

Robust and Efficient Dimension Reduction

Metodi robusti ed efficienti per la riduzione delle dimensioni

Marco Riani

Dipartimento di Economia
Università di Parma
mriani@unipr.it

Matilde Bini¹

Dipartimento di Statistica G. Parenti
Università di Firenze
bini@ds.unifi.it

Riassunto: L'obiettivo di questo lavoro è quello di estendere l'algoritmo di *forward search* (Atkinson and Riani (2000)) alle componenti principali. L'approccio suggerito parte da un *data set* ridotto privo di outliers ed include sequenzialmente le rimanenti osservazioni in base ad una misura via via crescente di anomalia delle stesse. La procedura proposta consente di ottenere un approccio unificato all'analisi delle trasformazioni multivariate, all'individuazione degli outliers e al trattamento delle osservazioni influenti nelle componenti principali. Per illustrare il nuovo metodo si utilizza un *data set* che si riferisce alla valutazione della qualità delle Università italiane.

Keywords: Principal components, robust methods, multivariate transformations, forward search, masked outliers, robustness.

1. Introduction

Principal component analysis (PCA) essentially uses eigenvalues and eigenvectors of covariance or correlation matrices, whose statistical properties are difficult to study. Most results are obtained under normal assumptions (e.g. Mardia *et al.* (1979); p. 229). This implies that at present few methods are available to assess the stability of the results of PCA. Often the assumption of multivariate normality is approximately true only after that the data have been appropriately transformed. Unfortunately, in multivariate data is very difficult to test and validate a particular transformation due to the well known masking and swamping problems (Velilla (1995)). The lack of a proper choice of the most appropriate transformation may lead to overestimate (underestimate) the importance of particular variables (Riani and Atkinson (2001)), and to a wrong interpretation of the extracted components. The difficulties and intricacies of the choice of the best multivariate transformation usually lead the analyst to apply PCA to untransformed data. Finally, as concerns outlier detection in PCA, the traditional approach is based on robust estimates of the covariance and correlation matrix or on projection pursuit (Hubert *et al.* (2002)). These estimators, however, are seldom used in practice because they are difficult to compute and unacceptably inefficient.

In this paper we suggest a unified robust and efficient approach to outlier detection, multivariate transformations, robust estimation of covariance matrices and dimension reduction. The suggested approach is an extension to PCA of the forward search (FS) methodology as described by Atkinson and Riani (2000) for regression data.

The paper is structured as follows: in section 2 we introduce notation and recall the likelihood ratio test for multivariate transformations. In section 3 we describe the general

¹Acknowledgments: the dataset used in this work has kindly been made available by the National University Evaluation Committee-Ministry of Education, Universities and Research (MIUR).

principles of our robust diagnostic method. In section 4 we show how the FS, free from masking and swamping problems, can be used to find and test a set of transformation parameters, to reduce dimensionality and to monitor the stability of the extracted principal components. In order to show the power of our procedure, in section 5 we introduce and use a data set coming from the National University Evaluation Committee (NUEC).

2. Tests for transformation in multivariate analysis

Let Y be a sample data matrix of dimension $n \times p$ and let y_{ij} be the i th observation on response j . In the extension of the Box and Cox (1964) family to multivariate responses the normalized transformation of y_{ij} is

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

where $\dot{y}_j$ is the geometric mean of the j th response. The value $\lambda_j = 1$ ($j = 1, \dots, p$) corresponds to no transformation of any of the responses.

The likelihood ratio test to validate the hypothesis $\lambda = \lambda_0$, can be expressed as the ratio of the determinants of two covariance matrices (Atkinson (2002)):

$$T_{LR} = n \log \{ |\hat{\Sigma}(\lambda_0)| / |\hat{\Sigma}(\hat{\lambda})| \} \quad (2)$$

where $\hat{\lambda}$ is the vector of p parameter estimates found by numerical search maximizing the transformed normal log likelihood. The value of T_{LR} must be compared with the χ^2 distribution on p degrees of freedom. In order to find the most appropriate transformation we use both the maximum likelihood estimates (MLE) and the values of the test statistic. However, as we see in the next sections, in a confirmatory stage it is better to look at the value of the test rather than parameter estimates, because if the likelihood is flat, the estimates can vary widely without conveying any useful information about the transformation.

Remark: From equation (2) it is clear that if we want to use standard deletion methods to validate a particular set of transformations, we have to delete each observation in turn and every time remaximize the likelihood. Because of the necessity for remaximization, this approach becomes infeasible if multiple deletions are required. In addition, if there is a group of influential observations, the backwards approach may fail due to masking.

3. The Forward Search for PCA

Principal components analysis looks for a few linear combinations which can be used to summarize the data, losing in the process as little information as possible. Let $\tilde{Y}$ be the sample centered data matrix $\tilde{Y} = Y - 1\bar{y}'$, and v a standardized vector ($v'v = 1$). The linear combination of the columns of $\tilde{Y}$, $z = \tilde{Y}v$ which has the largest variance is obtained when v is the standardized eigenvector corresponding to the largest eigenvalue of $\hat{\Sigma} = V\Gamma V'$: V is the matrix which contains the eigenvectors ($V = (v_1, \dots, v_p)$) and Γ is the diagonal matrix containing the eigenvalues $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_p)$ in non decreasing order. The j -th principal component z_j is defined as $z_j = \tilde{Y}v_j$, $j = 1, \dots, p$. As is well known, the sum of the first k eigenvalues divided by the sum of all the eigenvalues

of the matrix $\hat{\Sigma}$ represents the proportion of variation explained by the first k principal components.

