A COMPARATIVE VISUALIZATION STUDY OF A GRAPHICAL MULTIVARIATE NORMALITY GOF TEST.

Jorge Luis Romeu
Department of Mathematics
SUNY Cortland, NY 13045
jromeu@suvm.bitnet

Abstract

An ordered sequence of plots of a new graphical multivariate normality test, performed on samples of selected distributions, is presented. The multivariate sample is transformed into a set of linked vectors in a bivariate space. According to where the vector endpoints fall, in relation to the confidence ellipse, multivariate normality is accepted or rejected. If normality is rejected, we visually analyze where the vector's endpoints fall, outside the confidence ellipse. We also compare the linked vector pattern, in relation to the null (solid line) pattern. This visual analysis provides an indication of how does the alternative non normal distribution looks like.

1.0 The Graphical Test.

The Multivariate Qn (Ozturk and Romeu, 1992) graphical procedure can be divided into three parts: (i) the confidence ellipse, (ii) the lower half of the pattern and (iii) the upper half (Figure 1). When a sample comes from a multivariate normal distribution, the corresponding linked vector closely follows both halves of the pattern and ends within the confidence ellipse. When the sample comes from another distribution, its vector endpoint falls outside the confidence ellipse. But the area, outside of the ellipse it falls, and the pattern it follows, is closely associated to the distribution it comes from. In this paper we present a study of test patterns from non normal alternative distributions.

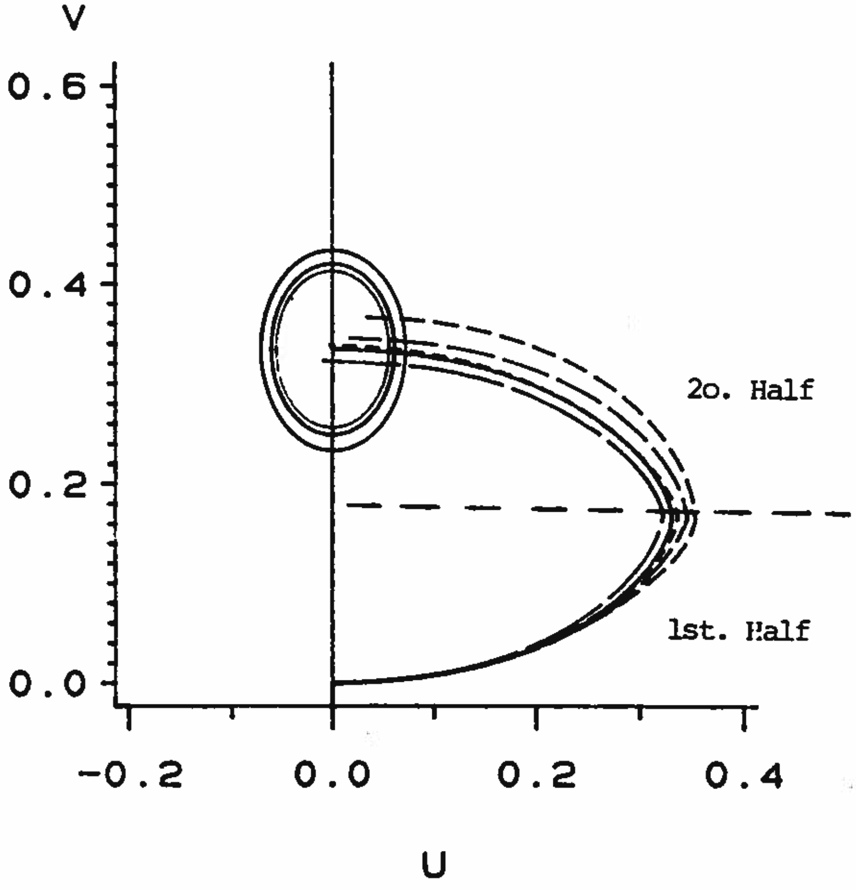


Figure 1.

