Application of Resampling Methods to the Choice of Dimension in Principal Component Analysis

Ph. Besse and A. de Falguerolles

Laboratoire de Statistique et Probabilités, U.A. CNRS D0745, Université Paul Sabatier, 31062 Toulouse cedex, France.

Abstract

This paper investigates the problem of the choice of dimension in Principal Component Analysis (PCA). PCA is introduced as a model; a loss function assessing the stability of the fit is considered. The choice of dimension then amounts to the minimisation of an expected loss which has to be estimated. This is achieved by resampling methods. Different bootstrap and jackknife estimates are presented. The behaviour of these estimates are investigated on artificial data and on real data. The resulting choices are confronted with those given by naïve rules.

Keywords: Principal Component Analysis; Optimal Dimension; Bootstrap; Jackknife; Perturbation Theory.

1 Introduction

Principal Component Analysis (PCA) is usually defined as the search for independent linear combinations of the initial variables with maximal variances. It can be also viewed as a dimension reduction of the data aiming at producing graphical displays in a low dimensional subspace. In accordance with Caussinus (1986), PCA is introduced in this paper as a fixed effect model. This classical modelling approach gives rise to a measure of fit which can be formulated in terms of a risk which is the expectation of a quadratic distance. This risk can then be used in order to optimise the choice of a parameter: a metric to measure distances between rows (see Besse et al. 1988), a smoothing parameter (see Besse and Pousse 1992) or the number of dimensions.

This paper discusses this last situation. In section 2, we introduce the model and the risk function. In this context, the choice of dimension amounts to the minimisation of an expected loss function. This is achieved by resampling methods which are presented in section 3. Bootstrap and jackknife estimates are reviewed. An approximation of the latter which is simply derived from the results of the PCA of the data is presented. In section 4, we illustrate the use of these estimates on artificial and real data. We compare their performances to those of certain naïve decision rules based on the results of a standard PCA. A new rule based on the shape of the parallel boxplots of the principal components is introduced. Finally, in section 5 we make some concluding remarks.

2 A PCA stability criterion

2.1 Model and notation

Let $\{Y_i; i = 1, \ldots, n\}$ be n independent vector valued random variables in $\mathbf{R}^p$. The fixed effect model assumes that each realization y_i can be decomposed as a fixed effect $\mu + z_i$ and an error term ε_i:

$$\begin{aligned} y_i &= \mu + z_i + \varepsilon_i \qquad (1) \\ \text{with } \mu, z_i \in \mathbf{R}^p, \sum_{i=1}^{n} z_i &= 0 \quad \textit{and}, \quad \forall i, E(\varepsilon_i) = 0, \mathrm{var}(\varepsilon_i) = \sigma^2 \mathbf{I}_p. \end{aligned}$$

Furthermore, we assume that vectors $\{z_i; i = 1, \ldots, n\}$ lie in a q dimensional subspace $\mathcal{E}_q$ of $\mathbf{R}^p$.

The least squares estimation of this model is obtained by solving the following minimisation problem:

$$\min_{\mathcal{E}_q, \mu, z_i} \left\{ \frac{1}{n} \sum_{i=1}^{n} \|y_i - \mu - z_i\|^2 ; \mu \in \mathbf{R}^p, \dim(\mathcal{E}_q) = q, z_i \in \mathcal{E}_q, \sum_{i=1}^{n} z_i = 0 \right\}. \qquad (2)$$

Let $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ be the column mean; then

$$\sum_{i=1}^{n} \|y_i - \mu - z_i\|^2 = \sum_{i=1}^{n} \|y_i - \overline{y} - z_i\|^2 + n \|\overline{y} - \mu\|^2$$

Hence, denoting by $\mathbf{X}$ the column centered data matrix and by $\mathbf{Z}$ the (n, p) matrix whose rows are the z_i, the minimization problem (2) is equivalent to:

$$\min_{\mathbf{Z}, \mu} \left\{ \frac{1}{n} \|\mathbf{X} - \mathbf{Z}\|_2^2 + \|\overline{y} - \mu\|^2 ; \mu \in \mathbf{R}^p, \mathrm{rank}(\mathbf{Z}) = q, \sum_{i=1}^{n} z_i = 0 \right\}, \qquad (3)$$

where $\|.\|_2$ is the SSQ norm on (n, p) matrices.

