Statistical Methods & Applications (2004) 13: 179–196
DOI: 10.1007/s10260-004-0095-1

SMA

Robust multivariate transformations to normality: Constructed variables and likelihood ratio tests

Marco Riani

Sezione di Statistica e Informatica, Dipartimento di Economia, Università di Parma, Italy
(e-mail: mriani@unipr.it)

Abstract. The assumption of multivariate normality provides the customary powerful and convenient ways of analysing multivariate data: if the data are not normal, the analysis may often be simplified by an appropriate transformation. In this context, the most widely used test is the likelihood ratio, which requires the maximum likelihood estimate of the transformation parameter for each variable. Given that this estimate can only be found numerically, when the number of variables is large (> 20) it is impossible or infeasible to compute the test. In this paper we introduce alternative tests which do not require the maximum likelihood estimate of the transformation parameters and prove algebraically their relationships. We also give insights both using theoretical arguments and a robust simulation study, based on the forward search algorithm, about the distribution of the tests previously introduced.

Key words: Box-Cox transformation, Forward search, Score statistic, Wald statistic, Transformation to normality, Robust methods

1 Introduction

The analysis of data is often improved by using a transformation of the response, rather than the original response itself. There are physical reasons why a transformation might be expected to be helpful in some examples. If the data arise from a counting process, they will have a Poisson distribution and the square root transformation will provide an approximately constant variance, independently of the mean. Concentrations are nonnegative variables and so cannot be subject to additive errors of constant variance. The effect is most noticeable if there are observations both close to, and far from, zero.

The parametric family of power transformations introduced by Box and Cox (1964) was extended to multivariate data by Andrews et al. (1971) and by Gnanadesikan (1977). Velilla (1993) compares marginal and joint transformations and gives

further references to related work. A full discussion of univariate transformations, including deletion diagnostics, is given by Atkinson (1985). Similar techniques for multivariate transformations have been introduced by Atkinson (1995).

In this paper we consider transformations for multivariate data and explore the theoretical relationships between a likelihood ratio test which requires the maximum likelihood estimate of the transformation parameters and other tests based on the use of constructed variables which lead to a multivariate multiple regression formulation in which the parameters are different for each response. The estimates are related solely through the correlation of the responses. This structure is known in the literature as the seemingly unrelated regression model (SUR regression). The theoretical contribution of the paper is a theorem which proves the asymptotical equivalence of two constructed variables tests.

Given that both the likelihood ratio and the constructed variable tests are not robust and suffer from masking and swamping problems, in order to provide information about the presence of influential observations and the inferential effect of each unit, we embed all the tests inside a forward search approach (Atkinson and Riani 2000). The application of the forward search to transformation of the response in univariate regression data is described in Riani and Atkinson (2000). Examples of applications of the forward search to transformations in multivariate and discriminant analysis are given by Riani and Atkinson (2001). Up to now in the literature, the results of the transformation tests have been compared using the asymptotic confidence bands of the χ^2 distribution. The computational contribution of this paper is a robust simulation study to investigate the small sample distribution induced by the forward search on all the tests which have been introduced.

The structure of the paper is as follows. In Sect. 2 we recall the likelihood ratio test for testing multivariate transformations. In Sect. 3 we introduce other tests which, unlike the likelihood ratio, do not require the maximum likelihood estimate of the transformation parameters and give a theorem which proves their relationships. In Sect. 4 we perform a robust simulation study in order to investigate both the small sample and the asymptotic distribution of the tests previously introduced. Section 5 concludes and gives suggestions for further research.

2 Multivariate transformations to normality

For multivariate data let y_i be the $v \times 1$ vector of responses at observation i with y_{ij} the observation on response j. In the extension of the Box and Cox (1964) family to multivariate responses the normalized transformation of y_{ij} is

$$\begin{aligned} z_{ij}(\lambda_j) &= (y_{ij}^{\lambda_j} - 1)/\lambda_j \dot{y}_j^{\lambda_j - 1} \quad (\lambda \neq 0) \\ &= \dot{y}_j \log y_{ij} \qquad\qquad\quad (\lambda = 0), \end{aligned} \tag{1}$$

where $\dot{y}_j$ is the geometric mean of the jth response. The value $\lambda_j = 1$ $(j = 1, \ldots, v)$ corresponds to no transformation of any of the responses. We assume a multivariate linear regression model of the form

$$Z(\lambda) = (X_1\beta_1, \ldots, X_v\beta_v) + \Xi$$

where $Z(\lambda)$ is a $n \times v$ matrix of normalized responses whose ij-th generic element $z_{ij}(\lambda_j)$ is defined in Eq. (1). The X_j, $j = 1, \ldots, v$ are $n \times p$ design matrices not necessarily equal. The β_j are unknown vector of parameters and Ξ is an $n \times v$ random matrix whose rows are i.i.d. If the transformed observations are normally distributed with mean μ_i for the i-th observation and covariance matrix Σ, twice the profile loglikelihood of the observations is given by

$$\begin{aligned} 2L_{max}(\lambda) &= \text{const} - n \log |\hat{\Sigma}(\lambda)| - \sum_{i=1}^{n} \{z_i(\lambda) - \hat{\mu}_i(\lambda)\}^T \hat{\Sigma}^{-1}(\lambda) \{z_i(\lambda) - \hat{\mu}_i(\lambda)\} \\ &= \text{const} - n \log |\hat{\Sigma}(\lambda)| - \sum_{i=1}^{n} e_i(\lambda)^T \hat{\Sigma}(\lambda)^{-1} e_i(\lambda). \end{aligned} \tag{2}$$

where $z_i(\lambda) = (z_{i1}(\lambda_1), \ldots, z_{iv}(\lambda_v))^T$ denotes the i-th row of matrix $Z(\lambda)$. In (2) $\hat{\mu}_i(\lambda)$ and $\hat{\Sigma}(\lambda)$ are derived from least squares estimates for fixed λ and $e_i(\lambda)$ is the $v \times 1$ vector of residuals.

The calculation of $\hat{\mu}_i(\lambda)$ and $\hat{\Sigma}(\lambda)$ is simplified when the matrix of explanatory variables X is the same for all responses. As a result, the least squares estimates are found by independent regression on each response, yielding the $p \times v$ matrix of parameter estimates $B(\lambda) = (\hat{\beta}_1, \ldots, \hat{\beta}_v) = (X^T X)^{-1} X^T Z(\lambda)$. In other words, if the explanatory variables are the same for all responses the maximum likelihood estimators of β_j can be obtained regressing the j-th column of $Z(\lambda)$ (say z_{C_j}) on X $(j = 1, \ldots, v)$ (see for example Hamilton 1994, p. 318). On the other hand, when the X_j are different, maximum likelihood estimation requires iteration. In Sect. 3 we will exploit this fact in connection with a SUR type model.

