Computers ind. Engng Vol. 21, Nos 1–4, pp. 191–195, 1991
Printed in Great Britain.

0360-8352/91 $3.00 + 0.00

A NEW MULTIVARIATE NORMALITY GOODNESS OF FIT TEST WITH GRAPHICAL APPLICATIONS

JORGE LUIS ROMEU*

Department of Mathematics
SUNY Cortland
Cortland, NY 13045

ABSTRACT. A new procedure for assessing distributional assumptions of multivariate data, with graphical applications, is presented and illustrated using a real industrial example. The procedure is based on transforming the multivariate sample into a set of uncorrelated samples and representing the order statistics of each transformed sample by linked vectors in a two dimensional space.

1.0 Introduction.

Often, the industrial engineer (I.E.) deals with sets of multivariate data and statistical procedures that require the data follows the multivariate normal distribution (MVN). For example, the I.E. may have to generate a stream of pseudorandom MVN data. Or the application of a given statistical model (e.g. regression, discriminant analysis) may require that the data is distributed as multivariate normal. In both examples, assessment of MVN is performed by submitting the data to one of the existing Goodness-of-Fit (GOF) procedures.

However, some procedures may handle only a reduced number of p-variates (e.g. Cox and Small (1978)), or require that the sample size is large (e.g. Hawkins (1981)). Others, may depend heavily on the underlying and unknown correlation among p-variates (e.g. Royston (1983)), or may require the calculation of empirical critical values at every instance (e.g. Malkovich and Afifi (1973)). Some may converge slowly to their asymptotic distribution (Mardia (1970)) or may exhibit algorithmic and computational problems (Koziol (1982)). Others, may be exclusively graphical and informal in nature (Andrews et al. (1973)) or may lack graphical capabilities at all, as most above. Or may exhibit more than one of the mentioned problems. For a state-of-the-art review of MVN GOF procedures, see Romeu (1990).

Our new MVN GOF procedure, a multivariate extension of the univariate graphical GOF procedure of Ozturk and Dudewicz (1990), overcomes or alleviates most of the above problems. It is based on representing the multivariate sample by linked vectors in a two dimensional space. It is graphical and analytical, and can handle any sample size and any number of p-variates.

In the rest of this paper we discuss our new procedure. In Section 2, we outline our test procedure, its properties and its implementation via the Choleski Decomposition of the covariance matrix. In Section 3, we analyze with it, some real industrial data. And in Section 4, we summarize our results.

2.0 Methods.

The univariate procedure of Ozturk and Dudewicz is thus defined:

Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be n ordered univariate observations from a continous distribution $\mathcal{F}$, with μ, σ the location and scale parameters. Let: $Y_{i:n} = (X_{i:n} - \overline{\mathbf{X}})/s, \quad i = 1, \ldots, n$, where $\overline{\mathbf{X}} = \sum_i^n X_i/n$ and $s = \{\sum_i^n (X_i - \overline{\mathbf{X}})^2/(n-1)\}^{\frac{1}{2}}$. Also, let $m_{i:n}$, $i = 1, \ldots, n$, be the expected value of the i^{th} order statistic from the standard normal. Then, define the i^{th} linked vector, $V_{i:n}$, by its length and direction in the following manner:

(1) length of $V_{i:n}$ is: $a_i = \frac{1}{n}|Y_{i:n}|$, i.e. the length of the i^{th} standardized order statistic, and,
(2) direction of $V_{i:n}$ is defined by:

$$\theta_i = \pi\Phi(m_{i:n}) = \pi \int_{-\infty}^{m_{i:n}} \frac{1}{\sqrt{2\pi}} exp\{t^2/2\}\, dt$$

where θ_i forms an angle in the range $(0, \pi)$, with the abscissas axis.

*Work performed under a Dr. Nuala McGann Drescher UUP/NYS SUNY Faculty Award and Cornell Supercomputer Facility grant, while the author was a Visiting Researcher at the CASE Center, Syracuse University, Syracuse, NY

Then, by beginning the construction of our linked vectors at the origin of the Euclidean plane, and linking each $V_{i:n}$, $i = 1, \ldots, n$, one after the other, we obtain a line vector chart with a resultant vector $\overrightarrow{OQ_n}$. Under the null hypothesis of normality, it follows a certain pattern with endpoint $Q_n = (U_n, V_n)$, and U_n and V_n are approximately distributed as a bivariate normal. Therefore, we can obtain an approximate $100(1-\alpha)\%$ confidence ellipse for Q_n (Johnson and Kotz (1970)) with the quadratic function g:

$$g(Q_n) = g(U_n, V_n) = \frac{U_n^2}{\sigma_U^2} + \frac{(V_n - \mathcal{V}_n)^2}{\sigma_V^2} \sim \chi_2^2$$

When testing $p \geq 2$ *independent* samples, the multisample univariate procedure Q_n, is implemented in the following way:

(i) For each sample, standardize and sort separately the data, as if performing the Q_n procedure on each sample, individually.