The small sample distribution of the eigenvalues and of the eigenvectors is very complicated even if all parent correlations vanish. The reason comes from the fact that the eigenvalues are non rational functions of the elements of $\hat{\Sigma}$. The large sample results stem from the theorem (Anderson (1963)) which states that for *normal data*, when the eigenvalues of Σ are distinct, the sample principal components and the eigenvalues are the maximum likelihood estimators of the corresponding population parameters. This implies that is very important to try to transform the data to reach approximate normality and to remove outliers. If the values of the parameters of the model were known ($\mu(\lambda)$ and $\Sigma(\lambda)$), there would be no difficulty in detecting the outliers, which would have large Mahalanobis distances. The difficulty arises because the outliers are included in the data used to estimate the parameters, which can then be badly biased.

Most methods for outlier detection therefore seek to divide the data into two parts, a larger “clean” part and the outliers. The clean data are then used for parameter estimation. The simplest example of this division of the data into two parts is in the use of single deletion diagnostics, where the division is into one potential outlier and the rest of the data. It is clear, however, that if multiple deletions are required there is a combinatorial explosion of the number of cases that have to be considered by such backwards working. Finally, the situation is complicated by the fact that outliers in one transformed scale may not be outliers in another scale.

In the FS such larger subsamples of outlier free observations are found by starting from small subsets and incrementing them with observations which have small Mahalanobis distances, and so are unlikely to be outliers. More precisely, the FS is made up of three steps: choice of the initial subset, progressing in the search and monitoring.

Step 1: Choice of the Initial Subset

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 and inside the univariate boxplot. There are two versions of the robust bivariate contour. The first uses convex hull peeling and B -spline smoothing (Zani *et al.* (1998)). The second, more simple but less robust, is based on robust ellipses (Riani and Zani (1997)). An important property of this method is that the size of the subset can easily be increased or decreased by changing the level of the contour. In the examples which follow we always use the quicker version, because we simply need to find an initial subset of a certain size which is not contaminated by outliers. Finally, since it is the extreme observations which provide the evidence for transformations, such a robust subset will provide a good start to the search for many values of λ .

Step 2: Adding Observations during the Forward Search

In every step of the forward search, given a subset $S_*^{(m)}$ 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.

More specifically, let $\hat{\mu}_{i,S_*^{(n)}} = \hat{\mu}_i$ be the estimated response using all the observations. The squared Mahalanobis distance for observation i is

$$d_i^2 = (y_i - \hat{\mu}_i)^T \hat{\Sigma}^{-1} (y_i - \hat{\mu}_i) = e_i^T \hat{\Sigma}^{-1} e_i, \quad (3)$$

where $\hat{\mu}_i$ is simply the arithmetic mean of $y_1, \dots, y_n$ or may come from regression and $\hat{\Sigma} = \hat{\Sigma}_{S_*^{(n)}}$ is the sample covariance matrix with

$$\hat{\Sigma}_{jk} = \sum_i (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik}) / (n - q) = \sum_i e_{ij} e_{ik} / (n - q), \quad (4)$$

and q is the dimension of the vector of regression parameters.

Now suppose that a subset $S_*^{(m)}$ of m observations is used to estimate the regression and covariances. Let the estimates be $\hat{\mu}_{i,S_*^{(m)}}$ and $\hat{\Sigma}_{S_*^{(m)}}$, yielding the set of squared Mahalanobis distances

$$d_{i,S_*^{(m)}}^2 = (y_i - \hat{\mu}_{i,S_*^{(m)}})^T \hat{\Sigma}_{S_*^{(m)}}^{-1} (y_i - \hat{\mu}_{i,S_*^{(m)}}) \quad i = 1, \dots, n \quad (5)$$

At step m we move to step $m + 1$ by selecting the $m + 1$ units with the smallest $m + 1$ $d_{i,S_*^{(m)}}^2$. Usually just one new unit joins the subset. It may also happen that two or more units join $S_*^{(m)}$ as one or more leave. However, our experience is that such an event is quite unusual, only occurring when the search includes one unit which belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other units then have to leave the subset.

Step 3 Monitoring the search

In multivariate analysis we first try to find a set of transformation parameters to reach approximate normality. In each step of the search (as m goes from r to n (where $p \ll r < n$)) we initially monitor the evolution of MLE and of the likelihood ratio test for transformations and of the Mahalanobis distances using the procedure described in section 4. The changes which occur, will be associated with the introduction of particular observations into the subset m used for fitting. Once a reasonable set of values of transformation parameters λ is found, if our purpose is dimension reduction (PCA), in order to find out if the results are stable, in each step of the search we monitor the elements of the eigenvalues:

$$\Gamma_{S_*^{(r)}}, \dots, \Gamma_{S_*^{(n)}} \quad (6)$$

and the eigenvectors

$$V_{S_*^{(r)}}, \dots, V_{S_*^{(n)}} \quad (7)$$

of the transformed variables. In general, the forward search estimator $\hat{\theta}_{FS}$ is defined as the collection of maximum likelihood estimators in each step of the forward search, that is:

$$\hat{\theta}_{FS} = (\hat{\theta}_r, \dots, \hat{\theta}_n). \quad (8)$$

We use a robust method for the estimation of $\hat{\theta}_r$, but maximum likelihood (that is fully efficient) estimators during the remainder of the search. The zero breakdown point of

maximum likelihood estimators, in the context of the forward search, has advantages. The introduction of atypical (influential) observations is signalled by sharp changes in the curves which monitor the percentage of variance explained by the first principal components, the elements of the eigenvectors or other statistics at every step. In this context, the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of the units into a subset which, in the first steps, is outlier free.