2.0 Non Normal Alternatives.

The bivariate distributions in this study (Figure 2) were chosen as to be (i) more skewed, (ii) more kurtic than a bivariate normal and (iii) a combination of these. In this paper we only present results for three special cases: skewed, kurtic and combination. Samples of size n=100 were drawn and submitted to our graphical test.

We used the Generalized Lambda Distribution (GLD) to obtain an increasing sequence of bivariate skewed distributions. We chose the bivariate Uniform to have a flatter distribution than the bivariate normal. We chose the bivariate T with 8 d.f., to have a peaked distribution. Both of these were purely Kurtic (i.e. no symmetry problems).

Finally, we chose the Chi Square distribution with 10 d.f. as one which would be skewed and kurtic at the same time. But the degree of skewness would be consistent with that of the GLD used. And the degree of kurtosis, with that of the t distribution.

We sampled these distributions extensively in a Monte Carlo power study for our test (Romeu, 1990). In this paper, we undertake a visual study of the patterns obtained when samples come from such distributions.

Existing multivariate normality GOF tests are not graphical. Ours allows the user not only to accept or reject, but to obtain a sense of where to go next (i.e. what does de alternative distribution looks like) in the last case.

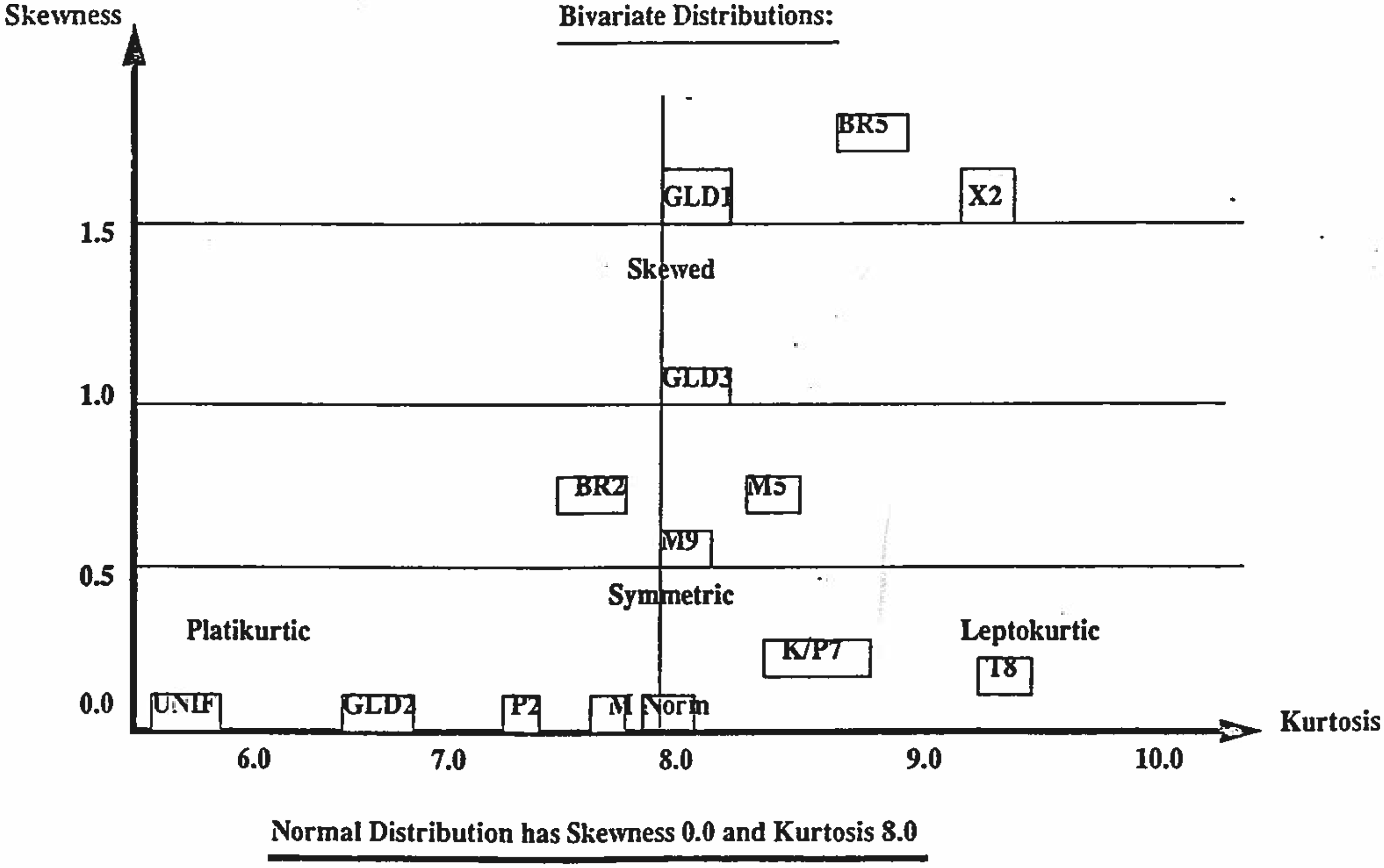


Figure 2. Statistical Distributions in the Skewness vs. Kurtosis plane.

3.0 Purely Skewed Distributions.

The pattern shown in Figure 3 corresponds to a sample taken from GDL1, a bivariate distribution with skewness of 1.5 and kurtosis of 8.0 (purely skewed). We see how the endpoint of the sample linked vector falls off the upper left quadrant of the confidence interval. This allows us to reject bivariate normality at all (90%, 95%, 99%) levels.

In addition, we observe the distinct pattern of the sample linked vector, sharply increasing in the first half, then changing direction before ever crossing the null distribution pattern.

We can use this test pattern to recognize a purely skewed non normal alternative.

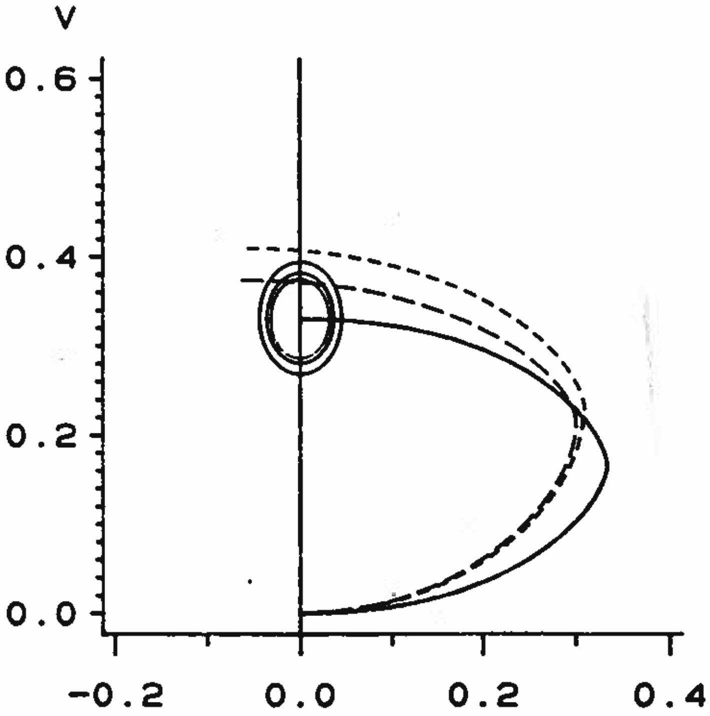


Figure 3.