(ii) For each sample, calculate the individual Q_n statistic, say $Q_n^j, \quad j = 1, \ldots, p$.

(iii) Obtain $\min_{1 \leq j \leq p}(g(Q_n^1), \ldots, g(Q_n^p))$, say $g(Q_n^*)$.

(iv) Solve, for α, in the equation $g(Q_n^*) = -2ln\alpha$, say α^*.

The univariate Q_n can be generalized to the multivariate case. Let $\mathbf{X} = (X_1, \ldots, X_p)$ denote a multivariate normal random vector with mean μ and covariance matrix Σ. Since Σ is positive definite we can define a $p \times p$ matrix L such that $LL' = \Sigma$. This decomposition of Σ is not unique, which makes our multivariate Q_n test a general procedure.

The **multivariate** Q_n (Ozturk and Romeu (1991)) is defined by the following two steps:

(1) perform a linear transformation:

$$\{(X_1, \ldots, X_p) \sim MVN_p(\mu, \Sigma)\} \quad \longmapsto \quad \{(Z_1, \ldots, Z_p) \sim MVN_p(0, I_p)\}$$

$$where \quad \mathbf{X} = CZ + \mu \quad and \quad CC' = \Sigma$$

(2) test these resulting p *independent samples* defined by $(Z_1, \ldots, Z_p)$, for joint (univariate) normality using the multisample univariate Q_n procedure above explained.

Our implementation of Q_n is via matrix L, the Cholesky decomposition of the covariance matrix Σ.

(1) L is lower triangular
(2) fulfilling: $LL' = \Sigma$

A step-by-step algorithm of Q_n for a p-variate sample of size n, $(X_1, \ldots, X_p)$ is:

(1) obtain $\overline{\mathbf{X}}$, the p-variate sample mean vector;
(2) obtain $\mathcal{S}$, the $p \times p$ sample covariance matrix;
(3) obtain L^*, the decomposition of $\mathcal{S}$, where $L^* L^{*\prime} = \mathcal{S}$.
(4) obtain L^{*-1}, the inverse of the decomposition of $\mathcal{S}$;
(5) obtain $Z_i = L^{*-1}(X_i - \overline{\mathbf{X}})$, $i = 1, \ldots, n$, the standardized values;
(6) Apply the *multisample univariate* Q_n procedure to this colection of p independent samples of size n.
(7) If the multisample univariate Q_n rejects the *joint* normality of the p univariate samples (i.e. if there is at least one endpoint lying outside the joint confidence ellipse for multisample univariate Q_n), then reject the hypothesis of *multivariate normality.*
(8) Otherwise, accept the multivariate normality of the sample under study.

Some properties of the multivariate Q_n are:

(1) it is location and scale independent, allowing the handling of multivariate data expressed in different scales and making our procedure very general.
(2) it is covariance independent, hence independent of the structure of the covariance or correlation matrices.
(3) it can handle an arbitrary number of p-variates.

3.0 An Industrial Example.

To illustrate the use and power of our Q_n procedure we analyze the dataset in Zemroch (1986). The data consists of 12 variables from a petroleum analysis problem, from which we have selected six: X_1: research octane number (RON); X_2: sensitivity (=RON - motor octane number); X_3 $\Delta R(100)$ (=RON - RON of fraction boiling below 100^0C); X_6: final boiling point; X_7: aromatic content and X_8: olefinic content.

These variables were selected to have enough ($p = 6$) as to preclude the use of several concurrent methods. We have a large enough dataset ($n = 88$) as to make Monte Carlo calculations of critical values expensive. Finally, the p-variates exhibit low, medium, high, positive and negative correlations among them. Such correlations affect several other MVN GOF methods (see Table 3.1).

TABLE 3.1: CORRELATIONS.

	x-1	x-2	x-3	x-6	x-7
x-2	0.573				
x-3	0.098	-0.053			
x-6	-0.013	0.085	0.076		
x-7	0.286	0.772	0.246	0.251	
x-8	0.080	0.460	-0.550	0.014	0.025

We start by looking at the marginals, since a necessary but not sufficient condition for MVN is that all marginals are univariate normal (Gnanadesikan (1977)). A descriptive analysis shows how X_6 and X_8 are right skewed (See Table 3.2) and a formal univariate GOF analysis using Ozturk and Dudewicz univariate procedure (Figure 3.1) confirms the non-normality of the marginal data. Corresponding p-values are shown next to the linked vectors endpoints.

TABLE 3.2: DESCRIPTIVE STATISTICS.

