

Assessing Multivariate Normality through Interactive Dynamic Graphics

Luca Scrucca

Dipartimento di Scienze Statistiche

Università degli Studi di Perugia

E-mail: luca@stat.unipg.it

Summary: The assumption of multivariate normality is often at the basis of many statistical analysis. In this article we propose a graphical method based on the characterization of the multivariate normal distribution in terms of univariate normality of all linear combinations of the variables in the set. We review some methods for choosing directions to look for departure from the hypothesis of normality, and we propose an interactive dynamic graphic approach for checking the joint distribution. Examples from simulated data and a real dataset will be used to show practical implementation of the ideas outlined.

Key words: Multivariate normal, Projections, Assessing normality, Probability plot, Interactive Graphical Methods.

1. Introduction

Many statistical methods for continuous variables assume an underlying multivariate normal distribution. For instance, classical multivariate methods rely on such hypothesis (Krzanowski, 1990), but also more recent contributions in the literature, such as Gaussian graphical models (Whittaker, 1990), deserve a central role to the multivariate normality.

Thus, several formal tests were developed for checking the multivariate normality of a set of random variables. Graphical techniques have

been proposed too, and they proved to be particularly useful in practical applications. Compared to methods based on hypothesis testing, graphical methods are more resistant to outlying observations and allow to evaluate the influence of any single observation on departures, if any, from the reference model.

Often in practical applications checking for multivariate normality is restricted to a control of the marginal distributions. This relies on the assumption that, except for pathological cases, the joint nonnormality is likely to be detected by studying the marginal normality of each variable. Moreover, as the dimension of the random variable \mathbf{x} increases, testing for joint normality becomes quite hard.

The literature on assessing multivariate normality is quite extensive. Some methods rely on evaluating marginal normality of each variable, such as D'Agostino and Pearson test (1973), Shapiro and Wilk test (1965), D'Agostino test statistic (1971), marginal Box-Cox transformation and associated likelihood ratio test (Box and Cox, 1964), normal QQ probability plot (Wilk and Gnanadesikan, 1968). Other methods try to evaluate the joint normality of a set of random variables through a single statistic which summarizes the whole distribution, such as: the nearest distance test and chi-square QQ plot proposed by Andrews et al. (1973), Mardia's measures of multivariate skewness and kurtosis (Mardia, 1970), beta probability plots (Fattorini, 1982), the generalization of Shapiro and Wilk test to the multivariate case proposed by Malkovich and Afifi (1973) with refinements discussed by Fattorini (1986), F-probability plot (Ahn, 1992), a generalization of the Box-Cox transformation to the multivariate case and the corresponding likelihood ratio statistic for testing a set of transformation parameters (Velilla, 1993). A further class of methods exploit the geometrical idea that a normal distribution in p dimensions must show normality in any direction.

In this article we propose a graphical method for assessing multivariate normality based on unidimensional views. In Section 2 we define notation and the building-block of our approach. In Section 3 we investigate some methods for choosing the directions to look for departures from the hypothesis of normality. Then, we propose a dynamic graphic approach for checking the joint distribution. An implementation of dy-

dynamic graphic using QQ-plot is discussed in Section 4. The final section presents examples from simulated data and a real dataset which will be used to show practical implementation of the ideas outlined.

2. Assessing multivariate normality

Let $\mathbf{X} = (X_1, X_2, \dots, X_p)$ be a vector of p random variables distributed as a multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, that is

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

If the matrix $\boldsymbol{\Sigma}$ has full rank p , then the joint density function exists and can be written as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Our graphical proposal relies on the characterization of the multivariate normal distribution in terms of univariate normality of all linear combinations of the variables. Formally, a random vector \mathbf{X} of dimension p is said to have a p -variate normal distribution if and only if the linear combination $\boldsymbol{\alpha}^\top \mathbf{X}$ has a univariate normal distribution for all vectors $\boldsymbol{\alpha} \in \mathbb{R}^p$. This means that if $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the following holds for all $\boldsymbol{\alpha} \neq \mathbf{0}$:

$$\boldsymbol{\alpha}^\top \mathbf{X} \sim N(\boldsymbol{\alpha}^\top \boldsymbol{\mu}, \boldsymbol{\alpha}^\top \boldsymbol{\Sigma} \boldsymbol{\alpha})$$

Thus, we can use the union-intersection principle of Roy (1953) for testing the hypothesis of joint multivariate normal distribution. Let $H_0(\boldsymbol{\alpha})$ be the null hypothesis under which $\boldsymbol{\alpha}^\top \mathbf{X} \sim N$. Taking all possible values of $\boldsymbol{\alpha}$, provided that $\|\boldsymbol{\alpha}\| = 1$, we could test all possible univariate directions and if and only if $H_0(\boldsymbol{\alpha})$ is true for every $\boldsymbol{\alpha}$, we could conclude that \mathbf{X} has a multivariate gaussian distribution. In other words, the null hypothesis can be expressed as $H_0 = \bigcap_{\boldsymbol{\alpha}} H_0(\boldsymbol{\alpha})$ and it must be rejected if at least one of the $H_0(\boldsymbol{\alpha})$ is rejected.

The main characteristic of this class of methods is that we move from a problem in high dimension to infinite problems in low dimension. This

strictly resembles the approach for checking regression models proposed by Cook and Weisberg (1997). Although in a different context, the idea is to take advantage of using dynamic graphics for checking several directions with relatively little effort. Of course, there is no guarantee we will detect a direction of nonnormality, but with a suitable choice of the directions we may be able to look along several “interesting” directions. Nevertheless, most tests for checking multivariate normality select just one direction based on a suitable measure of nonnormality.