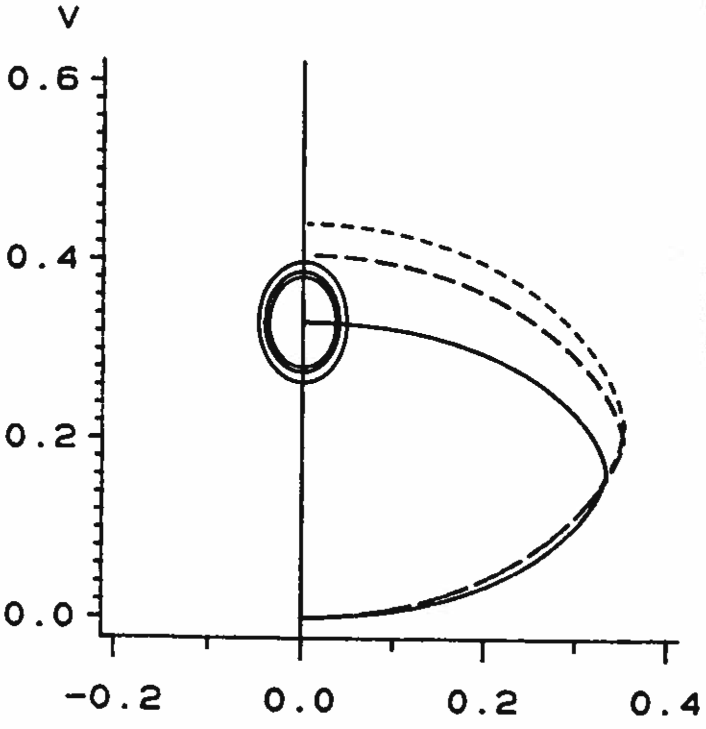


Figure 4.

4.0 Purely Kurtic Distributions.

In Figure 4 we show a sample from a bivariate Uniform, having skewness of 0.0 and kurtosis of less than 6.0 (purely kurtic).

We observe how the sample vector endpoint now falls way above the confidence ellipse, but on its vertical center axis. In addition, we can now observe a totally different vector pattern on the two halves of the linked vector path. The first half is much closer to the null pattern, crosses it and changes direction beyond the point where the null pattern does. Then, the upper half moves sharply upward.

This is a totally different pattern from the one presented in Figure 3, and allows us to recognize the sample as coming from a symmetric, flat, non normal distribution.

The pattern corresponding to a leptokurtic distribution is not presented for lack of space, but is also distinctive (Table 1).

5.0 Combined Skewed/Kurtic Distributions.

In Figure 5, we show a sample of pattern from a mixed skewed/kurtic distribution: the Bivariate Chi Square with 10 d.f.

We can see how the sample linked vector endpoint also falls off the confidence ellipse, entirely to its left. We can also notice how the vector pattern remains, during its entire trajectory, before (1st. half) or below (2nd half) the pattern of the null vector, which it never crosses. This characteristic pattern of the sample linked vector also provides a distinct test visualization.

It is with this graphical visualization that we recognize this sample as coming from a combined skewed and peaked non normal distribution.