The search which we use avoids, in the first steps, the inclusion of outliers and provides a natural ordering of the data according to the specified null model. It is therefore possible to know how many observations are compatible with a particular specification. Furthermore, the suggested approach enables us to analyze the inferential effect of the atypical units (outliers) on the results of statistical analyses (Cerioli and Riani (1999)).

4. Finding a transformation and reducing the dimension with the forward search

With just one variable for transformation it is comparatively easy to use our forward search to find satisfactory transformations, if such exist, and the observations that are influential in their choice. Atkinson and Riani (2000) perform a forward search using five standard values of λ ($-1, -0.5, 0, 0.5$ and 1) and monitor the likelihood ratio statistic for transformations for each search. However, with p variables for transformation there would be 5^p combinations of the standard values. Whether or not the calculations are time consuming, trying to absorb and sort the information would be difficult (Riani and Atkinson (2001)). We therefore suggest three steps to help structure the search for a transformation:

1. Run a forward search through the data, ordering the observations at each m by Mahalanobis distances calculated from untransformed observations. Estimate λ at each value of m . Use the results to select a set of transformation parameters.
2. Rerun the forward search using distances calculated with the parameters selected in the first step, again estimating λ for each m . If some change is suggested in λ , repeat this step until a reasonable set of transformations has been found. Let this be λ_R .
3. Test the suggested transformation. We expand each transformation parameter in turn around the five common values of λ ($-1, -0.5, 0, 0.5, 1$), using the values of the vector λ_R for transforming the other variables. In this way we turn a multivariate problem into a series of univariate ones. In each search we can test the transformation by comparing the likelihood ratio test with χ^2 on 1 degree of freedom. But we prefer to use the signed square root of the likelihood ratio in order to learn whether lower or higher values of λ are indicated.

If the final purpose of the analysis is dimension reduction, once a reasonable set of transformation parameters has been found and tested, we rerun the FS using the transformed variables and monitor the Mahalanobis distances, the percentage of variance explained by the different components and the stability of the eigenvectors. In the next sections we exemplify this procedure for a real data set concerning the evaluation activity of the Italian Universities.

5. The evaluation of university activity through the forward search

The university service can be modelled as a production process where inputs such as capital, labour and organisational factors, are used to produce teaching and research outputs and outcomes. Many stakeholders such as university policy-makers (i.e. Ministry of the Education - MIUR), teachers, researchers, technical and administrative staff and students, interact among them inside the process and are interested to measure the performance of the process. In the literature, in order to measure the productivity of the research and teaching, and to carry out comparative cost-efficiency analyses, it has been suggested to collect 29 indicators which can be grouped in the following four classes (Ewell (1999); Report n.11/98 in <http://www.mur.st.it/osservatorio/public.htm>): **Resources indicators**, i.e. indicators of resources as available funds, staff, etc.; **Contextual indicators**, i.e. indicators of the context where the university is working, that indicate the degree of the importance of the university, size, the socio-economic environment, etc.; **Process indicators**, that should inform about the organization, facilities and results of the teaching and research processes; **Outcome indicators**, that should inform about the final results and the degree of quality of the teaching and research activities (Biggeri and Bini (2000); <http://www.mur.st.it/valutazionecomitato/attivnuc.htm>).

In Italy starting from March 2000, the National University Evaluation Committee (NUEC) yearly conducts census surveys on all the 73 Italian universities to monitor the situation of the university system and to organize the statistical information system in order to carry out the requested evaluations. The information collected concerns the following aspects:

- the student and academic staff characteristics;
- the teaching service and the studying conditions of students (quality of teaching, classroom availability, laboratories, libraries, location of the offered services);
- the facilities (funds included) devoted to teaching and research activities;
- the managing and administrative organization as well as the working conditions of the technical and administrative staff.