When the X are the same, the maximum likelihood estimator of Σ is given by

$$\begin{aligned} n\hat{\Sigma}(\lambda) &= \sum_{i=1}^{n} e_i(\lambda) e_i(\lambda)^T \\ &= \{Z(\lambda) - XB(\lambda)\}^T \{Z(\lambda) - XB(\lambda)\}. \end{aligned} \tag{3}$$

When these estimates are substituted in (2), the profile loglikelihood reduces to

$$2L_{max}(\lambda) = \text{const}' - n \log |\hat{\Sigma}(\lambda)|. \tag{4}$$

So, to test the hypothesis $\lambda = \lambda_0$, asymptotically the statistic

$$T_{GLR} = n \log\{|\hat{\Sigma}(\lambda_0)| / |\hat{\Sigma}(\hat{\lambda})|\} \tag{5}$$

has a χ^2 distribution on v degrees of freedom. In Eq. (5) $\hat{\lambda}$ is the vector of v parameter estimates maximising (4), which is found by numerical search. Of course, it makes no difference in the (generalized) likelihood ratio test for the value of λ whether we use the maximum likelihood estimator of Σ or the unbiased estimator $\hat{\Sigma}_u(\lambda)$ where

$$(n - p)\hat{\Sigma}_u(\lambda) = \sum_{i=1}^{n} e_i(\lambda) e_i(\lambda)^T.$$

In the remaining part of this paper we will refer to Eq. (5) as the *generalized* likelihood ratio test. We will reserve the words likelihood ratio test, for the test in which we consider the log of the ratio of two likelihoods associated to two predetermined values of the vector λ.

3 Multivariate transformation tests based on constructed variables

In order to validate a particular transformation or to test its sensitivity to the presence of outliers or atypical observations, we require to find T_{GLR} for a variety of values of the vector $\lambda_0 = (\lambda_{01}, \ldots, \lambda_{0v})^T$ and for subsets of observations. Certainly, when the number of variables and/or the number of observations is large this may lead to computational problems. An advantage of the procedure of this section is that it brings the diagnostic method into the framework of multivariate regression analysis and does not require non linear numerical maximization procedures.

Constructed variables for the univariate Box-Cox transformation are described in detail in Atkinson (1985) and Atkinson and Riani (2000). The model is linearized by Taylor expansion which leads to inclusion in the regression of a constructed variable. More specifically, when the Box-Cox transformation is applied to multivariate data, linearization of the transformation (1) produces

$$z_{ij}(\lambda_j) \approx z_{ij}(\lambda_{0j}) + (\lambda_j - \lambda_{0j}) \left. \frac{\partial z_{ij}(\lambda)}{\partial \lambda_j} \right|_{\lambda_j = \lambda_{0j}} . \tag{6}$$

This leads to the nv values of the v constructed variables

$$\begin{aligned} w_{ij}(\lambda_0) &= \left. \frac{\partial z_{ij}(\lambda)}{\partial \lambda_j} \right|_{\lambda_j = \lambda_{0j}} \\ &= y_{ij}^{\lambda_{0j}} \log y_{ij} / (\lambda_{0j} \dot{y}_j^{\lambda_{0j}-1}) - z_{ij}(\lambda_0)(1/\lambda_{0j} + \log \dot{y}_j), \end{aligned} \tag{7}$$

in which the j-th response is differentiated with respect to λ_j. Provided the model for $z(\lambda)$ contains a constant, regression on (7) is equivalent, in the special cases of $\lambda = 1$ and 0, to regression on the variables

$$\begin{aligned} w(1) &= y\{\log(y/\dot{y}) - 1\} \qquad (\lambda = 1) \\ w(0) &= \dot{y} \log y (\log y/2 - \log \dot{y}) \ (\lambda = 0). \end{aligned} \tag{8}$$

In (8) the subscripts i and j have been omitted for typographic clarity.

If the explanatory variables are the same for each response, the combination of Eq. (6) and the regression model for the j-th response $z_{C_j}(\lambda) = X\beta_j(\lambda) + \epsilon_j$, where $z_{C_j}(\lambda_0)$ is the j-th column of matrix $Z(\lambda_0)$, yields

$$\begin{aligned} z_{C_j}(\lambda_0) &= X\beta_j(\lambda_0) - (\lambda_j - \lambda_{0j}) w_j(\lambda_0) + \epsilon_j \\ &= X\beta_j(\lambda_0) + \gamma_j w_j(\lambda_0) + \epsilon_j, \qquad j = 1, \ldots, v. \end{aligned} \tag{9}$$

where $w_j(\lambda_0)$ is the constructed variable for response j and $\gamma_j = -(\lambda_j - \lambda_{0j})$. Testing that λ_0 is the correct transformation of the response is equivalent to testing that the γ_j in (9) are zero. Note that, even if the matrix of explanatory variables X is

the same for all responses, Eq. (9) shows that inclusion of the constructed variables means that the variables for regression are no longer the same for all responses. As a result, the simplification of Sect. 2 no longer holds and the covariance Σ between the v responses has to be allowed for in estimation and independent least squares is replaced by generalized least squares. In the particular form (9) the parameters for each response are different, the estimates being related only through covariances of the $z_{C_j}(\lambda)$. This special structure is known as seemingly unrelated regression (SUR) (Zellner 1962). More in detail, if we suppose that matrix X is the same for all responses, the above set of v equations can be written as one combined equation

$$z^* = X^*\beta + W^*\gamma + \epsilon^* = (X^*,\ W^*)\delta + \epsilon^* \tag{10}$$

where z^* =vec$(z_{C_1}(\lambda_0), \ldots, z_{C_v}(\lambda_0))$,

$$X^* = I_v \otimes X, \qquad W^* = \begin{pmatrix} w_1(\lambda_0) & 0 & \ldots & 0 \\ 0 & w_2(\lambda_0) & \ldots & 0 \\ & & \ddots & \\ \ldots & \ldots & & \ldots \\ 0 & 0 & \ldots & w_v(\lambda_0) \end{pmatrix}, \tag{11}$$