The solution of that problem is well known and is given by considering the order q generalized singular value decomposition (SVD) of $\mathbf{X}$:

$$\widehat{\mathbf{Z}}_q = \sum_{k=1}^{q} \lambda_k^{1/2} u_k v_k', \qquad (4)$$

where u_k and v_k are respectively the orthogonal eigenvectors of $\mathbf{X}\mathbf{X}'/n$ and those of the empirical covariance $\mathbf{S} = \mathbf{X}'\mathbf{X}/n$ —or correlation matrix in case of standardized variables— associated with the eigenvalues λ_k arranged in decreasing order. We denote by $\mathbf{V}_q$ the (p, q) matrix which stores the column vectors $\{v_k; k = 1, \ldots, q\}$ and by $\mathbf{U}_q$ the (n, q) matrix which stores the column vectors $\{u_k; k = 1, \ldots, q\}$. The different estimates in problems (2) and (3) are then given by:

$$\begin{aligned} \widehat{\mu} &= \overline{y}, \\ \widehat{\mathcal{E}_q} &= \mathrm{vect}\left\{v_k; k = 1, \ldots, q\right\}, \text{ subspace spanned by the } v_k\text{s}, \\ \widehat{P_q} &= \mathbf{V}_q \mathbf{V}_q', \text{ the eigenprojector matrix}, \\ \widehat{z_i} &= \widehat{P_q} x_i, \\ \widehat{\mathbf{Z}_q} &= \mathbf{X} \widehat{P_q}. \end{aligned}$$

Moreover,

$$\mathbf{C}_q = \mathbf{U}_q \mathrm{diag}(\lambda_k; k = 1, \ldots, q) = \mathbf{X}\mathbf{V}_q$$

gives the q principal variables or component scores which lead to graphical displays for the rows.

In fact, the quality of all these estimates seriously depends on the choice of the number q of dimensions or components which are retained. Our aim is to define an approximation which should remain stable with respect to sampling alterations. Therefore a criterion of stability has to be defined.

2.2 Stability criterion

A natural criterion for the stability of the vectors z_i (see Besse et al. 1988) is given by the following loss function:

$$L_q = \frac{1}{n}\sum_{i=1}^{n} \|z_i - \widehat{z}_i\|^2_{\mathbf{A}},$$

where $\mathbf{A}$ defines any euclidean metric in $\mathbf{R}^p$. Its expectation, the associated risk function, is used to optimize the euclidean metric within the vector space $\mathbf{R}^p$; it leads to a Gauss-Markov-like property in the framework of PCA: the optimal metric is given by the inverse matrix of the covariance of the noise. When estimating the dimension, this approach unfortunately raises a problem of indetermination: it is not possible to estimate simultaneously σ and the optimal dimension q because the natural estimate for σ^2 is the average of the $(p-q)$ smallest eigenvalues.

This drawback can be overcome by considering the subspace $\widehat{\mathcal{E}_q}$, which is asymptotically consistant (see Fine and Pousse 1990) rather than the vectors z_i. The loss function then becomes:

$$L_q = Q(\mathcal{E}_q, \widehat{\mathcal{E}_q}) = \frac{1}{2}\left\|P_q - \widehat{P_q}\right\|_2^2 = q - \mathrm{tr}P_q\widehat{P_q}, \tag{5}$$

where Q is based on the usual SSQ norm of matrices which is applied to measure distances between projectors and thus to measure distances between the associated subspaces. In that context, $\mathrm{tr}P_q\widehat{P_q}$ is also the sum of the squared canonical correlation coefficients between the component sets which respectively span $\mathcal{E}_q$ and $\widehat{\mathcal{E}_q}$. This approach follows the principle which underlies indices based on the sum of correlation coefficients between pairs of principal components which have been introduced by Daudin *et al.* (1988,1989): in a PCA, the results are assumed to be reliable if the representation subspace is stable.