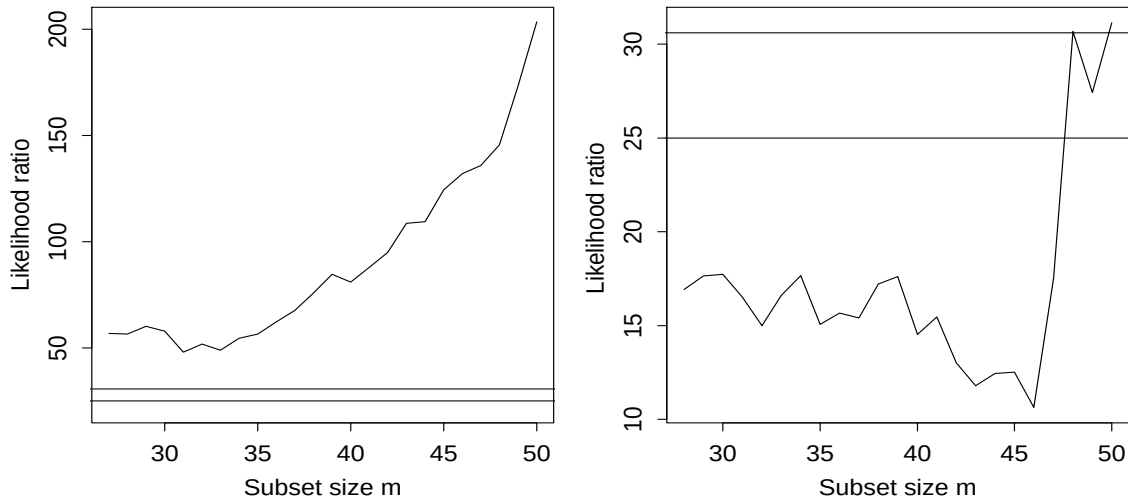
In our study we constructed 15 indicators ($p = 15$) for 50 public Universities ($n = 50$)². More precisely the variables which have been considered are the following:

- **Outcome:** graduation rate within institutional time (y_1); retention rate = 1- drop out rate (y_2).
- **Resources:** faculty/students (y_3); university current expenditure per student (y_4); research funds (y_5); administrative staff per faculty (y_6).
- **Contextual:** university size (y_7); enrollment rate of the best freshmen (y_8); enrollment rate of freshmen from classic and scientific high schools (y_9); extra-regional attraction index (y_{10}).
- **Process:** non institutional time students ratio (y_{11}); private research grants per single member of faculty (y_{12}); financial self-sufficiency (y_{13}); course completing ratio (y_{14}); expenditure for technical staff per ordinary funds (y_{15}).

²We have only included public Universities since the private Universities have a different organization of their structures and are evaluated using different criteria of judgement from the public institutions for receiving funds and incentives from MIUR.

The scatter plot matrix of these data (not given here for lack of space) shows that the distribution of some variables is highly skewed and that maybe some outliers are present in the data. The MLE of the transformation parameters using all the observations are $\hat{\lambda} = (0.19, 0.02, 0.65, 1.22, 0.14, -1.24, -0.16, -0.01, 0.78, 0.49, 1.70, 0.46, 0.37, 1.71, 2.01)$. The null hypothesis of no transformation, that is all $\lambda_j = 1$, yields a value of $T_{LR} = 203.37$. Given that this value must be compared with a χ^2_{15} , it is clear that the p -value is virtually zero. The question is the following: is this large value of T_{LR} due to outliers or is the need of transformation spread throughout the data? In addition: what is the effect of each unit on the MLE of the transformation parameter for each variable? The left hand panel of Figure 1, which reports a FS for the hypothesis of no transformation enables us to state that the need for transformation is spread throughout the data, because the hypothesis of no transformation is always rejected in each step of the forward search. The monitoring of the maximum likelihood estimates of the transformation parameters as

Figure 1: Forward search for the Likelihood ratio test. $H_0 : \lambda = 1$ for all variables. The evidence of transformation is spread throughout the data (left). $H_0 : \lambda = (0.5, 0, 0.5, 0.5, 0.5, -1, 0, 0, 0.5, 0.5, 1, 0.5, 0.5, 1, 1)$. The transformation is acceptable up to step $m = 47$ (right). The two horizontal lines are respectively the 5% and 1% points of the associated χ^2 distribution



described in step 1 of the previous section (not given here for lack of space), suggests that a set of reasonable values for λ is $\lambda_R = (0.5, 0, 0.5, 0.5, 0.5, -1, 0, 0, 0.5, 0.5, 1, 0.5, 0.5, 1, 1)$. The FS using these new values is shown in the 3 panels of Figure 2. This plot enables us to state what are the variables whose MLE are stable, what is the effect of the introduction of each unit on the MLE of the transformation parameters for each variable, and what is the impact of the outliers on the suggested transformation. For example, as concerns the MLE of λ for variable 4, up to $m = 46$ the curve is stable and always lies in the interval 0-0.5. At step $m = 47$ the curve shows a sudden upward jump and the final MLE becomes greater than 1. The same thing, even if less marked, happens for variable 3 (y_3). Finally, the curves for y_{11} , y_{14} and y_{15} , given in the right hand panel of Figure 2, vary widely and show that perhaps no transformation is necessary for these variables. The interpretation of the large changes in the values of the estimated parameters in Figure 2 is aided by considering the likelihood surface. Figure 3 shows how the loglikelihood changes when $m = 46$ and the other 14 values of λ are at their overall value of $\hat{\lambda}$. For example, the

Figure 2: Maximum likelihood estimates of the transformation parameters. The number by the curves refer to the variables