$\delta = (\beta^T, \gamma^T)^T = (\beta_1^T, \ldots, \beta_v^T, \gamma_1, \ldots, \gamma_v)^T$ and $\epsilon^* \sim N(0,\ \Sigma \otimes I_n)$. The symbol $\otimes$ denotes the Kronecker product. In this form the model is that for a vector of nv observations on a heteroscedastic univariate response. Twice the profile loglikelihood of the observations stacked in the combined equation is given by

$$\begin{aligned} 2L_{max}(\lambda) = const - \log|\Sigma^{-1}(\lambda) \otimes I_n| \\ - \{z^* - (X^*,\ W^*)\delta\}^T (\Sigma^{-1}(\lambda) \otimes I_n) \{z^* - (X^*,\ W^*)\delta\}. \end{aligned} \tag{12}$$

Differentiation of Eq. (12) with respect to δ yields

$$\hat{\delta} = \left\{(X^*,\ W^*)^T (\Sigma^{-1}(\lambda) \otimes I_n) (X^*,\ W^*)\right\}^{-1} (X^*,\ W^*)^T (\Sigma^{-1}(\lambda) \otimes I_n) z^*. \tag{13}$$

This implies that for a fixed value of λ the maximum likelihood estimator of δ is nothing but the GLS estimator resulting from Eq. (10). Note that the SUR model (10) is not motivated by heteroscedasticity between independent rows, but rather as a convenient way to adapt for different explanatory variables, as created by the presence of the constructed variables $w_j(\lambda)$.

Because Eq. (13) contains $\Sigma^{-1}(\lambda) \otimes I_n$, estimation of Σ is required for the procedure to be operational. The estimation proceeds in two or more steps:

1. Obtain $\hat{\Sigma}_0$, an estimate of Σ from independent regressions with $z_{C_j}(\lambda_0)$ regressed on X and $w_j(\lambda_0)$.

2. Obtain an estimate of δ based on Eq. (13) using $\hat{\Sigma}_0^{-1}$ (seemingly unrelated regression step).

3. Iteration in the estimation of Σ and δ is possible, starting with the estimate of δ obtained from step 2 and repeating the seemingly unrelated regression step until there is no significant change in the estimates.

In order to test $H_0 : \gamma = 0$ we can use a Wald criterion applied to generalized least squares or the likelihood ratio test. In both cases, under the alternative hypothesis, although not under the null, independent least squares is replaced by seemingly unrelated regression.

3.1 Wald type statistic

In the context of SUR regression, the Wald test assumes the form

$$T_W = (r - R\hat{\delta})^T \left[R \left\{ (X^*,\ W^*)^T (\hat{\Sigma}^{-1}(\lambda) \otimes I)(X^*,\ W^*) \right\} R^T \right]^{-1} (r - R\hat{\delta}), \tag{14}$$

where, as usual, $r = R\delta$ is a known q-element vector and R is a known matrix of full row rank of order $q \times v(p+1)$. If we are interested in testing that all the coefficients of the constructed variables are equal to zero, $\gamma_j = 0$, $j = 1, \ldots, v$, $q = v$, $r = 0$ and $R = (0, I_v)$ where 0 denotes a zero matrix of dimension $v \times pv$.

3.2 Likelihood ratio test

In order to test $\gamma = 0$ we can also use a modified version of the generalized likelihood ratio test introduced in the previous section. The two determinants in Eq. (5) can be replaced respectively by $|E^T E|$ and $|E^T(w)E(w)|$. $E = (e_{C_1}, \ldots, e_{C_v})$ is the matrix of the residuals obtained from independent regression of each $z_{C_j}(\lambda_0)$ on each set of explanatory variables

$$\begin{aligned} E &= (z_{C_1}(\lambda_0) - X\hat{\beta}_1(\lambda_0), \ldots, z_{C_v}(\lambda_0) - X\hat{\beta}_v(\lambda_0)) \\ &= Z(\lambda_0) - XB(\lambda_0). \end{aligned}$$

$E(w) = (e_{C_1}(w), \ldots, e_{C_v}(w))$ is the matrix of residuals obtained applying SUR to the regression model which includes X and w_j (Eq. 10). In other words: $\text{vec}(E(w)) = z^* - (X^*, W^*)\hat{\delta}$.

This leads to the following expression

$$T_{LR} = n \ln \frac{|E^T E|}{|E^T(w)E(w)|}. \tag{15}$$

Two versions of the statistics given in Eqs. (14) and (15) are possible. They differ in the estimated covariance matrix $\hat{\Sigma}(\lambda)$ which can either be calculated from the residuals of independent regression or iterated from this starting point using SUR regression.

It is clear that the two tests given in Eq. (14) and (15) tackle the problem of testing $\gamma_j = 0$, $j = 1, \ldots, v$ using different mathematical instruments. We have called these two tests likelihood ratio and Wald, but they should have been called Wald-score and likelihood ratio-score, because they use a constructed variable found using a Taylor series expansion (that is a score argument).

3.3 Asymptotic equivalence

The main theoretical contribution of this paper is the following theorem which states the relationship between the two tests.

Definition. *Asymptotic Equivalence*. Two sequences of tests $\{T_{1,n}, n = 1, 2, \dots\}$ and $\{T_{2,n}, n = 1, 2, \dots\}$, with a common parameter space $\mathcal{F}_n$ for each n, are asymptotically equivalent if (Le Cam 1986, Le Cam and Yang 1990, Cox and Hinkley 1974)

$$T_{1,n} - T_{2,n} \to o_p(1) \qquad \text{as} \qquad n \to \infty$$

where the general notation $o_p(n^a)$ means a random variable Z_n such that for any $\epsilon > 0$

$$\lim_{n\to\infty} \text{pr}(n^{-a}|Z_n| > \epsilon) = 0$$

and $o_p(1)$ denotes a random variable that is $o(1)$ in probability.

Theorem. *for testing the null hypothesis $H_0 : \gamma_j = 0$, $j = 1, \dots, v$, the sequence of tests $T_W(n)$ is asymptotically equivalent to the sequence $T_{LR}(n)$.*

The proof which is given in the Appendix is based on a preliminary result proved as a lemma to the theorem which states that the variance-covariance matrix of the residuals from regressing Y on X and W can be partitioned as the sum of two idempotent quadratic forms. The first comes from the matrix X alone, the second is a function of both X and W. The core of the proof uses a Taylor series expansion and the matrix operators of linear algebra. The impact of this result on asymptotic equivalence, is that an investigation using Eq. (14) automatically yields asymptotically analogous results to an investigation based on Eq. (15).

