$$\Phi(\hat{\zeta}_{ij}) = \frac{j}{d} + O_p(n^{-1/2}) \quad \text{and} \quad \hat{\zeta}_{ij} = \zeta_{ij}^0 + O_p(n^{-1/2}) \quad \text{for all } i, j. \quad (4)$$

We do not immediately need the full strength of (4), but only the fact that $\hat{\zeta}_n \xrightarrow{p} \zeta_0$. Since we also have $\theta_n \xrightarrow{p} \theta_0$, we conclude easily that Λ_n converges to Λ_0 in the sense that $P(\Lambda_\pi(\theta_n, \hat{\zeta}_n) \triangle \Lambda_\pi(\theta_0, \zeta_0), \theta_0) \xrightarrow{p} 0$ for all cells π . Now the class of all sets of the form $\Lambda_\pi(\theta, \zeta)$ is a Donsker class since $\Lambda_\pi(\theta, \zeta)$ can be expressed as an intersection

of $2p$ open or closed half-spaces. This implies that

$$V_n(\Lambda_n) = V_n(\Lambda_0) + o_p(1). \quad (5)$$

Our final two facts concern the multivariate normal distribution. The proofs are straightforward and we shall omit them. First, there exists $c_1 > 0$ such that for all measurable $A \subset R^p$,

$$|\mathbb{P}(A, \theta) - \mathbb{P}(A, \theta_0) - D(A)(\theta - \theta_0)| \leq c_1 \|\theta - \theta_0\|^2. \quad (6)$$

Secondly, there exists $c_2 > 0$ such that for all measurable $A, B \subset R^p$,

$$\|D(A) - D(B)\|^2 \leq c_2 \mathbb{P}(A \Delta B, \theta_0). \quad (7)$$

Here, $\|\cdot\|$ is the usual Euclidean norm.

We now prove our lemma by applying facts (5), (6), (7) in sequence to obtain

$$\begin{aligned} n^{-1/2} (U_n(\Lambda_n) - n\mathbb{P}(\Lambda_n, \theta_n)) &= V_n(\Lambda_n) + \sqrt{n}(\mathbb{P}(\Lambda_n, \theta_0) - \mathbb{P}(\Lambda_n, \theta_n)) \\ &= V_n(\Lambda_0) + \sqrt{n}(\mathbb{P}(\Lambda_n, \theta_0) - \mathbb{P}(\Lambda_n, \theta_n)) + o_p(1) \\ &= V_n(\Lambda_0) - D(\Lambda_n)\sqrt{n}(\theta_n - \theta_0) + o_p(1) \\ &= V_n(\Lambda_0) - D(\Lambda_0)\sqrt{n}(\theta_n - \theta_0) + o_p(1). \end{aligned}$$

□

Proof of Theorem 1.1

Suppose the coordinates of $\theta = (\mu, \Sigma)$ are arranged so that $\theta = (\mu, \boldsymbol{\sigma}, \boldsymbol{\rho})$, where $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \dots, \sigma_{pp})$ and $\boldsymbol{\rho} = (\sigma_{12}, \sigma_{13}, \dots, \sigma_{(p-1)p})$ are the diagonal and off-diagonal elements of Σ respectively. Then it is easy to show that $D(\Lambda_0) = (D_1/d^{p-1}, D_2/(2d^{p-1}), D_3/d^{p-2})$, where D_1, D_2, D_3 are defined just prior to Theorem 1.1. For any vector $x = (x_1, x_2, \dots, x_p)$,

we define the column vectors $s(x) = (x_1^2, x_2^2, \dots, x_p^2)$, and $r(x) = (x_1x_2, x_1x_3, \dots, x_{p-1}x_p)$ with dimensions p and $p(p-1)/2$ respectively. Then since $S = Y^tY/n + o_p(n^{-1/2})$, Lemma 3.2 leads to

$$n^{-1/2} (U_n(\Lambda_n) - nP(\Lambda_n, \theta_n)) = \tag{8}$$

$$V_n(\Lambda_0) - \sqrt{n} \left(D_1 \bar{y}/d^{p-1} + D_2 \sum_{i=1}^n (s(y_i) - e)/(2nd^{p-1}) + D_3 \sum_{i=1}^n r(y_i)/(nd^{p-2}) \right) + o_p(1).$$