Finally, the risk function is defined by considering the following expectation:

$$R_q = EQ(\mathcal{E}_q, \widehat{\mathcal{E}_q}). \tag{6}$$

To compute estimates for this risk function, we resort to computer intensive methods. R_q being symmetrically defined since its value is invariant under any permutation of the observations y_i, resampling methods such as the bootstrap and the jackknife are natural candidates. This strategy has been previously advocated in a different context by Beran and Srivastava (1985) in order to estimate confidence regions for differentiable functions of the covariance matrix or by Daudin et al. (1988, 1989) to study certain stability indices of PCA.

3 Risk estimates

3.1 Bootstrap

According to Efron (1982), a bootstrap estimate of a mean square risk is defined by:

$$\widehat{R_{Bq}} = \frac{1}{B}\sum_{b=1}^{B}(q - \mathrm{tr}P_q^{*b}\widehat{P_q}) = q - \frac{1}{B}\sum_{b=1}^{B}\mathrm{tr}P_q^{*b}\widehat{P_q}, \tag{7}$$

where B is the number of bootstrap replications, P_q^{*b} is the eigenprojector which is obtained by computing the b$^{\text{th}}$ sample PCA.

3.1.1 Bootstrap of residuals

In model (1) the errors ε_i are assumed to be i.i.d. Thus the most natural way to compute an estimate for R_q is to bootstrap the residuals. This leads to the following algorithm[1]:

[1]all algorithms have been programmed in S language (see Becker et al. 1988).

```
Compute the PCA of X
For q = 1 to (p − 1) do

    Calculate residuals:  e_i = λ_{q+1}^{1/2} V u_i; i = 1, ..., n
    whose covariance matrix is λ_{q+1} I_p
    Repeat B times

        Draw the e_i^* with replacement from the e_i
        Construct a bootstrap sample X*:  x_i^* = \hat{P}_q x_i + e_i^*
        Compute the PCA of X*
        Cumulate:  tr \hat{P}_q P_q^{*b}

    End
    Calculate:  \widehat{R_{Brq}} by (7)

End
Compare \widehat{R_{Brq}} values
```

Unfortunately, this requires a large amount of computation: B times $(p-1)$ singular value decompositions (SVD) must be performed. Indeed this can be computationaly expensive and time consuming. Hence we propose a moderate value for B, say 30 .

3.1.2 Bootstrap of data

A less expensive estimate can be computed by naïvely bootstrapping the data. It should be emphasized that this is somewhat contradictory with model (1) since a bootstrap of the data assumes that the Y_i are i.i.d.. The algorithm becomes:

```
Compute the PCA of X
Repeat B times

    Construct X* by drawing the x_i^* with replacement from the x_i
    Compute the PCA of X*
    For q = 1 to (p − 1) do

        Cumulate:  tr \hat{P}_q P_q^{*b}

    End

End
Calculate:  \widehat{R_{Bdq}} by (7)
Compare \widehat{R_{Bdq}} values
```

The algorithm performs B times a SVD to compute the naïve bootstrap estimate. Hence we propose the value $B = 100$.

3.2 Jackknife estimate

Under the same rough assumptions as those considered for the bootstrap of the data, we can also compute the jackknife estimate of R_q. The jackknife estimate (see Efron 1982) is defined by:

$$\widehat{R_{JKq}} = (n-1)\left(q - \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr} P_q^{(i)} \widehat{P_q}\right). \tag{8}$$

where we denote by $P_q^{(i)}$ the eigenprojector obtained by performing a PCA of the data matrix $\mathbf{X}^{(i)}$ obtained from $\mathbf{X}$ by deleting the ith row.

This estimate is obtained by the following algorithm:

```
Compute the PCA of X
For i = 1 to n

    Compute the PCA of X^(i)
    For q = 1 to (p − 1) do
        Cumulate:  tr P_q^(i) \hat{P}_q
    End

End
Calculate:  \hat{R}_{JKq} by (8)
Compare \hat{R}_q values
```

This risk estimate requires the computation of n SVDs but, if n is too large, the jackknife replications can be performed only $B < n$ times for a random sample of rows and thus with the same computational cost as that of bootstrap. Moreover by using the so-called jackknife-after-bootstrap (Efron, 1992), an approximation of the jackknife can be derived from the bootstrap of the data with supplementary data management: each bootstrap sample gives an approximation of the estimate jackknifed with respect to all units which are not sampled.