4 The forward distribution of the tests

None of the tests introduced in the previous sections is robust to the presence of outliers. In order to provide information about the effect of outliers on multivariate transformations we can use a forward search through the data and monitor the values of the tests. A full description of the forward search for the analysis of multivariate transformations can be found in Riani and Atkinson (2001). It is made up of three steps: 1) choice of the initial subset, 2) progressing in the search and 3) monitoring the search. We find an initial subset of moderate size by robust analysis of the matrix of bivariate scatter plots. The initial subset of r observations consists of those observations which are not outlying on any scatter plot, found as the intersection of all points lying within a robust contour containing a specified portion of the data (Riani and Zani 1997) and inside the univariate boxplot. As concerns 2), in every step of the forward search given a subset of size m ($m = r, \dots, n-1$), we move to a subset of size $(m+1)$ by selecting the $(m+1)$ units with the $(m+1)$ smallest Mahalanobis distances. These distances are calculated using in every step the variables transformed as suggested by the null hypothesis. For multivariate

transformations Riani and Atkinson (2001) monitor a sequence of forward plots of test statistic (5) and of parameter estimates to obtain transformations which describe most of the data, with the outliers entering at the end of the searches. The monitoring of the units entering at each step enables us to evaluate the effect of the introduction of each observation on the results of the test statistics. Up to now the results of the generalized likelihood ratio test (5) have been compared with the asymptotic χ^2_v distribution. The new contribution given by this section is the investigation, through the use of empirical confidence envelopes, not only of the asymptotic distribution of the test statistics introduced in the previous sections, but also of their small sample distribution induced by the application of the forward search.

4.1 Heads data

In order to illustrate the properties of the different tests we start by using a data set of six readings on the dimensions of the heads of 200 twenty year old Swiss soldiers. The data are described by Flury and Riedwyl (1988, p. 218) and also by Flury (1997, p. 6). The variables are:

y_1: minimal frontal breadth
y_2: breadth of angulus mandibulae
y_3: true facial height
y_4: length from glabella to apex nasi
y_5: length from tragion to nasion
y_6: length from tragion to gnathion.

Diagrammatic front and profile views of a head illustrating these measurements are on p. 223 of Flury and Riedwyl (1988).

The data were collected to determine the variability in size and shape of heads of young men in order to help in the design of a new protection mask for the Swiss army. Because of the variations in human heads, it was clear that one mask could not be satisfactory for all soldiers. The aim was to find a few typical head sizes and shapes which, it was hoped, would make it possible to provide satisfactory masks for all soldiers. If the data have a multivariate normal distribution, the standard techniques of multivariate data analysis can be used to determine the best few standard types.

Figure 1 gives the forward plot of the generalized likelihood ratio test for the hypothesis of no transformation. The horizontal lines are the 5%, 10%, 25%, 50%, 75%, 90%, 95% and 99% confidence bands of the theoretical asymptotic χ^2_6 distribution. The dotted lines are the empirical confidence bands based on 1000 simulations constructed using 1000 independent forward searches. In order to generate each set of simulated data we multiplied the $n \times v$ random numbers extracted from $N(0,1)$ by the Choleski decomposition of the covariance matrix of the transformed observations $(y_{C_1}^{\lambda_{01}}, \ldots, y_{C_v}^{\lambda_{0v}})$ and added the mean of the transformed observations obtaining, say, a matrix $(y_{C_1}^{0}, \ldots, y_{C_v}^{0})$. Finally, for each simulated variable we

considered the inverse transformation $(y_{C_j}^0)^{(1/\lambda_{0j})}$, $j = 1, 2, \ldots, v$. It is clear that if $\lambda_{0j} = 0$, the transformation is $\log y_{C_j}$ and the inverse is $\exp(y_{C_j}^0)$. To these simulated data we applied the forward search in order to obtain for each step m the statistic of interest (in this case the generalized ratio test under the null hypothesis $H_0 : \lambda = \lambda_0$). We repeated this procedure for each set of simulated data and finally we sorted the simulations in each step of the forward search to compute the empirical quantiles.

Figure 1 clearly shows that the forward distribution of the empirical curves is increasing and always below the theoretical ones with the difference getting smaller as the search progresses. The forward search orders the data according to their agreement with the suggested model with remote observations entering the subset in the last steps.

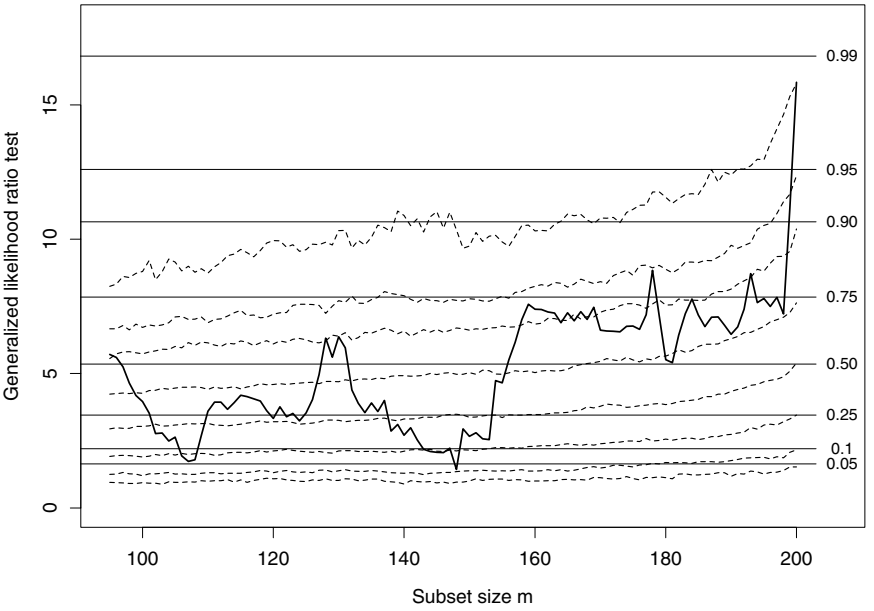

Fig. 1. Heads data: forward plot of the generalized likelihood ratio test for the hypothesis of no transformation. The horizontal lines are the 5%, 10%, 25%, 50%, 75%, 90%, 95% and 99% confidence bands of the χ_6^2 distribution. The dotted lines are the corresponding empirical confidence bands