VAR.	MIN	Q1	MEDIAN	MEAN	Q3	MAX
X-1	88.3	91.23	93.60	94.03	97.28	99.10
X-2	5.2	6.80	7.85	8.40	9.97	12.60
X-3	0.1	4.90	8.60	7.81	10.30	15.20
X-6	160.8	177.90	183.75	193.36	207.00	262.50
X-7	12.7	26.57	33.95	34.14	41.67	57.20
X-8	0.4	1.13	2.70	4.44	6.60	21.60

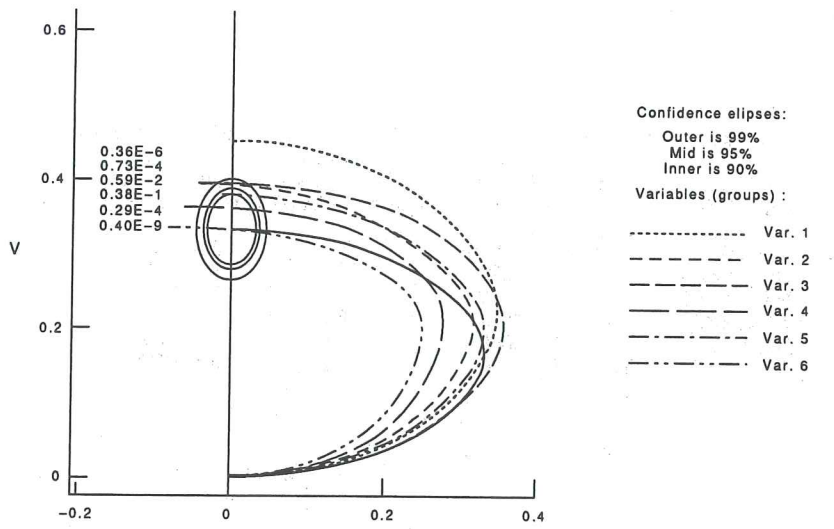

FIGURE 3.1: UNIVARIATE GOF OF THE MARGINALS.

Figure 3.2 shows the MVN GOF graphical analysis of the data using our Q_n. We see how two of the linked vectors fall way outside the confidence ellipses, thus rejecting the multinormality of the data.

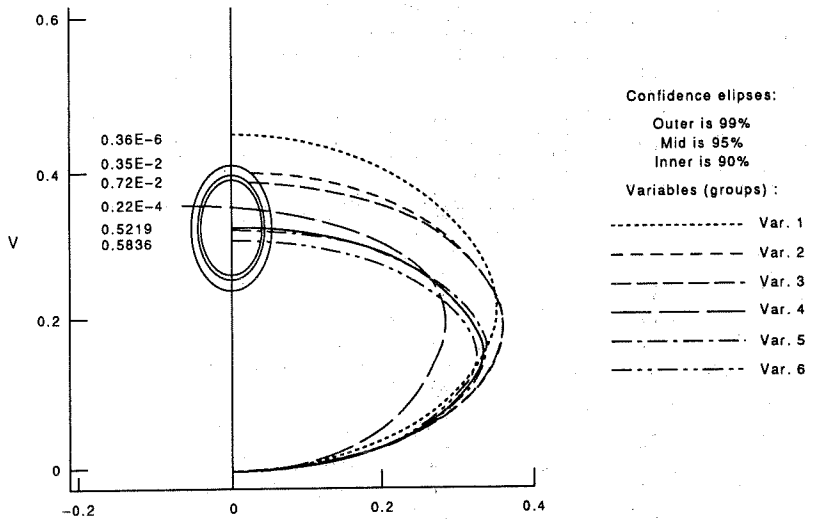

FIGURE 3.2: Q_m GRAPHICAL MULTIVARIATE NORMALITY GOF.

For comparison, we reanalyze the data using Mardia's Skewness and Kurtosis, Malkovich-Afifi's, Hawkins' and Royston's MVN GOF tests. The first two converge slowly to their asymptotic critical values, the next two are empirical and Royston's showed dependence on correlation among p-variates.

We use the critical values obtained by Monte Carlo, in Romeu (1990). Results are shown in Table 3.3.

TABLE 3.3: MULTIVARIATE ANALYSIS BY DIFFERENT METHODS.

MVN GOF METHOD	TEST STAT.	PVALUE -	EMPIRICAL C.VAL.(0.01)	DECISION: MULTINORMAL
Q(N)	N/A	0.36E-6	N/A	REJECT
MAR-SKEW	143.35	N/A	56.77	REJECT
MAR-KURT	0.648	N/A	-2.22/2.24	ACCEPT
MALK-AFIFI	0.874	N/A	0.920	REJECT
HAWKINS	1.571	N/A	1.681	ACCEPT
ROYSTON	159.09	N/A	15.34	REJECT

4.0 Conclusions.

We have discussed and illustrated the use of our new MVN GOF Q_n procedure. We have seen how it is used where others fail to be applicable and how it succeeds where some applicable ones fail. In addition, we have shown the use of the empirical (Monte Carlo) critical values obtained by Romeu (1990). Empirical critical values are suggested when using MVN GOF methods under conditions where the asymptotic ones are invalid or unavailable.