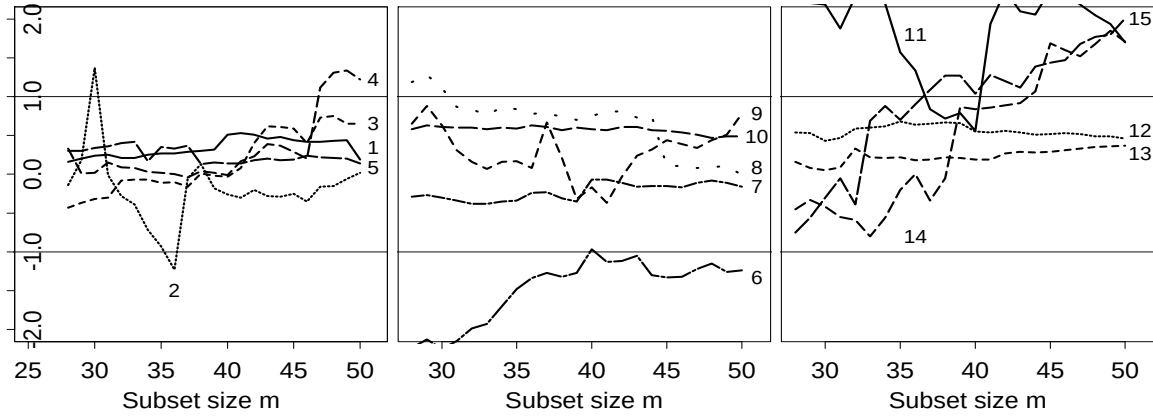
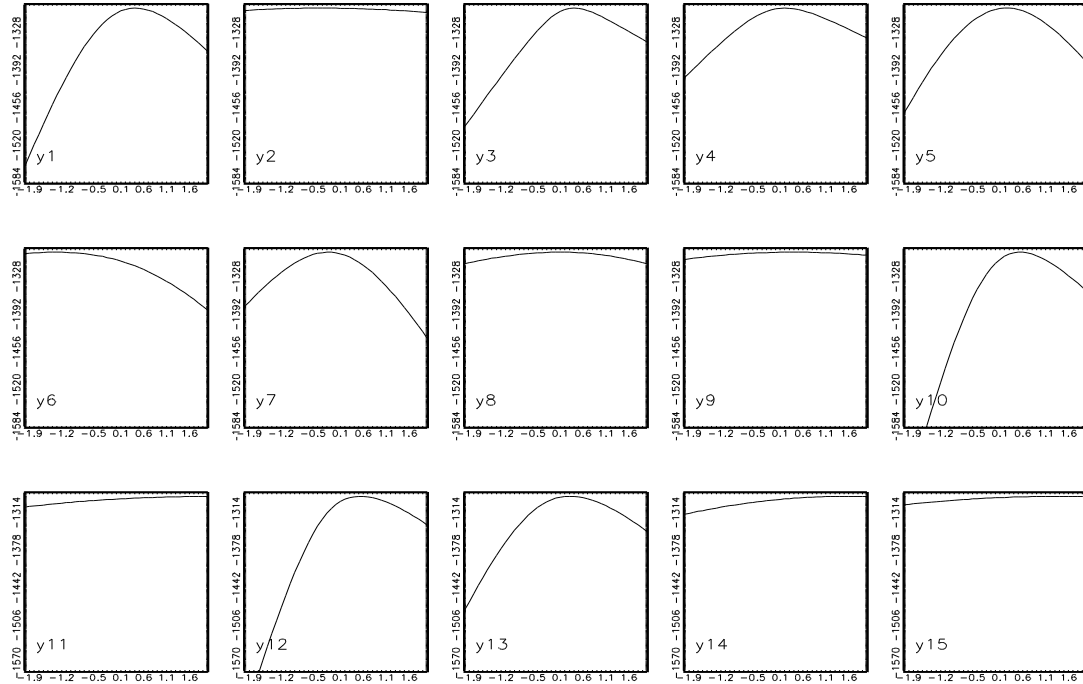


Figure 3: Likelihoods as function of the 15 components of λ when $m = 46$ and $H_0 : \lambda = \lambda_R$. Each panel refers to a variable



surfaces for y_{11} , y_{14} and y_{15} are relatively flat so that a small change in the shape by the addition of one unit can cause a large change in the position of the maximum and so of the value of $\hat{\lambda}$. In Figure 3 we have chosen to show step $m = 46$, because the monitoring of maximum and minimum Mahalanobis distances for the units not in the subset (see Atkinson and Riani (2000) or Atkinson (2002) for a detailed description of these plots) clearly shows that the last 4 units which enter the search (12, 23, 30 and 37), must be considered as atypical. The effect of these 4 units is also clear in the right hand panel of Figure 1 which monitors the likelihood ratio test for the final suggested transformation. Notice that working backwards it would have been impossible to discover the suggested combination of values of λ . In fact, using standard deletion diagnostics, the deletion of 1

or 2 observations does not enable us to see at all the tremendous effect these 4 units have on the likelihood ratio test.