These remote observations are those most likely to produce evidence for transformation, so we expect the increasing smooth behaviour of the empirical confidence bands. The plot shows that during the central part of the forward search the value of the test is generally around the 75% quantile suggesting that these data do not have to be transformed. In the final two steps of the search ($m = 199$ and $m = 200$), with the introduction of the units 104 and 111, the generalized likelihood ratio statistic shows a sudden jump causing the value to be significant at 0.05 level, when compared with a χ_6^2. The empirical superimposed forward confidence bands show that the effect of the introduction of the last two units is even more

pronounced than it would seem using the theoretical asymptotic bands, because they move the value of the test outside the 99% empirical confidence band. Units 104 and 111 have large values of y_4, more remote from this distribution than any other units for any variable, but they are not outlying in any other marginal distribution. Our conclusions about this data set is that multivariate normality provides a useful model and that there may be two people for whom the measurements of y_4 is incorrect.

Let us now study the forward behaviour of the tests introduced in Sect. 4 based on the use of constructed variables. Figure 2 shows the monitoring of the results of the constructed variable tests.

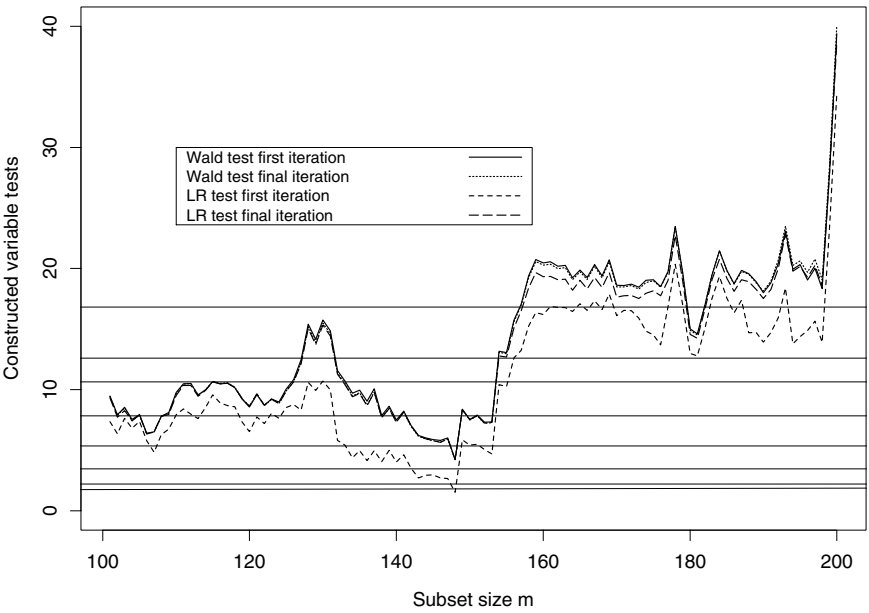

Fig. 2. Heads data: forward plot of the constructed variables tests for the hypothesis of no transformation. The horizontal lines are the 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99% confidence bands of the χ^2_6 distribution

For each of the two tests (Wald and likelihood ratio) we computed both the values obtained using the estimate of $\hat{\Sigma}(\lambda)$ coming from the first iteration and those obtained after convergence. This picture clearly shows that in this example there is close agreement between the results of the likelihood ratio test and of the Wald test. The other point to notice is that the results of the two tests do not seem to change appreciably if we use an estimate of Σ based on the first or final iteration. Another thing which is worthwhile to remark is that the shape of the forward constructed variables tests seems to follow closely that of the generalized likelihood ratio given in Fig. 1 as can be seen by plotting one line against the other. However, the results of the constructed variables tests are always much higher than those in Fig. 1. In Fig. 2 are also given the reference quantiles of a χ^2_6 distribution.

Using such a reference distribution we would claim that in this data set the evidence of transformation is spread throughout the data. This conclusion would be strongly in disagreement with the results of the generalized likelihood ratio test given in Fig. 1.

Figure 3 gives again the monitoring of the likelihood ratio test based on the final iteration and the associated forward confidence bands. This picture clearly shows that simulation envelopes always lie above the associated quantiles of the χ^2 distribution. The gap between the two curves seems to increase in the final part of the search.

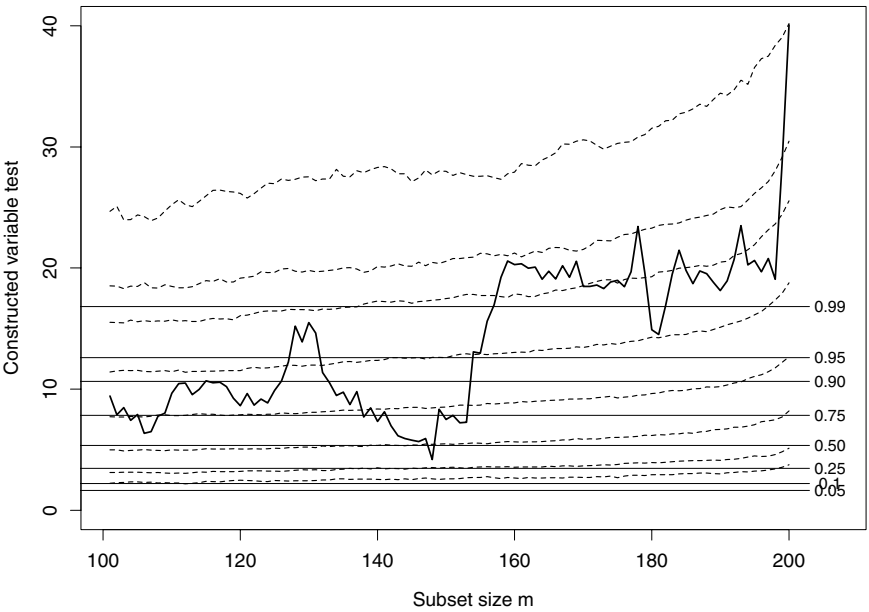

Fig. 3. Heads data: forward plot of the likelihood ratio test based on constructed variables using an estimate of Σ coming from final iteration. The horizontal lines are the 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99% confidence bands of the χ^2_6 distribution. The dotted lines are the corresponding empirical confidence bands

This heavy tails phenomenon has been analysed for the score test for transformations of univariate data by Atkinson and Riani (2002). These authors conclude that the longer tails are due to the presence of the observations in the constructed variables on which the response is regressed. Note that the longer tail effect becomes more pronounced in the final part of the search when the most remote observations are included in the subset. The simulations show that the forward values of the likelihood ratio test are never significant at the 5% level in the central part of the search and that only in the final two steps (introduction of units 104 and 111) they become significant at 1%. Note that these results are perfectly in agreement with those previously obtained from the monitoring of the generalized likelihood ratio test (see Fig. 1). The conclusion which comes from the analysis of this example

is that the multivariate transformation tests based on constructed variables have a forward distribution which is always above the reference χ^2 distribution and this difference increases in the final part of the search. The judgement about the significance of the tests coming from the use of constructed variables must be done only after superimposing forward simulation envelopes.