TABLE 3.4: TABLE OF EMPIRICAL CRITICAL VALUES (ROMEU (1990)).

TABLE NO. C-7 CRITICAL VALUES FOR THE CASE P = 5 VARIATES.

RHO=0.5 N	%	SKEWNESS TEST	KURTOSIS LOWER	KURTOSIS UPPER	ROYSTON W	MALKOVICH AFIFI	KOZIOL CHI-SQR.	COX-SMAL REG	HAWKINS TEST	KOZIOL ANGLES
25	90	35.29	-1.59	0.26	8.25	0.827	0.155		0.924	11.54
25	95	38.72	-1.70	0.50	10.07	0.806	0.190		1.165	17.64
25	99	46.25	-1.93	1.07	14.26	0.752	0.283		1.717	>1000
50	90	40.90	-1.62	0.77	8.45	0.919	0.150		0.950	13.67
50	95	44.66	-1.77	1.11	10.37	0.906	0.186		1.146	24.30
50	99	54.57	-2.07	1.86	15.09	0.878	0.268		1.668	>1000
75	90	43.16	-1.68	1.01	8.37	0.947	0.151		0.961	12.23
75	95	47.37	-1.86	1.34	10.42	0.940	0.186		1.185	19.98
75	99	56.45	-2.28	2.11	15.34	0.921	0.283		1.681	>1000
100	90	43.50	-1.70	1.16	8.36	0.960	0.151		0.969	11.49
100	95	47.82	-1.91	1.55	10.31	0.954	0.193		1.155	15.89
100	99	57.44	-2.32	2.33	14.74	0.940	0.268		1.627	>1000
125	90	44.07	-1.73	1.21	8.17	0.969	0.153		0.962	11.79
125	95	48.50	-1.94	1.56	9.92	0.964	0.191		1.189	16.35
125	99	57.19	-2.32	2.29	14.46	0.953	0.280		1.773	>1000
150	90	44.72	-1.70	1.28	8.53	0.973	0.150		0.956	12.66
150	95	48.86	-1.89	1.68	10.75	0.969	0.188		1.189	18.43
150	99	57.10	-2.28	2.45	15.20	0.959	0.269		1.688	>1000
175	90	44.77	-1.73	1.31	8.83	0.976	0.153		0.969	11.86
175	95	48.85	-1.98	1.70	10.97	0.973	0.190		1.204	15.80
175	99	57.10	-2.50	2.47	15.79	0.965	0.282		1.754	>1000
200	90	45.06	-1.69	1.28	8.97	0.979	0.147		0.933	12.96
200	95	48.98	-1.93	1.60	11.09	0.976	0.180		1.154	22.23
200	99	57.69	-2.36	2.43	15.80	0.970	0.264		1.599	>1000

REFERENCES

Andrews, D. F., Gnanadesikan R. and J. L. Warner, *Methods for Assessing Multivariate Normality*, Multivariate Analysis, Academic Press, 1973.

Cox, D. R. and N. J. H. Small, *Testing Multivariate Normality*, Biometrica **65** (1978), 263–272.

Gnanadesikan, R., *Methods of Statistical Data Analysis of Multivariate Observations*, Wiley, 1977.

Hawkins, D. M., *A New Test for Multivariate Normality and Homoscedasticity*, Technometrics **23** (1981), 105–110.

Koziol, J. A., *A Class of Invariant Procedures for Assessing Multivariate Normality*, Biometrika **69** (1982), 423–427.

Malkovich, J. F. and A. A. Afifi, *On Tests for Multivariate Normality*, JASA **68** (1973), 176–179.

Mardia K. V., *Measures of Multivariate Skewness and Kurtosis With Applications*, Biometrika **57** (1970), 519–530.

Ozturk, A. and E. J. Dudewicz, *A New Statistical Goodness-of-fit Test Based on Graphical Representation*, Technical Report No. 52, Department of Mathematics, Syracuse University, 1990.

Romeu, J. L., *Development and Evaluation of A General Procedure for Assessing Multivariate Normality*, Ph.D. Dissertation, Syracuse University (Published as CASE Center Technical Report No. 9022), 1990.

Ozturk, A and J. L. Romeu, *A New Graphical Test for Multivariate Normality*, (submitted for publication).

Royston, J. P., *Some Techniques for Assessing Multivariate Normality Based on the Shapiro Wilk W*, Applied Statistics **32** (1983), 121–133.

Zemroch, P. J., *Cluster Analysis as an Experimental Design Generator, With Application to Gasoline Blending Experiments*, Technometrics **28(1)** (1986), 39–49.