3.3 Approximation of the jackknife estimate

All estimates above are computationaly expensive. Hence we now consider an approximation of the jackknife which does not require much computational effort. The elimination of the i^{th} row in the data set $\mathbf{X}$ provides an empirical covariance matrix $\mathbf{S}^{(i)}$ which can be decomposed as follows:

$$\mathbf{S}^{(i)} = \mathbf{S} + (n-1)^{-1}(\mathbf{S} - x_i x_i') - (n-1)^{-2} x_i x_i'. \tag{9}$$

If n is large enough, it is quite acceptable to consider that any row elimination introduces only a small perturbation in further computations. Perturbation theory (Kato, 1966) then enables us to write Taylor's expansions of the eigenelements of $\mathbf{S}$ and thus of the eigenprojectors which are introduced in (8). These results lead to the following Taylor's expansion of the jackknife estimate (see Besse, 1992 for more details):

$$\widehat{R_{JKq}} = \widehat{R_{Pq}} + O((n-1)^{-2}).$$

$\widehat{R_{Pq}}$, an analytic approximation of the jackknife estimate is given by:

$$\widehat{R_{Pq}} = \frac{1}{n-1} \sum_{k=1}^{q} \sum_{j=k+1}^{p} \frac{\tau_{jk}}{(\lambda_j - \lambda_k)^2} \text{ with } \tau_{jk} = \frac{1}{n} \sum_{i=1}^{n} c_{ik}^2 c_{ij}^2 \tag{10}$$

where c_{ij} denotes the general entry of the principal components matrix $\mathbf{C}$. It is clear that this approximation is easily computed from the results of the PCA of the initial data set.

The validity condition for this expansion is easy to check. It relates n and q as follows:

$$n > \frac{\|\mathbf{S}\|_2^2}{\inf\{(\lambda_k - \lambda_{k+1}); k = 1, \ldots, q\}}.$$

The results above show the importance of the gap between successive eigenvalues. The leading term in $\widehat{R_{Pq}}$ depends on the difference between the eigenvalues associated with the last retained dimension and the first neglected one. This is consistant with intuitive considerations: if the difference $(\lambda_q - \lambda_{q+1})$ is large enough, data perturbations cannot lead to the swapping of the associated eigenvectors v_q and v_{q+1}. Formula (10) also highlights the influence of the fourth order moments of the principal components: high values for the average τ_{jk} of cross products of squared principal components increase the estimated risk.

Finally this approximation simply extends to the PCA of standardized variables. Indeed, perturbation of the correlation matrix instead of the covariance matrix (9) introduces extra terms, but Besse (1992) shows that they all vanish in the approximation calculus. Thus, Taylor's expansion of the risk estimate (10) also holds for centered and standardized PCA.

Note that Daudin et al. (1989) have also suggested in a similar context the use of the infinitesimal jackknife to avoid large computational costs and to obtain analytic expansions. Unfortunately, in their approach, it is impossible to check the validity of the results since functions of ordered eigenvalues do not have continuous derivatives.

4 Illustration

4.1 Motivation

We now turn our attention to data and present the results obtained in considering artificial and real data. The aim of this illustration is to investigate the behaviour of the different estimates of the risk function and to confront the resulting decisions with those of standard naïve rules. The naïve decision rules which we use as benchmarks are:

- the screegraph (search for a knee),
- the parallel boxplots of the principal components (search for a reduction of the boxes visualising the quartiles of principal components),
- Kaiser's rule as implemented in SAS (1989) automatic selection procedure (selection of eigenvalues larger than average).

For an extensive discussion of such rules and some others, a good reference is Jolliffe (1986).