4.2 Emilia-Romagna data

The data set we consider in this section is made up of 341 observations and 28 variables. The 341 observations are from all the municipalities of Emilia-Romagna, a region of Italy. Nearly all the variables are indices related to different aspects of the quality of life. A full description of these data can be found in Atkinson et al. (2004). It is clear that in this example with 28 variables it is impossible to use the generalized likelihood ratio test. Atkinson et al. (2004) tackled the problem of finding the best transformation parameters to achieve approximate normality by dividing the 28 variables into three categories (demographic, income and wealth and industrial production) and analyzing each set of variables separately.

The contribution of this section is to investigate the set of transformation parameters found by Atkinson et al. (2004) by analyzing each set of variables separately using a constructed variable approach and the tests given in Sect. 3. Given that we expect that the reference distribution is not a χ^2_{28}, we have computed forward simulation confidence envelopes. Figure 4 gives the forward plot of the likelihood ratio which comes from final iteration with the theoretical and empirical confidence bands. This figure shows that the simulated quantiles are well above those of the χ^2_{28} distribution. The value of the test seems generally to lie between the 95% and 99% envelopes with a decrease before the final part of the search. It is interesting to notice the sudden upward jump due to the inclusion of the outliers in the final part of the search. The last 21 units which enter the forward search are all (except two) poor rural communities belonging to the mountainous area of the region. The monitoring of Mahalanobis distances during the forward search shows that these municipalities are quite different from the rest of the data. These units tend to have similar problems with an aging population, and low indexes of wealth, education, housing and industrial development. The conclusion is that in this data set using the transformation parameters we can claim to have reached only approximate normality, because the values of the test are at the boundary of significance. Finally, the forward search reveals a series of outliers which have an enormous effect on the choice of the transformation.

5 Conclusions

In this paper we have introduced different tests based on constructed variables for the analysis of multivariate transformations. The main theoretical contribution of this paper is a theorem which states the asymptotic equivalence of two constructed variable tests.

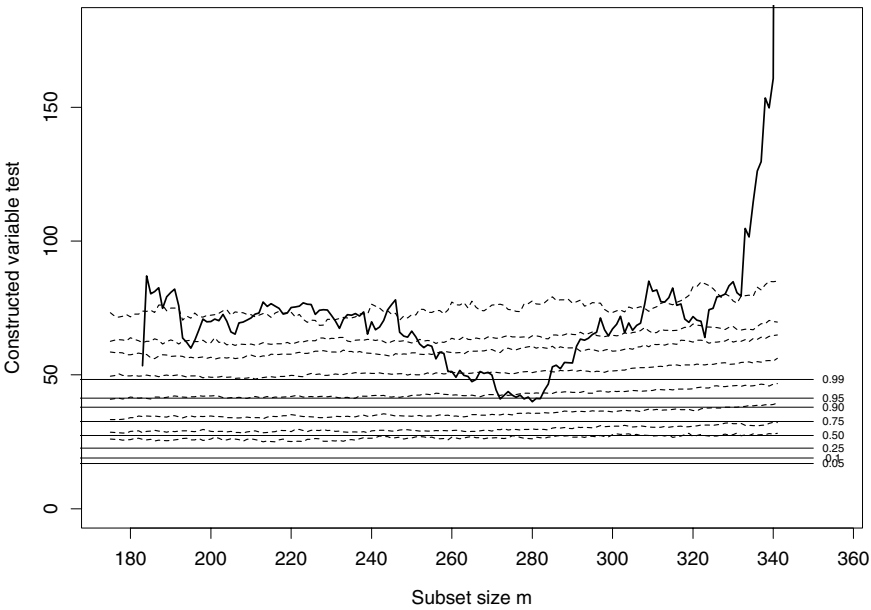

Fig. 4. Emilia Romagna data: monitoring of the likelihood ratio test based on the final iteration. The horizontal lines are the 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99% confidence bands of the χ^2_{28} distribution. The dotted lines are the corresponding empirical confidence bands

Given that both the tests based on the likelihood ratio and those based on the use of constructed variables are not robust to the presence of outliers, in order to provide information about the effect of outliers on multivariate transformations we embedded the tests in a forward search context and monitored their values. Up to now in the literature the results of the tests have been compared with the theoretical asymptotic χ^2 distribution.

The computational contribution given in this paper has been the investigation of the small sample distribution of the tests induced by the forward search. In all the examples which have been considered we saw that the constructed variable tests showed a good agreement in each step of the forward search irrespective of the fact of using an estimate of the covariance matrix of the residuals coming from first or final iteration. As concerns the forward empirical confidence bands of the tests, we showed that the generalized likelihood ratio tests are characterized by a smooth increasing behaviour as the search progresses and asymptotically tend to the nominal values. On the other hand, we showed that the tests based on constructed variables tend to have a forward heavy tailed distribution. This implies that, if the researcher decides to use constructed variable tests, the judgement about the eventual significance of the values which are obtained must be made only using simulation envelopes. Conversely, the use of horizontal asymptotic confidence bands to judge the significance of the forward values of the generalized likelihood ratio test implies that we shall be conservative with respect to the null hypothesis.

Acknowledgements The author wishes to thank the two anonymous referees for their constructive criticism that led to improvements in the exposition and corrected some typos. Furthermore, suggestions and comments from Anthony Atkinson, Andrea Cerioli and Sergio Zani are appreciated.

Appendix: proof of the asymptotic equivalence of the constructed variable tests

In order to show the relationship between the two tests given in Eqs. (14) and (15), we first have to find an expression relating the variance-covariance matrix of the residuals from regressing Y on X and W, to the variance covariance matrix of residuals from regressing only on X, where W is the $n \times v$ matrix which contains the constructed variables.