3. *Choosing the directions*

In the previous section we show that the problem of assessing multivariate normality may be reformulated in terms of assessing the normality of unidimensional projections along all possible directions. The problem may be further simplified by looking at some specified directions rather than all possible directions with a suitable choice of the “interesting” directions. In this section we discuss how to choose such directions in order to detect possible lack of normality. In the next section, a dynamic graphic approach which leads to easily analyze several projections in a relatively small amount of time (much will depend on the computer power and on the dimension of data set) will be discussed.

Suppose $\tilde{\mathbf{X}} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^\top$ is a $n \times p$ matrix of observations on p variables, hence in our notation each row \mathbf{x}_i is a vector of length p , for $i = 1, \dots, n$. As long as we are interested in assessing multivariate normality, the first step is checking the *marginal distributions*, that is the distribution of each variable ignoring the remaining $(p - 1)$ variables. In terms of the coefficient of the vector $\boldsymbol{\alpha}$, this simply means to assign 1 to element corresponding to the selected variable and zero elsewhere.

Orthogonal directions of maximum variance can be obtained through the principal components analysis (Krzanowski, 1990, pp. 48–74). *Principal component directions* represent projections into lower-dimensionality subspaces, with decreasing variability. Hence, it is reasonable to look at such directions, where possible nonnormal features are likely to be detected.

Let

$$\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$$

and

$$\mathbf{S} = (n - 1)^{-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

be the sample mean vector and sample covariance matrix respectively. The p principal components are given by the columns of the $n \times p$ matrix

$$\mathbf{Y} = \tilde{\mathbf{X}}\mathbf{V}$$

where \mathbf{V} is the $p \times p$ matrix of eigenvectors of \mathbf{S} , that is $\mathbf{S} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$. Furthermore, since \mathbf{V} is an orthogonal matrix, the principal components are uncorrelated.

This basic idea was also used by Srivastava and Hui (1987), who proposed a generalization of the Shapiro-Wilk statistic based on principal components. They obtained two test statistics as a function of the transformed Shapiro-Wilk statistic computed on each component. Furthermore, they provided the approximated null distribution for both test statistics.

Selection of specific *directions of nonnormality* were discussed by Andrews et al. (1971). They obtained a projection onto the unidimensional subspace which has been chosen to be sensitive to particular types of nonnormality. The method is also discussed by Gnanadesikan (1977), and it is outlined in the following.

Let the matrix of standardized values be

$$\mathbf{Z} = \mathbf{H}\tilde{\mathbf{X}}\mathbf{S}^{-1/2}$$

where $\tilde{\mathbf{X}}$ is again the $(n \times p)$ matrix of n observations on p variables, \mathbf{H} is the $(n \times n)$ centering matrix, and $\mathbf{S}^{-1/2}$ is the inverse of square-root of the covariance matrix, whose dimension is $(p \times p)$. The suggested method relies on studying a normalized weighted sum of the observations in the transformed scale. Let \mathbf{z}_i^\top be a vector of p measurements for the

i -th observation, so we can define

$$\mathbf{d}_\delta^\top = \frac{\sum_{i=1}^n w_i \mathbf{z}_i^\top}{\left\| \sum_{i=1}^n w_i \mathbf{z}_i^\top \right\|}$$

where $w_i = \|\mathbf{z}_i\|^\delta$ are the weights for each observation. The vector \mathbf{d}_δ depends on the parameter δ : if $\delta > 0$, the greater the distance of a point in the space \mathbb{R}^p is from the center of the points (which is a vector of zeroes, since variables have been standardized), the more weight it will receive. The converse is true for $\delta < 0$, whereas for $\delta = 0$ all the observations will have equal weight. Thus, we can say that for positive values of δ the vector \mathbf{d}_δ will tend to point toward any clustering of observations distant from the mean, whereas for negative values of δ it will point toward any clustering close to the center of the data. The vector \mathbf{d}_δ is defined in the space of the standardized values, then a projection in the original space is given by $\mathbf{S}^{1/2} \mathbf{d}_\delta$.

A further set of directions we may want to check are those given by *residuals from conditional regressions*, that is by taking the residuals from the regression of each variable on the remaining. In fact, it is well known that if a set of variables are jointly normally distributed, then the residuals from the linear regression of one variable on the remaining should show no pattern. Otherwise, any non linear dependencies should be detected by a non linear shape on the plot of the residuals vs any linear combination of the predictors. This idea was also used by Cox and Small (1978) in their test for normality.

Let \mathbf{X} be partitioned as $\mathbf{X} = [X_1, \mathbf{X}_2]^\top$ where $X_1 \in \mathbb{R}^1$ and $\mathbf{X}_2 \in \mathbb{R}^{p-1}$. The corresponding vector of means and the covariance matrix are

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} v_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{21} & \boldsymbol{\Sigma}_2 \end{bmatrix}$$

where v_{11} is a scalar, $\boldsymbol{\Sigma}_2$ is $(p-1) \times (p-1)$ matrix and \mathbf{v}_{21} is a vector of dimension $(p-1) \times 1$ such that $\mathbf{v}_{12} = \mathbf{v}_{21}^\top$. Under the hypothesis that $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the conditional distribution of X_1 given \mathbf{X}_2 is

$$X_1 | (\mathbf{X}_2 = \mathbf{x}_2) \sim N(\mu_1 + \mathbf{v}_{12} \boldsymbol{\Sigma}_2^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2), v_{11} - \mathbf{v}_{12} \boldsymbol{\Sigma}_2^{-1} \mathbf{v}_{21})$$

The mean function $\mathbb{E}(X_1 | \mathbf{X}_2 = \mathbf{x}_2) = \mu_1 + \mathbf{v}_{12} \boldsymbol{\Sigma}_2^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)$ is linear and, under the hypothesis of multivariate normal distribution, the residuals should be normally distributed. Clearly, this must hold for any variable selected as response in the regression. Therefore, we can think to regress any variable on the remaining ones, and then check for the normality of the residuals.