4.2 Artificial data

The 5 artificial data sets each consist of a $(100, 10)$ matrix: $p = 10$ variables observed on $n = 100$ units. They are obtained by adding to the same matrix of fixed effects with known rank ($q = 4$) a simulated white noise of controlled standard deviation:

1. The fixed effects have been constructed as follows:
   - first, a $(100, 10)$ matrix of pseudo-random numbers uniformly and independently distributed in the range $[0, 20]$ was generated;
   - then the $(100, 10)$ matrix of non centered fixed effects was derived from its SVD by constructing its best least squares approximation of rank $q = 4$.
2. The noise was simulated as follows:
   - 1000 observations of independently and identically normally distributed pseudo-random variables with mean 0 and variance 1 were generated;
   - 5 simulated noises of standard deviation σ ($\sigma = 1, 3, 5, 7, 10$) were then obtained by multiplication by σ.

Since the data were simulated, the true quadratic risk can be computed and reported. To mimic practical situations, in the formula giving the quadratic loss L_q we have used the order q approximations of the vectors z_i which coincide with the z_i for q greater or equal to the true

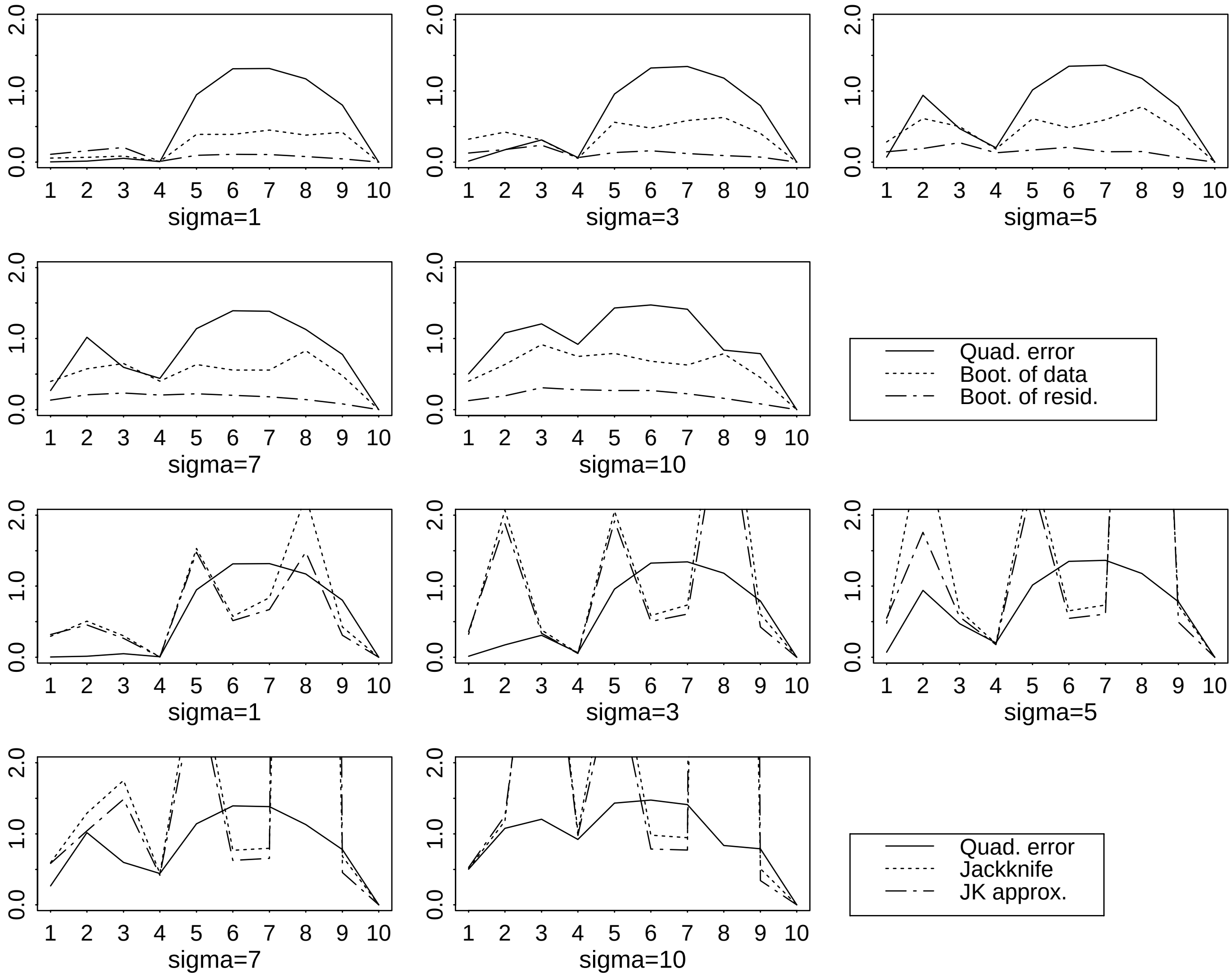

Figure 1: *Artificial data: risk estimates.*

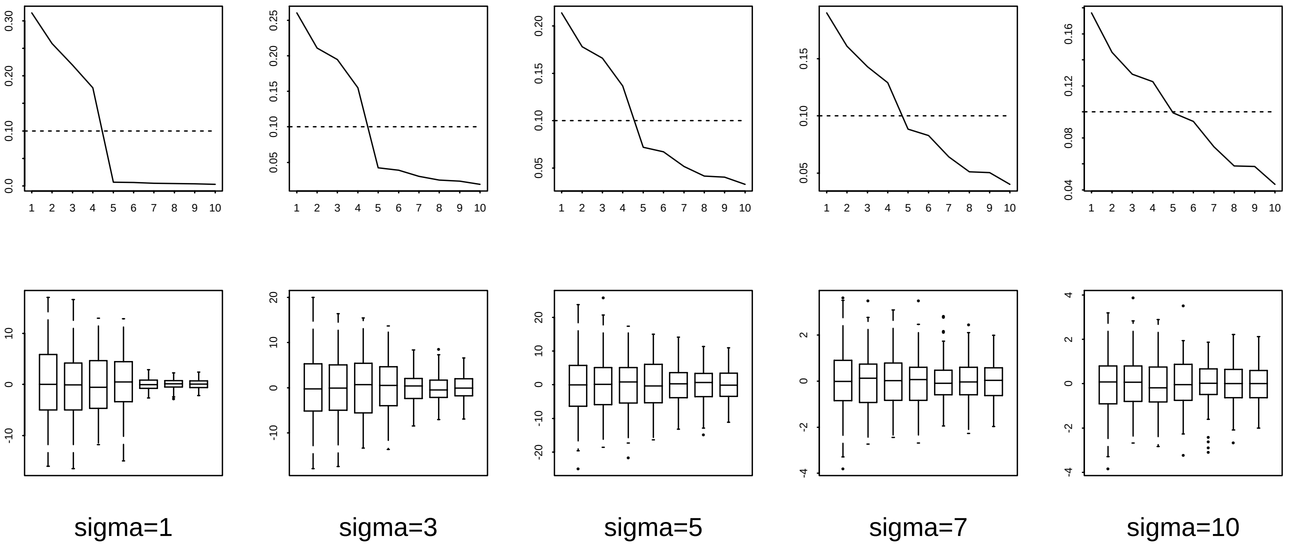

Figure 2: *Artificial data: screegraphs (with Kaiser's cut-off value) and parallel boxplots of components.*

dimension. The resulting quadratic risk (see Figure 1) presents a noticeable local minimum for the true dimension, althought it reaches its minimum for $q = 1$. This last solution is not very interesting for graphical displays since it is often linked to some trivial general effect. All estimations of the risk function behave quite well and point out the true dimension for moderate noise ($\sigma \leq 5$). This is also the case for the naïve rules which clearly recognize the true dimension (see Figure 2). In presence of larger noise ($\sigma \geq 7$) the pattern is less clear. Although the decision rules do not reject the true dimension ($q = 4$), they show a somewhat contrasted pattern: while the bootstrap of residuals is rather uninformative, the jackknife, its approximation and the bootstrap of data also suggest a value of 6 or 7 for q (possibly an artefact introduced by the noise).

For these simulated data sets, the unanimity of the decision rules likely stems from the smoothness of the data: there is no contamination, outlier,... which is not usually the case for real data.

4.3 Pollution data

We now consider the data that McDonald and Schwing (1973) assembled to study the effects of air pollution on mortality. The $p = 16$ variables were measured at each of $n = 60$ areas in the United States. Our outlook being exploratory, we fit a (standardized) PCA to all variables after a logarithmic transform of the 3 pollution potential variables.