Lemma. *In the multiple multivariate regression model* $Y = XB + W\Gamma + \Xi$, *where* $\Gamma = diag\,(\gamma_1, \ldots, \gamma_v)$ *is a diagonal matrix, the total sum of squares of the residuals can be partitioned as*

$$(n-p-1)S_w = E^T(w)E(w) = Y^T A_X Y - Y^T F Y \tag{A.1}$$

where $F = A_X W\{W^T A_X W\}^{-1} W^T A_X$, $A_X = I_n - X(X^T X)^{-1} X^T$. $Y^T A_X Y$ *is the matrix of residual sum of squares after regressing only on* X, *and* $Y^T A_X W\{W^T A_X W\}^{-1} W^T A_X Y$ *is the reduction in the matrix of residual sum of squares due to* Γ *after adjusting for* B.

Proof.

$$\begin{aligned} E^T(w)E(w) &= (Y - XB - W\Gamma)^T(Y - XB - W\Gamma) \qquad \text{(A.2)} \\ &= Y^T Y - Y^T XB - Y^T W\Gamma + B^T(X^T XB + X^T W\Gamma - X^T Y) \\ &\quad + \Gamma^T(W^T XB + W^T W\Gamma - W^T Y). \end{aligned}$$

After differentiating with respect to B and Γ and equating to zero, we obtain the following system of equations

$$X^T XB + X^T W\Gamma = X^T Y \tag{A.3}$$

$$W^T XB + W^T W\Gamma = W^T Y. \tag{A.4}$$

Rearranging Eq. (A.3) we obtain

$$\hat{B} = (X^T X)^{-1} X^T Y - (X^T X)^{-1} X^T W\hat{\Gamma}. \tag{A.5}$$

Substitution of this value into (A.4) yields, after rearrangement, to

$$\hat{\Gamma} = (W^T A_X W)^{-1} W^T A_X Y. \tag{A.6}$$

Using the expressions just found for $\hat{B}$ and $\hat{\Gamma}$ and the constraints implied by Eqs. (A.3) and (A.4) we can write

$$\begin{aligned} (n-p-1)S_w = E^T(w)E(w) &= Y^T Y - Y^T X\hat{B} - Y^T W\hat{\Gamma} \\ &= Y^T A_X Y - Y^T A_X W\{W^T A_X W\}^{-1} W^T A_X Y. \end{aligned}$$

This is the multivariate generalization of Eq. (2.29) given in Atkinson and Riani (2000):

$$(n-p-1)s_w^2 = y^T Ay - (y^T Aw)^2/(w^T Aw). \tag{A.7}$$

Remark. the projection matrix $F = A_X W\{W^T A_X W\}^{-1} W^T A_X$ is symmetric and idempotent.

Theorem. *for testing the null hypothesis $H_0 : \gamma_j = 0$, $j = 1, \ldots, v$, the sequences of tests $T_W(n)$ is asymptotically equivalent to the sequence $T_{LR}(n)$.*

We begin the proof by showing that, under the null hypothesis $H_0 : \gamma = 0$, Eq. (14) reduces to a difference in residual sum of squares. If we apply the Choleski decomposition to the matrix $\hat{\Sigma}$ as follows: $\hat{\Sigma} \otimes I_n = E(w)^T E(w)/(n-p-1) \otimes I_n = (C^T C)^{-1} \otimes I_n$, where C is lower triangular, we can rewrite model (10) as

$$z_C^* = X_C^* \beta + W_C^* \gamma_C + \epsilon_C^*, \tag{A.8}$$

where $z_C^* = (C \otimes I_n) z^*, X_C^* = C \otimes X, W_C^* = (C \otimes I_n) W^*$ and $\epsilon_C^* = (C \otimes I_n)\epsilon^* \sim N(0, I_{n \times v})$.

Now, given that

$$R = (0, I_v) \tag{A.9}$$

where v is the number of rows of γ, we obtain that $R\delta = \gamma$. Similarly, the expression $\left[R\left\{(X^*, W^*)^T(\Sigma^{-1} \otimes I)(X^*, W^*)\right\}^{-1} R^T\right]$ simply extracts the last v rows and columns of the matrix

$$\begin{pmatrix} X_C^{*T} X_C^* & X_C^{*T} W_C^* \\ W_C^{*T} X_C^* & W_C^{*T} W_C^* \end{pmatrix}^{-1}. \tag{A.10}$$

Applying the rules of the inverse of a partitioned matrix (see for example Mardia et al. 1979, p. 459) it follows that

$$R\left\{(X^*, W^*)^T(\hat{\Sigma}^{-1} \otimes I)(X^*, W^*)\right\}^{-1} R^T = \left(W_C^{*T} A_{X_C^*} W_C^*\right)^{-1}, \tag{A.11}$$

where $A_{X_C^*} = I - X_C^*(X_C^{*T} X_C^*)^{-1} X_C^{*T}$.

Finally, the null hypothesis $\gamma_j = 0$ implies that:

$$r = (0, \ldots, 0)^T. \tag{A.12}$$

We obtain

$$\begin{aligned} T_W &= (r - R\hat{\delta})^T \left[R\left\{(X^*, W^*)^T(\hat{\Sigma}^{-1} \otimes I)(X^*, W^*)\right\}^{-1} R^T\right]^{-1} (r - R\hat{\delta}) \\ &= (W_C^* \hat{\gamma})^T A_{X_C^*} W_C^* \hat{\gamma} \\ &= (z_C^* - X_C^* \hat{\beta}_C^* - e_C^*)^T A_{X_C^*} (z_C^* - X_C^* \hat{\beta}_C^* - e_C^*), \end{aligned}$$

where $e_C = A_{X^*_C,W^*_C} z^*_C$ is the vector of residuals of the model which contains both β and γ. Now, given that $X^{*T}_C A_{X^*_C} = 0$ and $A_{X^*_C} A_{X^*_C,W^*_C} = A_{X^*_C,W^*_C}$,

$$T_W = z^{*T}_C A_{X^*_C} z^*_C - z^{*T}_C A_{X^*_C,W^*_C} z^*_C. \tag{A.13}$$

The first argument in the final equation is the residual sum of squares in the model with only X^*_C and the second is the residual sum of squares in the full model. Alternative ways of writing Eq. (A.13) are:

$$T_W = \text{trace}\left\{E^T_C E_C - E(w)^T_C E(w)_C\right\} \tag{A.14}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{v}\left\{e^2_{C_{ij}} - e(w)^2_{C_{ij}}\right\}. \tag{A.15}$$