Additional directions can be obtained randomly, requiring that the vector of coefficients have unit length. These *random directions* do not rely on any measure of nonnormality, but it is likely that an inspection of several of them might be able to detect departures from normality, if any.

4. Dynamic probability plot

Given the linear combination $y_i = \boldsymbol{\alpha}^\top \mathbf{x}_i$ ($i = 1, \dots, n$), we could use a wide range of graphical techniques in order to assess univariate normality, for instance histograms, box-plots, probability plots, and kernel density estimation. Our graphical approach will use the well-known QQ probability plot (Wilk and Gnanadesikan, 1968).

A normal probability plot allows to graphically compare the ordered observations $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ with the quantiles obtained from the normal distribution corresponding to a cumulative probability p_i . If we note by Φ the cumulative density function for a normal distribution, the theoretical quantile \tilde{y}_i is the value corresponding to $\Phi(\tilde{y}_i) = p_i$, where p_i is usually defined as $i/(n+1)$ or $(i-0.5)/n$. A QQ probability plot is then obtained plotting $y_{(i)}$ vs \tilde{y}_i , and when the hypothesis of normality holds it should be approximately straight.

Assuming we know the vector $\boldsymbol{\alpha}$, we can compute the projection $y_i = \boldsymbol{\alpha}^\top \mathbf{x}_i$ for all i . Geometrically, this is a direction in the space \mathbb{R}^p , and a normal probability plot is therefore useful for assessing whether or not the distribution is normal in specified directions. Dynamic graphic methods will help us to recover as many directions as we want with a little effort.

A visual enhancement may be added to the probability plot in order to help in assessing departure from a straight line. Recalling an idea in-

roduced by Atkinson (1981) for checking residuals in linear regression models, we constructed via simulation a reference envelope for normality.

Let $y_{(i)}$ ($i = 1, \dots, n$) be the ordered empirical quantiles corresponding to the selected direction. Assume the hypothesis of normality holds, i.e. y is sampled from a normal random variable with mean μ_y and variance σ_y^2 . Plugging-in the usual unbiased estimates for the mean and the variance, we generate n values from a normal distribution with mean \bar{y} and variance s_y^2 and we order them. If we repeat the latter step m times, we obtain a sequence of samples as

$$(y_{(1)j}, y_{(2)j}, \dots, y_{(n)j}) \quad j = 1, \dots, m$$

For a fixed i , the sequence

$$(y_{(i)1}, y_{(i)2}, \dots, y_{(i)m})$$

provide the empirical distribution of the i -th ordered statistic. Then, we pick up the smallest and the largest values for each plotting position, that is for each $i = 1, \dots, n$ we define

$$y_{(i)L} = \min y_{(i)m}$$

$$y_{(i)U} = \max y_{(i)m}$$

Joining all the smallest $y_{(i)L}$ and the largest values $y_{(i)U}$ for $i = 1, \dots, n$, we obtain a simulated envelope under the normality assumption. If this hypothesis holds we expect that all the points in the probability plot will fall within the envelope a fraction $m/(m+1)$ of the times; for $m = 19$ this means that 95% of the times we should observe all the points lying within the envelope.

5. Examples

In this section we provide examples that will be used to discuss the proposed method in practical applications. The first two examples concern artificial datasets, while the third concerns a real data set.

A computer implementation based on the ideas outlined was also developed. The computer code was written in the Xlisp-Stat programming language (Tierney, 1990), and using the statistical software called Arc. This is a computer software developed by Cook and Weisberg (1999), and can be downloaded from the web-site <http://www.stat.umn.edu/arc>. Because of the high level dynamic plots we can obtain with Xlisp-Stat, the choice of the environment has been straightforward. In addition, with Arc we took advantage of an excellent software and its libraries of functions.

Among the main features of the code is the chance to select the desired directions for checking. Then, they are computed simultaneously and a probability plot for the first direction selected is drawn. A slider on the plot allows the user to cycle through all the chosen directions. Moreover, a button is available for drawing the simulated envelope discussed in Section 4. The Xlisp-Stat code that implements the dynamic probability plot is available from the the web page: <http://www.stat.unipg.it/~luca>.

Nevertheless, the graphical approach introduced in this paper does not depend on the specific computer environment.

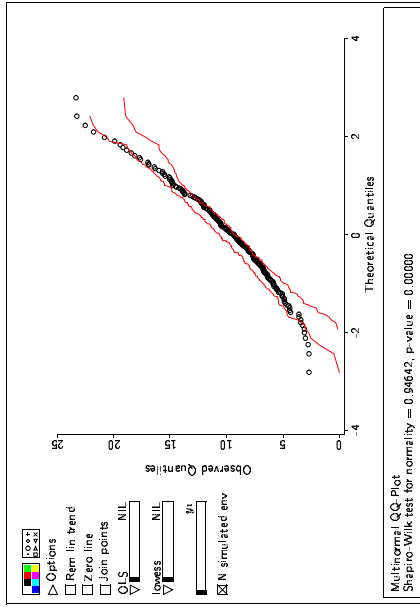
5.1 Simulated data

As first example, we generated a dataset of 200 observations from a bivariate correlated $\chi^2_{(10)}$ distribution. The simulation scheme starts by generating two independent variables from a nonnormal distribution (z_1, z_2) , and then combining z_1 and z_2 to form correlated variables, y_1 and y_2 , by using the linear transformation:

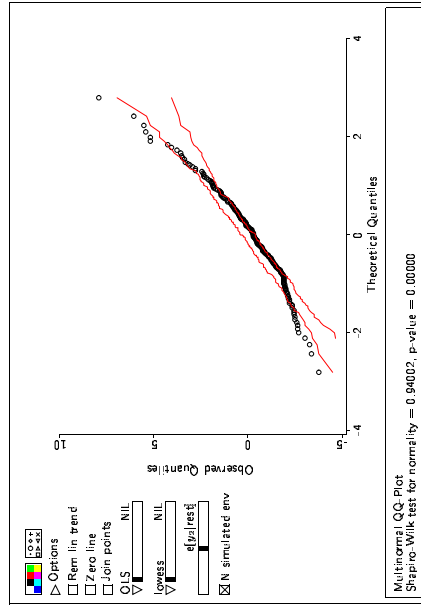
$$\begin{cases} y_1 = & z_1 \\ y_2 = & \rho z_1 + \sqrt{1 - \rho^2} z_2 \end{cases} \quad (1)$$

where ρ is the correlation coefficient between y_1 and y_2 .

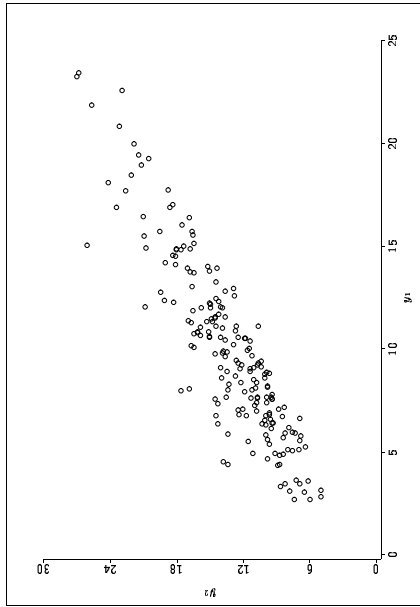
In our example we generated two independent chi-square random variables, which are then correlated following (1) with $\rho = 0.9$, such that (y_1, y_2) have marginal distributions relatively close to normal, the second being expected to be more nearly normal than the first.



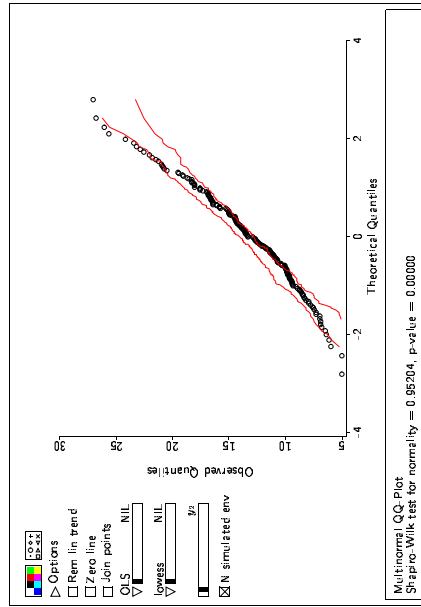
b) marginal direction for y_1



d) direction for the residuals from $E(y_2|y_1)$



a) Scatterplot of y_1 vs y_2



c) marginal direction for y_2

Figure 1. Scatterplot and QQ probability plots for some directions for the example on bivariate correlated $\chi^2_{(10)}$ variables.

Panel a) of Figure 1 shows a scatterplot for the data generated, while the remaining panel contains QQ probability plots for some directions. Departures from normality are evident in all checking plots.

For the second example we generated data from a multivariate normal distribution with mean and covariance matrix equal to

	Mean	Covariance Matrix				
x_1	10	10	2	4	5	3
x_2	0	2	50	13	8	1
x_3	5	4	13	12	1	4
x_4	-10	5	8	1	4	10
x_5	100	3	1	4	10	20

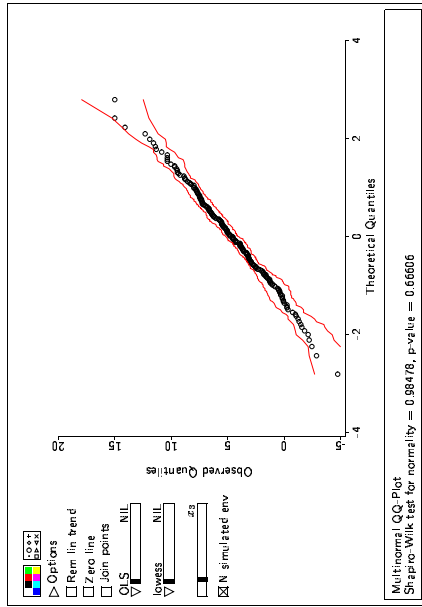
Graphs on Figure 2 show the QQ probability plots for some selected directions. In all the directions checked no departure from the hypothesis of normality has been detected.