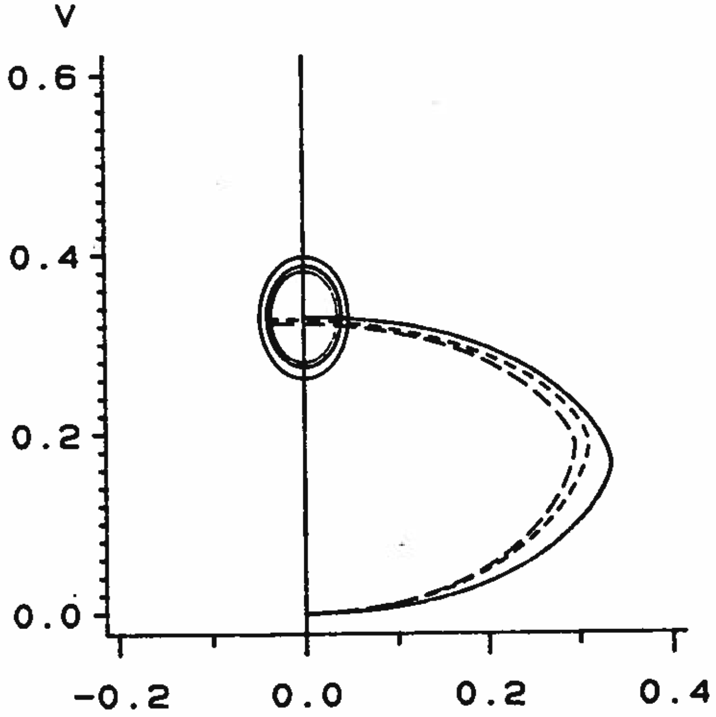


Figure 5.

6.0 Results.

On Table 1, we show a graphical comparison of six patterns of the sample linked vectors. The patterns have been subdivided into three parts: (i) lower and (ii) upper halves of the vector trajectory and (iii) endpoint position with respect to the confidence ellipse.

We have classified the six non normal distributions under study into three groups: (i) skewed, (ii) kurtic and (iii) combination. GLD-2 and GLD-3 are two bivariate distributions generated using the Generalized Lambda Distribution. Its objective is to provide, for comparison, an intermediate (skewed/kurtic) result between, respectively, the flat bivariate Uniform and the highly skewed GLD-1.

We can observe three distinctly different linked vector patterns, corresponding to these three non overlapping distribution classes. Hence, when rejecting multivariate normality, it is now also possible to identify the pattern of the sample and then to, (i) make an educated guess as to which (non normal) alternative distribution we want to test next. Or we may want to (ii) assess the type of non normal departure we are dealing with in order to implement the transformations necessary to redress the problems.

None of the other multivariate normality tests studied provides this capability.

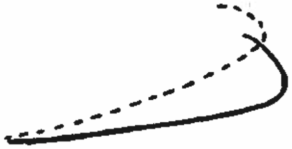
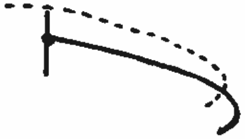
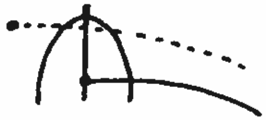
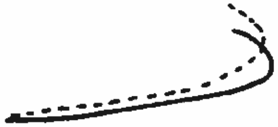
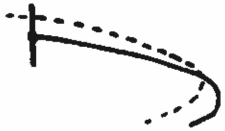
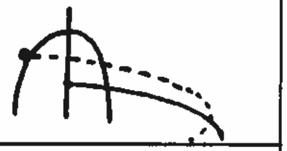
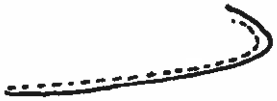
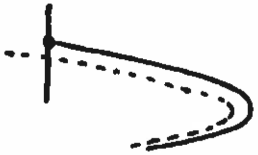
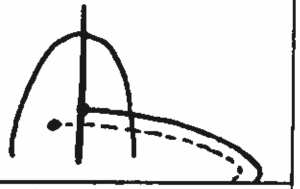
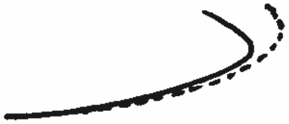
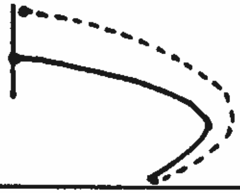
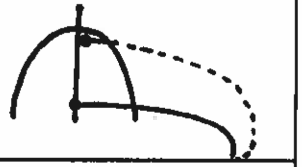
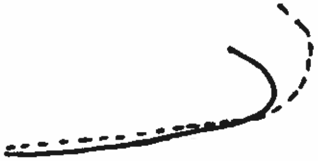
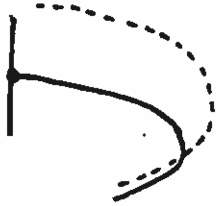
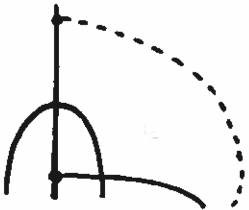
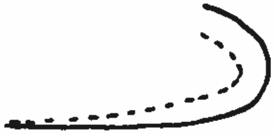
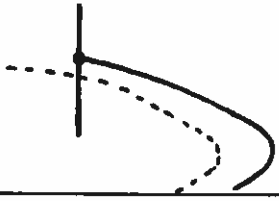
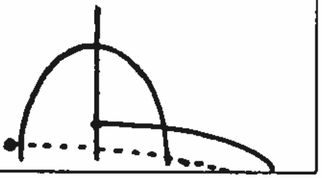
6.0 Bibliography.