The likelihood ratio test for $H_0 : \gamma = 0$ is defined as

$$T_{LR} = n \ln \frac{|E^T E|}{|E^T(w)E(w)|}. \tag{A.16}$$

Using the results of Lemma 1 we can write the former expression as

$$= n \ln \frac{|Y^T A_{X,W} Y + Y^T A_X W\{W^T A_X W\}^{-1} W^T A_X Y|}{|Y^T A_{X,W} Y|}. \tag{A.17}$$

If we define $G = FY = A_X W\{W^T A_X W\}^{-1} W^T A_X Y$, Eq. (A.17) can be rewritten as:

$$= n \ln \frac{|E(w)^T E(w) + G^T G|}{|E(w)^T E(w)|} \tag{A.18}$$

$$= n \ln |I_v + (E(w)^T E(w))^{-1} G^T G|. \tag{A.19}$$

Using the Taylor series expansion of the log of the determinant the former expression can be written as:

$$T_{LR} = \text{ trace}\{(E^T(w)E(w))^{-1} G^T G\} + o_p(1). \tag{A.20}$$

Now, since $\hat{\Sigma}^{-1} = C^T C = (n-p-1)\{E^T(w)E(w)\}^{-1}$,

$$T_{LR} \approx \text{trace}(C^T C G^T G) = \text{trace}\left\{C^T C (E^T E - E(w)^T E(w))\right\}. \tag{A.21}$$

Now, since for 2 generic matrices A and B, $(A \otimes I)\text{vec}(B) = \text{vec}(BA^T)$ and since taking the trace of $(D^T D)$ corresponds to computing the sums of squares of all the elements of the matrix, i.e. $\sum_{ij} d^2_{ij}$, the former expression can be written as

$$T_{LR} = (\text{vec}E_C)^T(\text{vec}E_C) - (\text{vec}E_C(w))^T(\text{vec}E_C(w)) + o_p(1) \tag{A.22}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{v}\left[e^2_{C_{ij}} - e(w)^2_{C_{ij}}\right] + o_p(1), \tag{A.23}$$

where $\text{vec}E_C = (C \otimes I)\text{vec}E$ and $\text{vec}E_C(w) = (C \otimes I)\text{vec}E(w)$.

The last expression apart from the term $o_p(1)$ coincides with (A.15).

Corollary. in the univariate case the likelihood ratio test defined as

$$T_{LR} = n \ln \frac{y^T A_X y}{y^T A_{X,w} y}$$

is asymptotically equivalent to the Wald test.

$$T_W = \frac{y^T A_X y - y^T A_{X,w} y}{y^T A_{X,w} y/(n-p-1)}$$

Proof. Using Eq. (A.7), the Wald test can be rewritten as

$$\begin{aligned} T_W &= \frac{y^T A_X y - y^T A_{X,w} y}{y^T A_{X,w} y/(n-p-1)} \\ &= \frac{(y^T A_X w)^2 / w^T A_X w}{y^T A_{X,w} y/(n-p-1)}. \end{aligned} \tag{A.24}$$

The likelihood ratio test becomes

$$\begin{aligned} T_{LR} &= n \ln \frac{y^T A_X y}{y^T A_{X,w} y} \\ &= n \ln \frac{y^T A_X y}{y^T A_X y - (y^T A_X w)^2/(w^T A_X w)} \\ &= n \ln \left\{ 1 + \frac{(y^T A_X w)^2}{w^T A_X w \; y^T A_X y - (y^T A_X w)^2} \right\} \\ &= n \frac{(y^T A_X w)^2 / w^T A_X w}{y^T A_X y - (y^T A_X w)^2/(w^T A_X w)} + o_p(1) \\ &= \frac{(y^T A_X w)^2/(w^T A_X w)}{y^T A_{X,w} y/n} + o_p(1). \end{aligned} \tag{A.25}$$

When the sample size n tends to infinity, Eq. (A.25) becomes equivalent to Eq. (A.24) and this completes the proof.

References

Andrews DF, Gnanadesikan R, Warner JL (1971) Transformations of multivariate data. Biometrics 27: 825–840

Atkinson AC (1985) Plots, Transformations, and Regression. Oxford University Press, Oxford

Atkinson AC (1995) Multivariate transformations, regression diagnostics and seemingly unrelated regression. In: Kitsos CP, Müller WG (eds) MODA 4 – Advances in Model-Oriented Data Analysis, pp 181–192. Physica-Verlag, Heidelberg

Atkinson AC, Riani M (2000) Robust Diagnostic Regression Analysis. Springer, Berlin Heidelberg New York

Atkinson AC, Riani M (2002) Tests in the fan plot for robust, diagnostic transformations in regression. Chemometrics and Intelligent Laboratory Systems 60: 87–100

Atkinson AC, Riani M, Cerioli A (2004) Exploring Multivariate Data with the Forward Search. Springer, Berlin Heidelberg New York

Box GEP, Cox DR (1964) An analysis of transformations (with discussion). Journal of the Royal Statistical Society, Series B 26: 211–246

Cox DR, Hinkley DV (1974) Theoretical Statistics. Chapman and Hall, London

Flury B (1997) A First Course in Multivariate Statistics. Springer, Berlin Heidelberg New York

Flury B, Riedwyl H (1988) Multivariate Statistics: A Practical Approach. Chapman and Hall, London

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

Hamilton JD (1994) Time Series Analysis. New Jersey: Princeton University Press, Princeton

Le Cam L (1986) Asymptotic Methods in Statistical Decision Theory. Springer, Berlin Heidelberg New York

Le Cam L, Yang GL (1990) Asymptotics in Statistics: Some Basic Concepts. Springer, Berlin Heidelberg New York

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. Academic Press, London

Riani M, Atkinson AC (2000) Robust diagnostic data analysis: Transformations in regression (with discussion). Technometrics 42: 384–398

Riani M, Atkinson AC (2001) A unified approach to outliers, influence, and transformations in discriminant analysis. Journal of Computational and Graphical Statistics 10: 513–544

Riani M, Zani S (1997) An iterative method for the detection of multivariate outliers. Metron 55: 101–117

Velilla S (1993) A note on the multivariate Box-Cox transformation to normality. Statistics and Probability Letters 17: 259–263

Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. Journal of the American Statistical Association *57*: 348–368