5.2 Haematology data

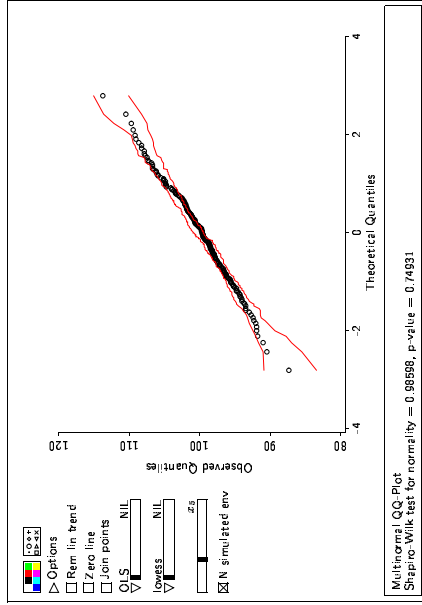
This dataset is taken from a health survey of paint sprayers in a car assembly plant. The 103 observations were measured on black workers. The variables are:

H	haemoglobin concentration
PCV	packed cell volume
WBC	white blood cell count
LY	lymphocyte count
N	neutrophil count
L	serum lead concentration

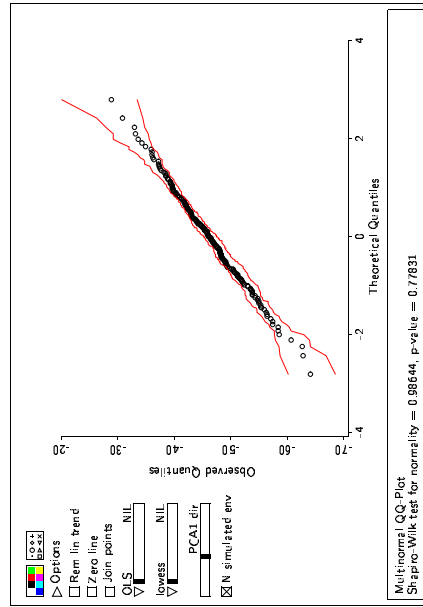
Royston (1983) analyzed these data for testing multivariate normality. Variates WBC, LY, N, and L were logarithmically transformed before the analysis since they showed skewed distributions, and because is a common practice with this kind of data. Royston noted that the six marginal distributions look reasonably linear by inspection of normal probability plots, but the joint distribution does not appear to be normal. In addition, he outlined the presence of three outliers (cases number 21, 47, 52)



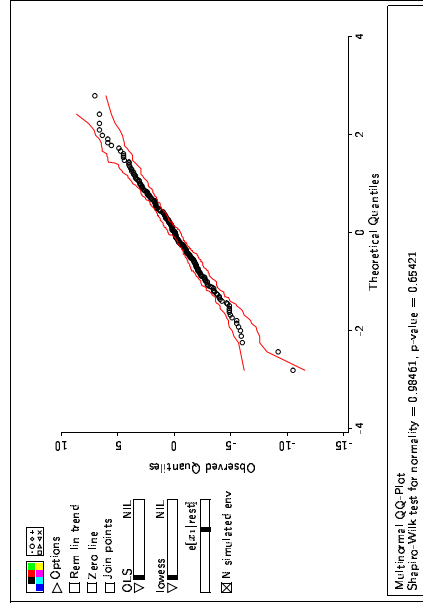
a) marginal direction for x_3



b) marginal direction for x_5



c) first principal component direction



d) direction for the residuals from $E(x_1 | \text{rest})$

Figure 2. Some QQ probability plots for a sample from a multivariate normal distribution.

and, after removing them, checking for joint distribution does not show evidence of departure from normality.

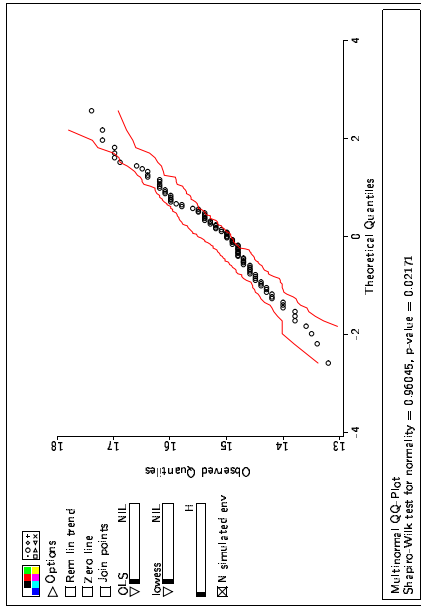
Li, Fang and Zhu (1997) analyzed this dataset too. Their goal was testing elliptically symmetry for some subsets of the six variables, and they concluded that all the subsets considered were from elliptical distributions even though some were not from normal distribution.

As a first step we graphically checked the set of variables proposed by Royston. The marginal distribution of \mathbf{H} and some directions, particularly those given by the residuals of the conditional regressions, showed evidence against the hypothesis of normality. Figure 3 reports some of these plots, together with a simulated envelope and the Shapiro-Wilk statistic. It is quite clear that the hypothesis of normality is not sustained, in particular in the directions given by the regression residuals. Moreover, the observations number 10, 21, 47, 52, and 80 have the largest Mahalanobis distance and appear to be outliers. Removing them do not improve the normality, except perhaps in the marginal distribution of the variable \mathbf{H} .

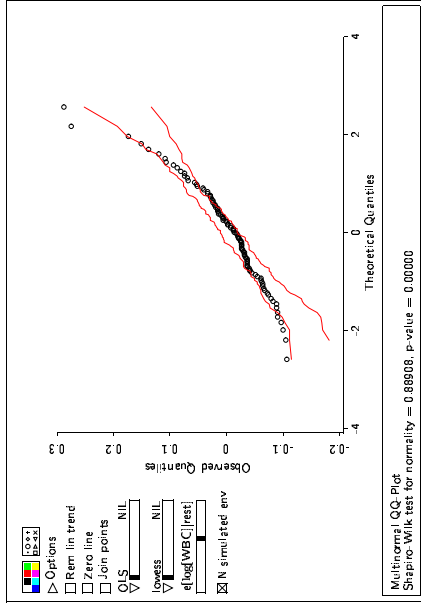
The transformations proposed by Royston rely on medical practice, but do not seem to achieve joint normality. Therefore, we look for a better set of simultaneous power transformations using the multivariate Box-Cox approach proposed by Velilla (1993). Basically, the method assumes there exists a vector $\Lambda = (\lambda_1, \dots, \lambda_p)^\top$ of transformation parameters such that when we transform each X_j ($j = 1, \dots, p$) the following model holds:

$$\mathbf{X}^{(\lambda)} = (X_1^{(\lambda_1)}, \dots, X_p^{(\lambda_p)}) \sim N_p(\boldsymbol{\mu}^{(\lambda)}, \boldsymbol{\Sigma}^{(\lambda)})$$