Using the notation E_i from Theorem 1.1, we define the $d^p \times d^p$ matrix Π by

$$\Pi = I + ee^t(p-1)/d^p - \sum_{i=1}^p E_i E_i^t / d^{p-1}.$$

It is easy to verify that Π is a projection matrix with rank $d^p - 1 - p(d-1)$. If we view the d^p coordinates as being arranged in a p -way table, then Π is simply the operator which removes the marginal means. We shall now apply Π to both sides of (8). It is easy to show that $\Pi U_n(\Lambda_n) = U_n(\Lambda_n) - (n/d^p)e + O(1)$, $\Pi D_1 = 0$, $\Pi D_2 = 0$, and $\Pi D_3 = D_3$. (Note that the equality $\Pi U_n(\Lambda_n) = U_n(\Lambda_n) - (n/d^p)e$ is exact when n is divisible by d .) Since $P(\Lambda_\pi(\theta, \zeta), \theta)$ does not depend on θ , that is, $P(\Lambda_\pi(\theta, \zeta), \theta) = P(\Lambda_\pi(\theta_0, \zeta), \theta_0)$, we have

$$P(\Lambda_{n\pi}, \theta_n) = \prod_{i=1}^p [\Phi(\hat{\zeta}_{i,\pi(i)}) - \Phi(\hat{\zeta}_{i,\pi(i)-1})] = \prod_{i=1}^p (d^{-1} + \xi_{i\pi}) \tag{9}$$

where we have defined $\xi_{i\pi} = \Phi(\hat{\zeta}_{i,\pi(i)}) - \Phi(\hat{\zeta}_{i,\pi(i)-1}) - d^{-1}$. Equation (4) implies $\xi_{i\pi} = O_p(n^{-1/2})$ so that

$$P(\Lambda_{n\pi}, \theta_n) = d^{-p} + d^{-(p-1)} \sum_{i=1}^p \xi_{i\pi} + O_p(n^{-1}).$$

Let $\xi_i = (\xi_{i\pi})$. Since $\Pi \xi_i = 0$ for all i , we see that $\Pi P(\Lambda_n, \theta_n) = o_p(n^{-1/2})$. Putting this all together we have

$$n^{-1/2} (U_n(\Lambda_n) - (n/d^p)e) = \Pi V_n(\Lambda_0) - n^{-1/2} D_3 \sum_i r(y_i)/d^{p-2} + o_p(1). \tag{10}$$

The finite dimensional central limit theorem tells us that $(V_n(\Lambda_0), n^{-1/2} \sum_i r(y_i))$ converges in distribution to a normal distribution with mean vector 0 and covariance matrix (obtained after a short calculation) given by

$$\begin{pmatrix} d^{-p}(I - ee^t/d^p) & d^{-(p-2)}D_3 \\ d^{-(p-2)}D_3^t & I \end{pmatrix}.$$

Combining this with (10) and doing another short calculation (using the fact $\Pi D_3 = D_3$), we see that $n^{-1/2}(U_n(\Lambda_n) - (n/d^p)e)$ converges in distribution to a normal distribution with mean vector 0 and covariance matrix $d^{-p}(\Pi - d^{-(p-4)}D_3D_3^t)$. This implies the desired result. \square