Kaiser's rule (see figure 3) selects $q = 5$ whereas the jump in the screegraph and the size reduction in the parallel boxplots only suggest $q = 3$. Interestingly, the parallel boxplots of the principal components provide some insight into the stability of the first $q \leq 3$ dimensional principal subspaces: moderate box but heavy whiskers of extreme data for the first component, large boxes for the second and the third; smaller boxes with influential points in their whiskers for the last $(p - q)$ components. This is supported by the risk function estimated by bootstrapping or jackknifing the data: the corresponding estimates also point out the stability of the first three-dimensional subspaces. It appears that Kaiser's rule overestimates the dimension needed to obtain a stable principal subspace.

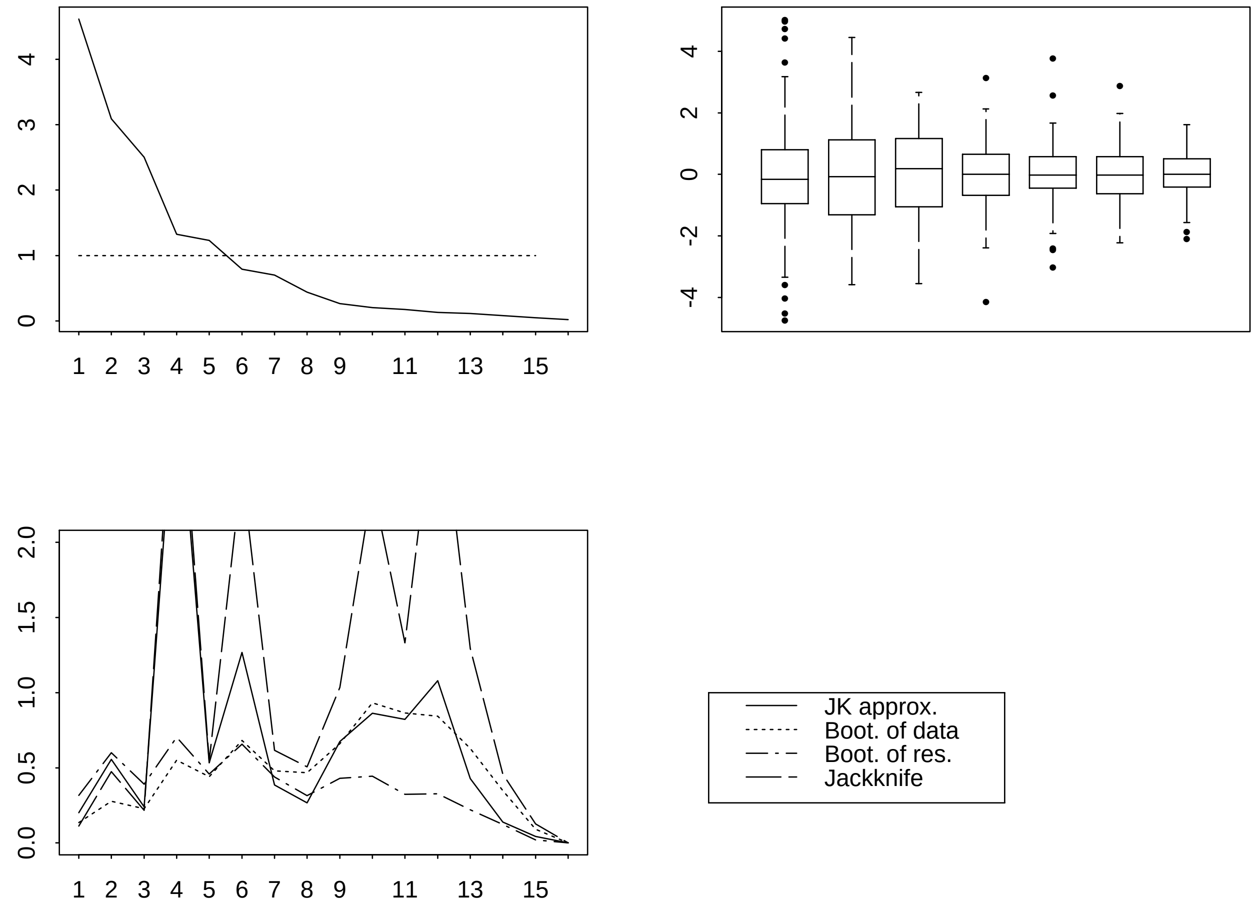

Figure 3: *Pollution data: screegraph (with Kaiser's cut-off rule), parallel boxplots and risk estimates.*

5 Concluding remarks

In the light of our experience of PCA, we recommend the following strategy.