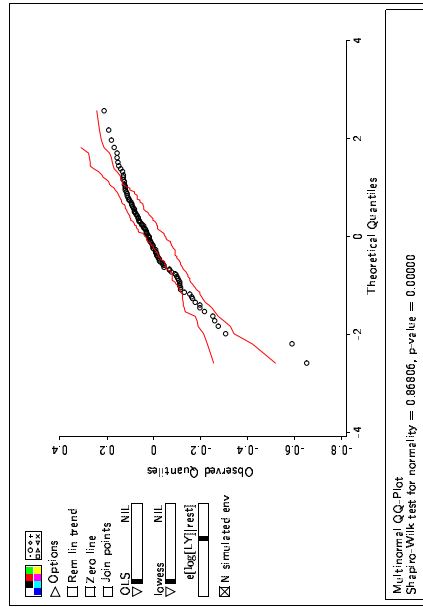
where $\boldsymbol{\mu}^{(\lambda)}$ and $\boldsymbol{\Sigma}^{(\lambda)}$ are respectively the mean vector and the covariance matrix in the transformed scale. The estimates for the λ s are obtained numerically maximizing the corresponding log-likelihood profile. A likelihood ratio test is also available for testing the hypothesis of a simultaneous set of transformation parameters.



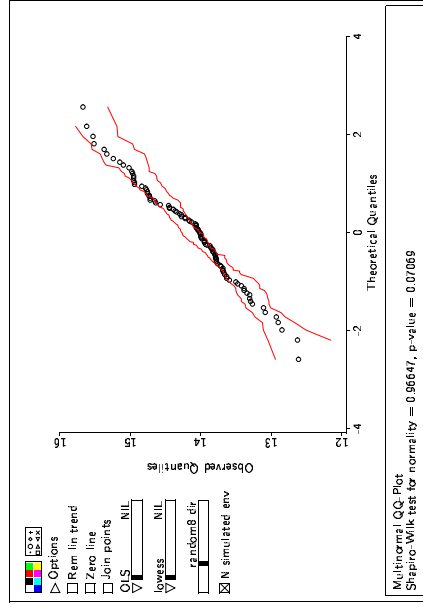
a) marginal direction for H



b) direction for the residuals from $E(\log[WBC] | \text{rest})$



c) direction for the residuals from $E(\log[LY] | \text{rest})$



d) a random direction

Figure 3. *QQ probability plots for some specified directions for assessing the multivariate normality of the set (H, PCV, log[WBC], log[LY], log[N], log[L])*

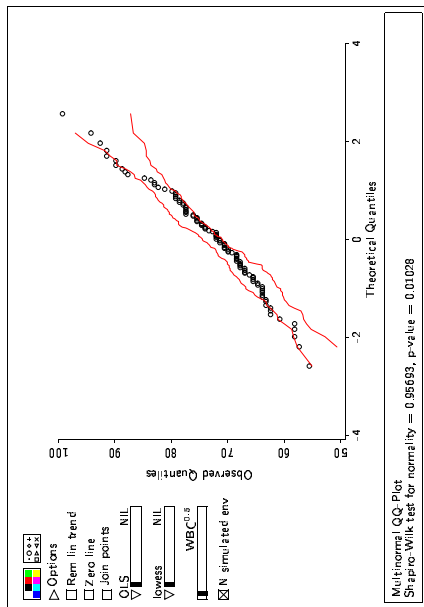
Table 1. Summary of transformations to normality

Variable	$\hat{\lambda}_i$	SE	Wald test for $\lambda_i = 0$	Wald test for $\lambda_i = 1$
H	-2.035	1.254	-1.62	-2.42
PCV	0.758	1.058	0.72	-0.23
WBC	0.401	0.210	1.91	-2.85
LY	0.479	0.116	4.12	-4.48
N	0.491	0.245	2.01	-2.08
L	-0.698	0.456	-1.53	-3.73

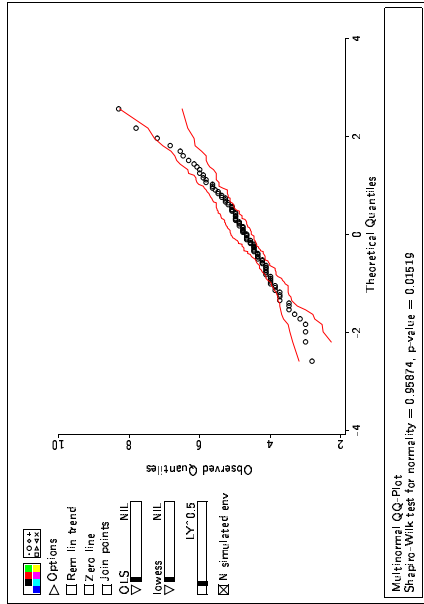
LRT for all $\lambda_i=0$: 23.554 on 6 df, p-value = .001
LRT for all $\lambda_i=1$: 44.155 on 6 df, p-value = .000
LRT for $\Lambda = (1, 1, 0, 0, 0, 0)$: 25.574 on 6 df, p-value = .000
LRT for $\Lambda = (-2, 1, 0.5, 0.5, 0.5, -1)$: 0.742 on 6 df, p-value = .994
LRT for $\Lambda = (-2, 1, -0.5, 0.5, 0.5, -1)$: 28.193 on 6 df, p-value = .000