Figure 4: Transformed data using λ_R . Forward plots of (left) maximum Mahalanobis distance (MD) inside subset and (right) minimum Mahalanobis distance not in the subset used to compute the centroid and the sample covariance matrix

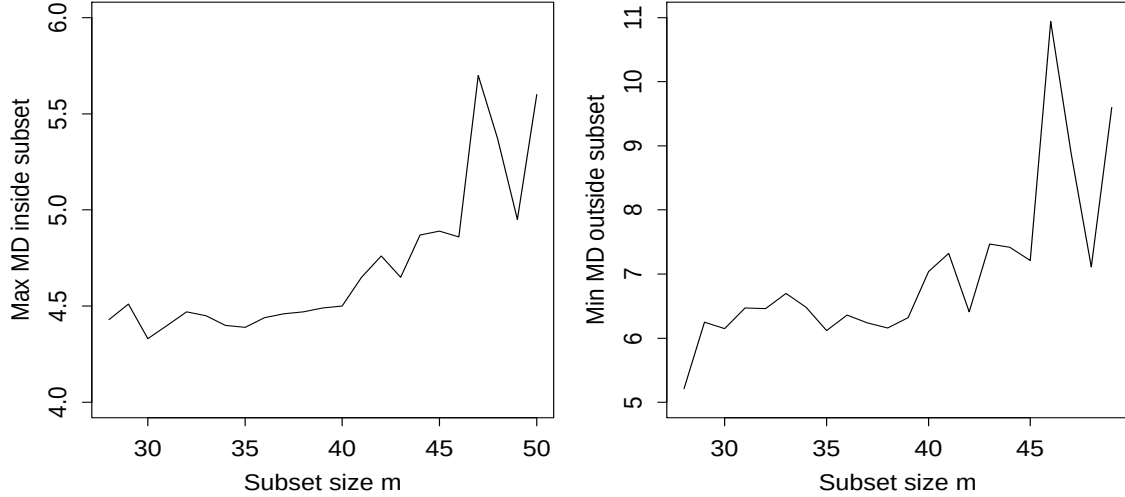
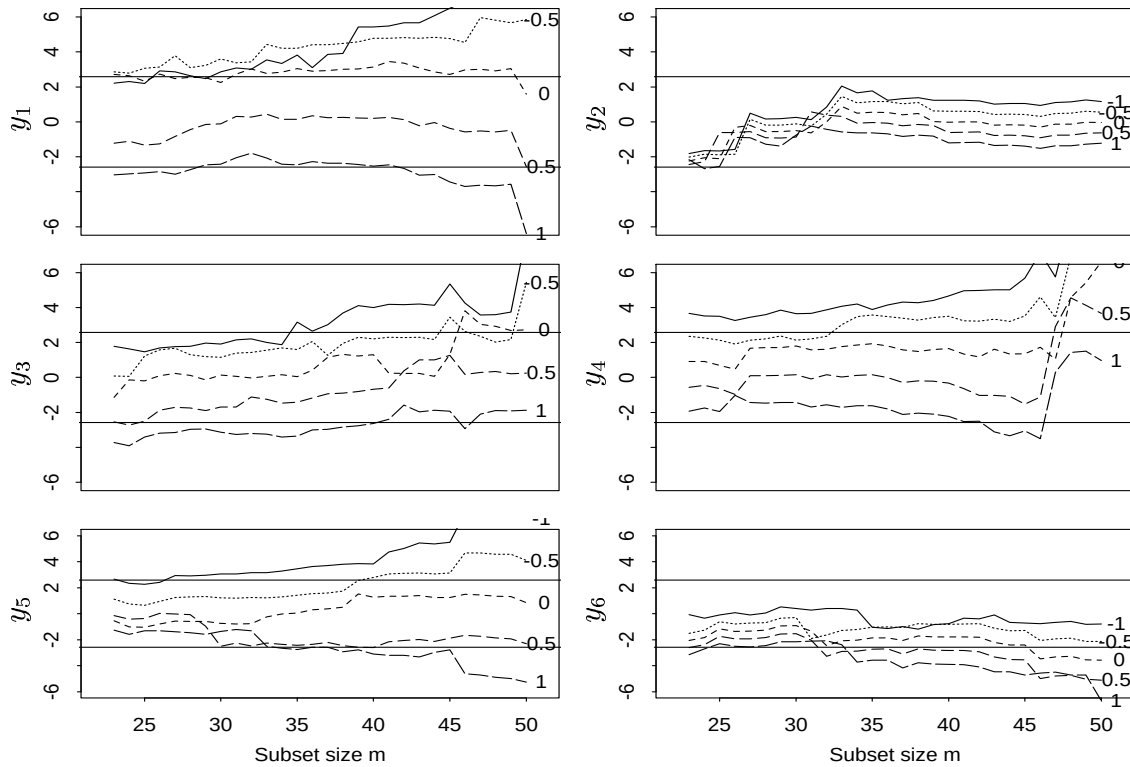


Figure 5: Signed square root of the likelihood ratio test for transformation, confirming $\lambda_R = (0.5, 0, 0.5, 0.5, 0.5, -1, 0, 0, 0.5, 0.5, 1, 0.5, 0.5, 1, 1)$. The numbers by the curves are the values of λ . Due to lack of space we only show the results for the first 6 variables