Proof of Corollary 1.1

According to standard theory, if $W \sim N(0, A)$ with A as given in Theorem 1.1, then $W^tW \stackrel{d}{=} \sum_i \lambda_i Z_i^2$ where Z_i are i.i.d. $N(0, 1)$ and λ_i are the eigenvalues of A . To determine the eigenvalues of A we rewrite it in the form

$$A = I - (I - \Pi) - cd^2\Omega$$

where $\Omega = D_3D_3^t/(cd^{p-2})$ and c is as in Corollary 1.1. Using the facts $\Pi D_3 = D_3$ and $D_3^tD_3 = cd^{p-2}I$, we can verify that $(I - \Pi)\Omega = 0$, and that $I - \Pi$ and Ω are symmetric and idempotent with ranks $1 + p(d - 1)$ and $p(p - 1)/2$ respectively. This implies the nonzero eigenvalues of A are: 1 with multiplicity $d^p - 1 - p(d - 1) - p(p - 1)/2$, and $1 - cd^2$ with multiplicity $p(p - 1)/2$. The remaining $1 + p(d - 1)$ eigenvalues are zero. This completes the proof. \square

Proof of Theorem 1.2 and Corollary 1.2

The proof of Theorem 1.2 is essentially the same as that of Theorem 1.1 except that we do not need the final projection argument. We use the notation given prior to Lemma

3.2 with one small change; we define $\Lambda_{n\pi} = \Lambda_\pi(\theta_n, \zeta_0)$ so that now $\tilde{U}_n = U_n(\Lambda_n)$. With this change, Lemma 3.2 remains valid with the same proof; simply substitute $\hat{\zeta}_n$ by ζ_0 throughout. Similarly, equation (8) still holds. Theorem 1.2 follows directly from equation (8) if we make the following three observations. First, with the new definition of Λ_n , we have $P(\Lambda_n, \theta_n) = e/d^p$. Secondly, the central limit theorem implies that

$$\mathbf{W}_n = \left(V_n(\Lambda_0), \quad n^{1/2}\bar{y}, \quad n^{-1/2}\sum_i(s(y_i) - e), \quad n^{-1/2}\sum_i r(y_i) \right) \xrightarrow{d} N(0, \mathbf{C}) \quad (11)$$

$$\text{where } \mathbf{C} = \begin{pmatrix} d^{-p}(I - ee^t/d^p) & d^{-(p-1)}D_1 & d^{-(p-1)}D_2 & d^{-(p-2)}D_3 \\ d^{-(p-1)}D_1^t & I_p & 0 & 0 \\ d^{-(p-1)}D_2^t & 0 & 2I_p & 0 \\ d^{-(p-2)}D_3^t & 0 & 0 & I_{p(p-1)/2} \end{pmatrix}.$$

Finally, the covariance matrix of the linear transformation of \mathbf{W}_n indicated in (8) may be computed using (11) and the fact that $D_i^t D_j = 0$ for $i \neq j$.

The proof of Corollary 1.2 is similar to that of Corollary 1.1. The result follows immediately from the fact that ee^t/d^p , $D_1 D_1^t/(ad^{p-1})$, $D_2 D_2^t/(bd^{p-1})$, $D_3 D_3^t/(cd^{p-2})$ are mutually orthogonal idempotent matrices of ranks 1, p , p , $p(p-1)/2$ respectively, where a, b, c are as defined in Corollary 1.2. (See Park (1992) for a detailed proof of this.)

There is an alternative method of proving Theorem 1.2 which is in some ways simpler and more intuitive. In particular, this method does not require any use of the general theory of empirical processes. We shall briefly sketch this approach. We continue using the modified definition $\Lambda_{n\pi} = \Lambda_\pi(\theta_n, \zeta_0)$ given above. The vector $\tilde{U}_n = U_n(\Lambda_n)$ is independent of $\theta_n = (\bar{y}, S)$ since \tilde{U}_n is ancillary (by Lemma 3.1) and θ_n is a complete sufficient statistic for $\theta = (\mu, \Sigma)$. This implies that $\mathcal{L}(\tilde{U}_n) = \mathcal{L}(\tilde{U}_n | \theta_n = \theta_0)$. But $\theta_n = \theta_0$ implies $\Lambda_n = \Lambda_0$. Thus, noting that $P(\Lambda_0, \theta_0) = e/d^p$, we can write

$$\mathcal{L}[n^{-1/2}(\tilde{U}_n - (n/d^p)e)] = \mathcal{L}[V_n(\Lambda_0) | \theta_n = \theta_0]. \quad (12)$$