Table 1 shows the results from applying the multivariate Box-Cox transformation. Based on the likelihood ratio test (LRT) reported, it seems clear that the previous logarithmic transformations are not supported by the data, whereas a large p-value can be obtained by the following transformations: H^{-2} , PCV, $WBC^{0.5}$, $LY^{0.5}$, $N^{0.5}$, L^{-1} . Checking for the normality of this set seems to be supported through several directions, but not in the directions involving $WBC^{0.5}$ and $LY^{0.5}$ (see Figure 4). The power transformation of WBC that marginally lead to normality is -0.5, but this does not allow to get an an approximate joint normality for the whole set of variables. For the variable LY the marginal power transformation is the logarithmic. If we remove this two variables from the set and we check the joint normality of the remaining variables, the hypothesis of multivariate normality is now sustained.

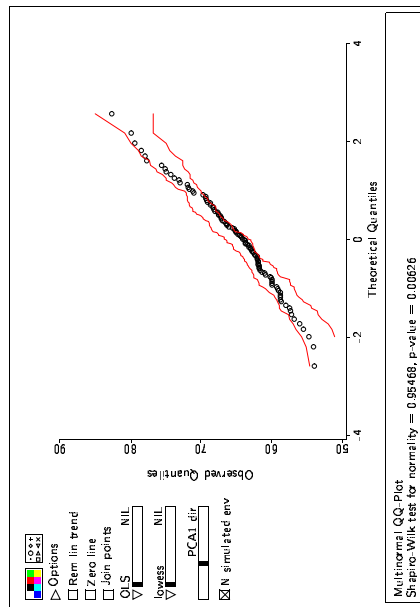
As further refinement, we added in turns one of the two variables removed and we check for multivariate normality. It turns out that the variable LY in the logarithmic scale may be added to the previous set without violating the hypothesis of normality (see plots on Figure 5).



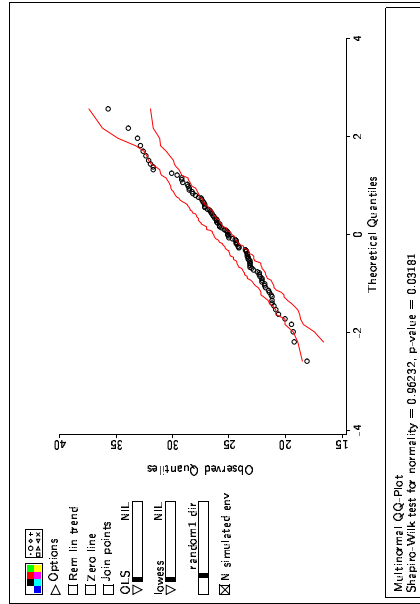
a) marginal direction for WBC^{0.5}



b) marginal direction for LY^{0.5}

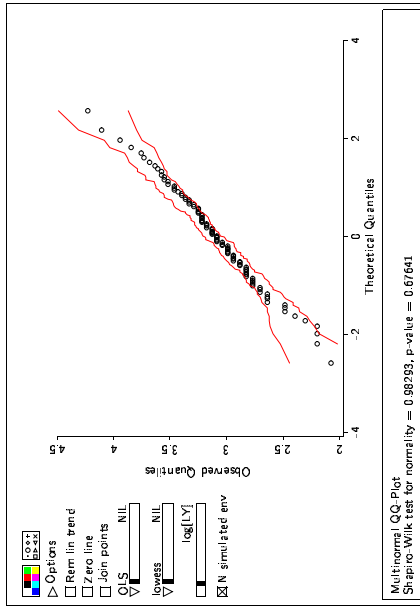


c) first principal component direction

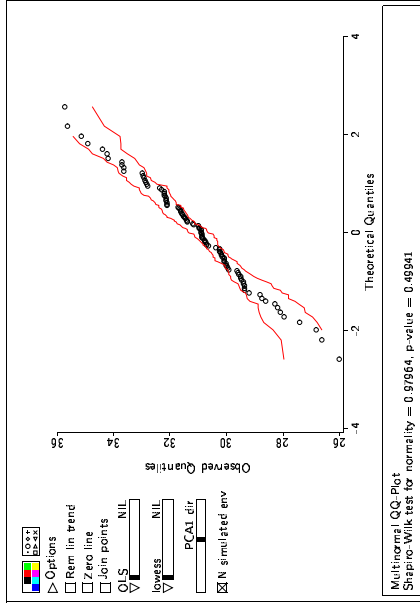


b) a random direction

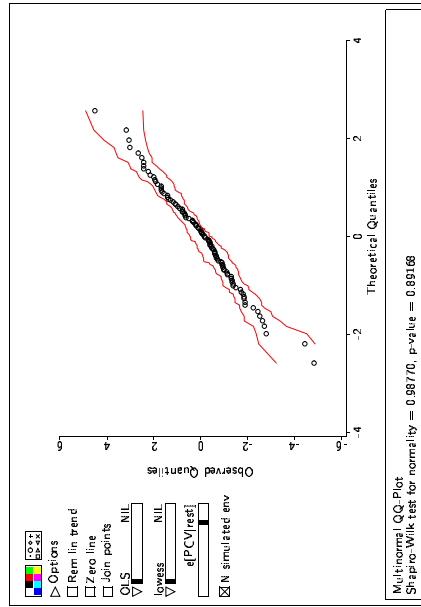
Figure 4. QQ probability plots for some specified direction for assessing the multivariate normality of the set $(H^{-2}, PCV, WBC^{0.5}, LY^{0.5}, W^{0.5}, L^{-1})$