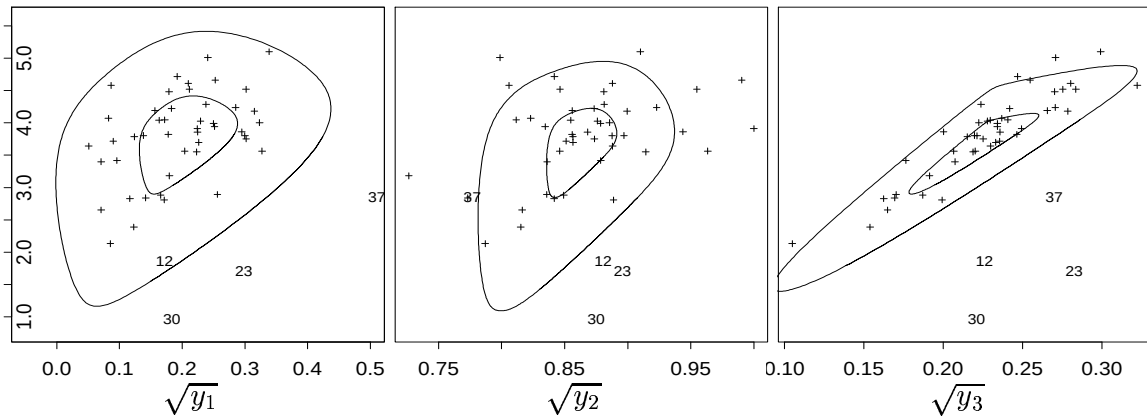


After finding a reasonable set of transformation parameters, we can use a series of forward searches to determine which observations are influential in determining the differ-

ences between the parameter estimates. For each search with each of the 5 most relevant values of the vector $\lambda_k = (-1, -0.5, 0.0.5, 1)$, we use the other 14 values from vector λ_R . We have thus replaced the 5^{15} factorial search by a one variable at a time search with five factors (the λ 's), each at five levels. We test the transformation by use of the likelihood ratio for each value of λ_k , which is on one degree of freedom. Calculation of each value of the statistic in the forward search requires a numerical maximization. The fan plots of the 5 forward searches for the first 6 variables are given in Figure 5 (those for the remaining variables are not given for lack of space). Within each panel we give a plot of the signed square root of the likelihood ratio statistic for the usual five values of λ . Use of the signed square root gives plots which cogently illustrate whether lower or higher values of λ are preferred. For example, the top left panel of Figure 5 shows that the value of 0.5 is the only one which is acceptable for y_1 . Notice at the end the effect of the outliers which make $\lambda_1 = 0$ acceptable and $\lambda_1 = 0.5$ unacceptable. As concerns y_2 , as the flat surface of the likelihood had already suggested (Figure 3), even if the log is the best transformation we cannot reject any value of λ in the interval $(-1, 1)$. Notice that in our suggested approach we not only suggest a transformation which does not suffer from masking and swamping problems, but we are also able to measure the impact of each unit and to illustrate what are the other values of λ which are acceptable for each variable. When multiple values of λ are acceptable we can both use a priori information and choose the value of λ which is in accordance with the other variables belonging to the same group. For example, given that $\lambda = 0.5$ is the only reasonable value of λ for y_1 , we choose this value also for y_2 in order to have the same transformation parameter for the two variables referred to *Outcome*.

Finally, our method provides a link between the evidence for a transformation and the scatter plots of the data. For example, Figure 6 shows the scatter diagrams of $\sqrt{y_4}$ versus $\sqrt{y_1}$, $\sqrt{y_2}$ and $\sqrt{y_3}$ with superimposed robust bivariate boxplots (Zani *et al.* (1998)). This

Figure 6: Scatter plots $\sqrt{y_4} = \sqrt{\text{university current expenditure per student}}$ versus $\sqrt{y_1} = \sqrt{\text{graduation rate within institutional time}}$, $\sqrt{y_2} = \sqrt{\text{retention rate}}$ and $\sqrt{y_3} = \sqrt{\text{faculty/students}}$ with superimposed robust bivariate boxplots. The numbers refer to the 4 outliers



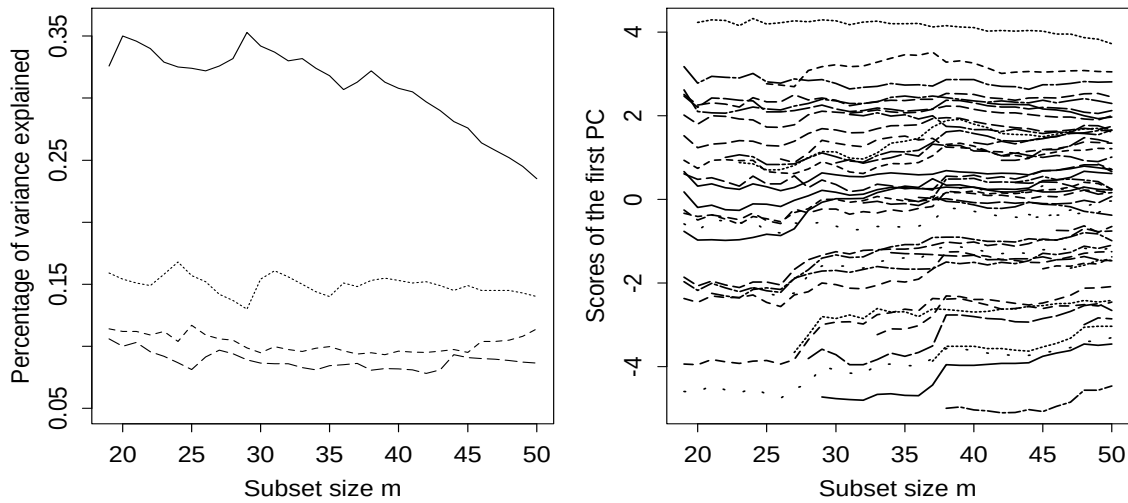
plot clearly shows that the 4 outliers (12 Catania, 23 Modena, 30 Palermo and 37 Roma Tor Vergata) are the Universities which have a combination of values of the first 4 indicators which do not follow the general shape. For example, they have current expenditures per student (y_4) which are much smaller than their corresponding ratio faculty/students