1. For starters, a classical PCA is to be considered. Without much computational effort and therefore on a routine basis, its output can include:
   - the screegraph,
   - the parallel boxplots of the components,
   - the approximation of the jackknife estimate along with its validity conditions relating q and n.
2. These intermediate results enable the stability of the embedded principal subspaces to be assessed. In this respect, we find the parallel boxplots and the approximation of the jackknife very helpful. The former highlights the potential effects of some extreme points while the latter stresses the importance of the gap between the last retained and the first neglected eigenvalues.
3. In case of doubt, a bootstrap approach may provide further insight. When the signal-to-noise ratio is moderate, the bootstrap of residuals should be prefered to the bootstrap of data since it takes advantage of the modelling framework which we are considering.

As seen in the examples, the decision rules may offer several values for the estimated rank q. In this case, we advocate the choice of a low value for the sake of stability. It is clear that the simulated data support our recommendation.

Finally, it is worthwhile noting that the general framework in which we have formulated the problem of dimension selection has some pedagogical implications. Our modelling and optimisation approach gives either natural grounds or insight into the resampling techniques used to estimate the risk function. It also provides useful guidelines for the selection problem, which can be used even in a purely exploratory context.

References

Becker, R.A., Chambers, J.M., Wilks, A.R. (1988). *The New S Language, a Programming Environment for Data Analysis and Graphics*, Wadsworth and Brooks/Cole, Pacific Grove, Ca 93950.

Beran, R., Srivastava, M.S. (1985). Bootstrap Tests and Confidence Regions for Functions of a Covariance Matrix. *The Annals of Statistics*, 13, 95-115.

Besse, Ph. (1992). PCA Stability and Choice of Dimensionality. *Statistics & Probability Letters*, 13, 405-410.

Besse, P., Caussinus, H., Ferré, L., Fine, J. (1988). Principal Components Analysis and Optimization of Graphical Displays. *Statistics*, 19, 301-312.

Besse, Ph., Pousse, A. (1992). Extension des Analyses Factorielles, in *Modèles pour l'Analyse des Données Multidimensionnelles*, J.J. Droesbeke et al. (eds.), Economica, Paris.

Caussinus, H. (1986). Models and Uses of Principal Component Analysis, in *Multidimensional Data Analysis*, J. de Leeuw et al. (eds.), DSWO Press, Leiden, 149-170.

Daudin, J.J., Duby, C., Trécourt, P. (1988). Stability of Principal Component Analysis Studied by Bootstrap, *Statistics*, 19, 241-158.

Daudin, J.J., Duby, C., Trécourt, P. (1989). PCA Stability Studied by the Bootstrap and the Infinitesimal Jackknife Method, *Statistics*, 20, 255-270.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Methods*, SIAM, Philadelphie.

Efron, B. (1992). Jackknife-after-Bootstrap Standard Errors and Influence Functions (with discussion), *Journal of the Royal Statistical Society*, series B, 54, 83-127.

Fine, J., Pousse, A. (1991). Asymptotic Study of the Multivariate Functional Model; Application to the Metric Choice in PCA, *Statistics*, to appear.

Jolliffe, I. (1986). *Principal Component Analysis*, Springer-Verlag, New-York.

Kato, T. (1966). *Perturbation Theory for Linear Operator*, Springer-Verlag, New-York.

McDonald, G.C., Schwing, R.C. (1973). Instabilities of Regression Estimates Relating Air Pollution to Mortality. *Technometrics*, 15, 463-481.

SAS (1989), SAS/STAT User's Guide, volume 2, Version 6, fourth edition, Sas Institute Inc, Cary.