We shall restate this fact using the notation in (11). Let us partition \mathbf{W}_n as $\mathbf{W}_n = (W_{n1}, W_{n2})$ where $W_{n1} = V_n(\Lambda_0)$. Then (12) can be written

$$\mathcal{L}[n^{-1/2}(\tilde{U}_n - (n/d^p)e)] = \mathcal{L}(W_{n1} | W_{n2} = 0). \quad (13)$$

Introduce $\mathbf{W} = (W_1, W_2)$ which is partitioned the same as (W_{n1}, W_{n2}) and satisfies $\mathbf{W} \sim N(0, \mathbf{C})$. Similarly partition \mathbf{C} into four blocks denoted C_{ij} . Since (11) says that $(W_{n1}, W_{n2}) \xrightarrow{d} (W_1, W_2)$, it is intuitively plausible that

$$\mathcal{L}(W_{n1} | W_{n2} = 0) \rightarrow \mathcal{L}(W_1 | W_2 = 0) = N(0, C_{11} - C_{12}C_{22}^{-1}C_{21}). \quad (14)$$

A rigorous proof of this statement can be given using the conditional limit theorem for exponential families in Holst (1981). Combining (13) and (14) (and doing a little calculation) leads immediately to the result in Theorem 1.2.

A detailed proof of Theorem 1.2 along the lines sketched above is given in Park (1992). We were unable to obtain a proof of Theorem 1.1 using the same approach, but this approach (used merely as a heuristic) did suggest the correct limiting distribution.

References

- Chernoff, H., and Lehmann, E.L. (1954). The Use of Maximum Likelihood Estimates in χ^2 Tests for Goodness of Fit. *Ann. Math. Statist.* **25**, 579-586.
- Dudley, R.M. (1978). Central Limit Theorems for Empirical Measures. *Ann. Probab.* **6**, pp. 899-929.
- Eaton, M.L. (1983). *Multivariate Statistics: a Vector Space Approach*. Wiley, New York.
- Giné, E., and Zinn, J. (1984). Some Limit Theorems for Empirical Processes. *Ann. Probab.* **12**, 929-989.

- Holst, L. (1981). Some Conditional Limit Theorems in Exponential Families. *Ann. Prob.* **9** 818-830.
- Huffer, F.W., and Park, C. (2000). A Test for Multivariate Structure, *Journal of Applied Statistics* **27**, 633-650.
- Moore, D.S., and Spruill, M.C. (1975). Unified Large-sample Theory of General Chi-squared Statistics for Tests of Fit. *Ann. Statist.* **3**, 599-616.
- Moore, D.S., and Stubblebine, J.B. (1981). Chi-squared Tests for Multivariate Normality with Application to Common Stock Prices. *Commun. Statist.; Theor. Meth.* **10**, 713-738.
- Park, C. (1992). A Preliminary Test for Structure. Ph.D. Dissertation, Dept. of Statistics, Florida State University.
- Pollard, D. (1979). General Chi-square Goodness-of-fit Tests with Data-dependent Cells. *Z. Wahrsch. verw. Gebiete* **50**, 317-332.
- Quiroz, A.J., and Dudley, R.M. (1991). Some New Tests for Multivariate Normality. *Probab. Th. Rel. Fields* **87**, pp. 521-546.
- Van der Vaart, A.W., and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer-Verlag, New York.
- Watson, G.S. (1957). The χ^2 Goodness-of-fit Test for Normal Distributions. *Biometrika* **44**, 336-348.