Table 1: Correlation matrix among the first 4 variables. Upper part: original untransformed variables including outliers. Lower part: transformed variables using $\lambda = (0.5, 0.5, 0.5, 0.5)$ and deleting the outliers

	y_1	y_2	y_3	y_4
y_1	1	0.09	0.35	0.11
y_2	0.39	1	0.26	0.27
y_3	0.36	0.35	1	0.65
y_4	0.43	0.36	0.91	1

(y_3). The sources of these observations should then be checked for anomalies and transcription errors. The 3 superimposed boxplots show that the spread of the data in the top right corner of the plots is slightly larger than that in the lower left part. The square root transformation has reduced the larger variability in correspondence of the high values. Finally, in order to show the effects of the outliers on the correlations, in Table 1 we show the 4×4 block of the correlation matrix referred to the first 4 variables before and after the transformation. The upper part shows the original correlations considering all the 50 units and without transforming the variables. The lower part shows the same correlations computed on the transformed variables after deleting the outliers. All the correlations increase and the average increase is 0.18. This implies that the variance of the first principal components will be much larger than that which uses untransformed variables.

Once a reasonable set of transformations has been found, if our purpose is dimension reduction, we can monitor the eigenvalues and the eigenvectors of the covariance and correlation matrices to check the stability of the extracted principal components. Figure 7(left) gives the 4 curves referred to the percentage of variance explained by the first 4 principal components. The trajectories are pretty stable, even if there is a slight decreasing pattern of the curve referred to the first eigenvalue in the last part of the search. However, this does not affect the ordering of the scores of the first principal component as Figure 7(right) clearly shows. This reflects the strong stability of the elements of the first



creasing pattern of the curve referred to the first eigenvalue in the last part of the search. However, this does not affect the ordering of the scores of the first principal component as Figure 7(right) clearly shows. This reflects the strong stability of the elements of the first

eigenvector. As concerns the interpretation of the 4 principal components, the monitoring of the eigenvectors (not given here for lack of space) suggests that the first principal component is positively correlated with all our transformed variables. This means that it can be interpreted as a global performance indicator. As concerns the second principal component, the Universities with high scores in this dimension are those which have high rates of graduation, retention and completion (y_1 , y_2 and y_{14}), high faculty per administrative staff ($1/y_6$) and extra regional attraction index (y_{10}), but low dimensions (y_7), research funds and grants (y_5 and y_{12}), enrollment rates of the freshmen (y_8 and y_9), and expenditures for technical staff (y_{15}). High values for the third principal component still identify universities with high values of y_1 , y_2 and y_{14} , but differently from before, high ratios of contextual indicators ($y_6 - y_9$) except y_{10} , and completion rate (y_{14}), and low values of resources indicators, grants and self sufficiency ($y_3 - y_6$, y_{12} , y_{13}) and extra regional attraction index (y_{10}). Finally, the fourth principal component is mainly determined (with positive sign) by research funds (y_6) and enrollment rate of the best freshmen (y_8) (with negative sign).

References

- Anderson T.W. (1963) *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.
- Atkinson (2002) Robust diagnostics for multivariate analysis and classification problems, in: *Proceedings of the XLI Riunione Scientifica della Società Italiana di Statistica*, Milano.
- Atkinson A.C. and Riani M. (2000) *Robust Diagnostic Regression Analysis*, Springer, New York.
- Biggeri L. and Bini M. (2000) The variability of the italian university cost per student: a multilevel cost function model, in: *ASA 2000, Proceedings of the Section on Statistical Education*, Washington, D.C.: The Association, Indianapolis.
- Box G.E.P. and Cox D.R. (1964) An analysis of transformations (with discussion), *Journal of the Royal Statistical Society, Series B*, 26, 211–246.
- Ceroli A. and Riani M. (1999) The ordering of spatial data and the detection of multiple outliers, *Journal of Computational and Graphical Statistics*, 8, 239–258.
- Ewell P.T. (1999) Linking performance measures to resource allocation: exploring unmapped terrain, *Quality in Higher Education*, 5, 191–208.
- Hubert M., Rousseeuw P. and Verboven S. (2002) A fast method for robust principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60, 101–111.
- Mardia K.V., Kent J.T. and Bibby J.M. (1979) *Multivariate Analysis*, Academic Press, London.
- Riani M. and Atkinson A.C. (2001) A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics*, 10, 513–544.
- Riani M. and Zani S. (1997) An iterative method for the detection of multivariate outliers, *Metron*, 55, 101–117.
- Velilla S. (1995) Diagnostics and robust estimation in multivariate data transformations, *Journal of the American Statistical Association*, 90, 945–951.
- Zani S., Riani M. and Corbellini A. (1998) Robust bivariate boxplots and multiple outlier detection, *Computational Statistics and Data Analysis*, 28, 257–270.