Ozturk, A. and J. L. Romeu (1992), A New Method for Assessing Multivariate Normality With Graphical Applications, Communications in Statistics; Simula., 21(1), pages 15-34.

Romeu, J. L. (1990), Development and Evaluation of a General Procedure for Assessing Multivariate Normality, CASE Center Technical Report No. 9022, CASE Center of Syracuse University. Syracuse, NY. 13244-4100.

Table 1.

Vector Patterns:

Shape	Distrib.	1st. Half	2nd. half	Endpoint
Skewed	**GLD - 1**			
	GLD - 3			
Kurtic	**t (8)**			
	GLD - 2			
	UNIFORM			
Both	**Chi Square**			

A Comparative Visualization Study of a Graphical Multivariate Normality GOF Test.

JORGE LUIS ROMEU

SUNY-CORTLAND

PRESENTED AT THE INTERFACE '92 SIMPOSIUM

TEXAS A. AND M. UNIVERSITY

MARCH 18-21, 1992

The present paper studies a statistical method for assessing distributional assumptions of multivariate data, with graphical applications.

This new GOF test fills an existing gap in the multivariate GOF area. We have developed a statistically powerful procedure, applicable to relatively small samples from multivariate populations of an arbitrary number of p-variates. In addition, our proposed procedure can also be graphically implemented.

Statistical graphical procedures (Wang (1978)) are principally used (a) to describe the data (e.g. histograms), (b) to ascertain informally their statistical hypotheses (e.g. residual plots in regression) and (c) to aid in the calculation of values (e.g. power nomograms in ANOVA). They are lacking, in the multivariate context, due to the higher dimensionality problem. Some current graphical procedures that describe the multivariate data include Chernoff faces, Andrews charts, line/star profiles, scattergrams and decomposition into principal/factor components.

But in general, multivariate graphical methods are few, complex to interpret, difficult to implement and quite constrained. There has been, however, increasing interest in these procedures, both in theoretical (Wilk and Gnanadesikan (1968), Cleveland (1987)) as well as in more applied data analysis (Fisher (1983), include the recent ASA paper) statistical procedures.

With the current development of fast computing capabilities, the situation has

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

improved significantly. Our graphical test statistic, denoted $Q_n = (U, V)$, where u and v are the coordinates of the point Q_n, is based on some functions of the ordered statistic. Q_n can be represented by a linked vector in a two dimensional space. And the test can also be analyzed graphically by examining whether the point Q_n falls within the confidence ellipse and whether the linked vector follows a well defined pattern (the null pattern), both derived under the null hypothesis of multivariate normality. Therefore, our test can be performed analytically and graphically in the same rigorous statistical way.

Some graphical examples developed for a set of bivariate alternatives, following a design pattern that studies increasing skewness, kurtosis and a combination of these, follows. We have developed a table of graphical characteristics of our statistic, according to each of these non normal alternatives. These serve the purpose of identifying a possible alternative when multivariate normality is rejected. Finally, the well known *Setosa Dataset* of Fisher (19xx), is used as an example of data set that does not conform to multivariate normality because it tends to be more leptokurtic and skewed than what would be expected.