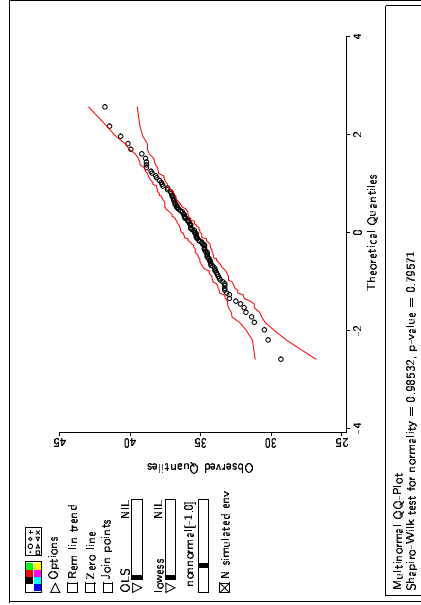
a) marginal direction for H



b) direction of residuals from $E(\log[WBC] | \text{rest})$



c) direction of residuals from $E(\log[LY] | \text{rest})$



d) a direction of nonnormality

Figure 5. *QQ probability plots for some specified directions for assessing the multivariate normality of the set $(H^{-2}, PCV, \log[LY], N^{0.5}, L^{-1})$*

6. Final remarks

The methodology discussed in this article tackles the relevant problem of assessing the multivariate normality of a set of variables. The approach proposed is based on the characterization of the multivariate normal distribution in terms of normality of any linear combinations of variables in the set. Each linear combination represents a direction in the space spanned by the set of variables under study. Some methods for choosing the directions to look for departures from the reference model have been discussed. Several directions may be easily checked using a dynamic graphic approach. In our proposal we used the well-known QQ probability plot for assessing the normality in any selected direction. Then, the Roy's union-intersection principle is used to establish whether the hypothesis under study is reasonable or not.

Examples based on simulated and real datasets have been provided, and they showed that our approach is particularly effective in detecting departures from normality. Furthermore, the graphical investigation might suggest a subset of the original variables as a candidate subset for normality.

Further developments of this approach could involve the selection of other relevant directions to look for departure from normality.

Acknowledgments: This work was supported by funds from the MURST.

References

D'Agostino R.B., Pearson E.S. (1971) An omnibus test of normality for moderate and large size samples, *Biometrika*, 58, 341–348.

D'Agostino R.B., Pearson E.S. (1973) Test for departure from normality. Empirical results for the distribution of $b_2/\sqrt{b_1}$, *Biometrika*, 60,

613–622.

Ahn S.K. (1992) F-probability plot and its application to multivariate normality, *Communication in Statistics - Theory and Method*, 21, 997–1023.

Andrews D.F., Gnanadesikan R., Warner J.L. (1971) Transformations of multivariate data, *Biometrika*, 27, 825–840.

Andrews D.F., Gnanadesikan R., Warner J.L. (1973) Methods for assessing multivariate normality, in *Multivariate Analysis III* (P.R. Krishnaiah, ed), Academic Press, New York.

Atkinson A.C. (1981) Robustness, transformations and two graphical displays for outlying and influential observations in regression, *Biometrika*, 68, 13–20.

Box G.E.P., Cox D.R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society*, B 26, 211–252.

Cook R.D., Weisberg S. (1997) Graphics for Assessing the Adequacy of Regression Models, *Journal of the American Statistical Association*, 92, 490–499.

Cook R.D., Weisberg S. (1999) *Applied Regression Including Computing and Graphics*, Wiley, New York.

Cox D.R., Small N.H. (1978) Testing multivariate normality, *Biometrika*, 65, 263–272.

Fattorini L. (1982) Assessing multivariate normality on beta plots, *Statistica*, 42, 251–257.

Fattorini L. (1986) Remarks on the use of the Shapiro-Wilk statistic for testing multivariate normality, *Statistica*, 46, 209–217.

Gnanadesikan R. (1977) *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York.

Krzanowski W.J. (1990) *Principles of multivariate analysis*, Oxford University Press, Oxford.

Li R.Z., Fang K.J., L.X. Zhu (1997) Some Q-Q probability plots to test spherical and elliptical symmetry, *Journal of Computational and Graphical Statistics*, 4, 435–450.

Malkovich J.F., Afifi A.A. (1973) On tests for multivariate normality, *Journal of American Statistical Association*, 68, 176–179.

Mardia K.V. (1970) Measures of multivariate skewness and kurtosis

with applications, *Biometrika*, 57, 519–530.

Roy S.N. (1953) On a heuristic method of test construction and its use in multivariate analysis, *Annals of Mathematical Statistics*, 24, 220–238.

Royston J.P. (1983) Some techniques for assessing multivariate normality based on the Shapiro-Wilk W , *Applied Statistics*, 52, 121–133.

Shapiro S.S, Wilk M.B. (1965) An analysis of variance test for normality, *Biometrika*, 52, 591–611.

Srivastava M.S., Hui T.K. (1987) On assessing multivariate normality based on Shapiro-Wilk W Statistic, *Statistics & Probability Letters*, 5, 15–18.

Tierney L. (1990) *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, Wiley, New York.

Velilla S. (1993) A note on the multivariate Box-Cox transformations to normality, *Statistics and Probability Letters*, 17, 315–322.

Whittaker J. (1990) *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.

Wilk M.B., Gnanadesikan R. (1968) Probability plotting methods for the analysis of data, *Biometrika*, 55, 1–17.