Draft from : COMPSTAT, Physica Verlag, Heidelberg, 1996, Alberto Prats, Editor, p 205 – 210.

Assessing Sample Variability in the Visualization Techniques related to Principal Component Analysis : Bootstrap and Alternative Simulation Methods.

Frederic Chateau[1], Ludovic Lebart[2]

[1] ENST, 46 rue Barrault; 75013 Paris, France. E-mail chateau@inf.enst.fr
[2] CNRS-ENST, 46 rue Barrault; 75013 Paris, France. E-mail lebart@eco.enst.fr.

Key words : Bootstrap, Principal component analysis, Correspondence analysis, Simulation

1.Introduction

Bootstrap distribution-free resampling technique (Efron, 1979) is frequently used to assess the variance of estimators or to produce tolerance areas on visualization diagrams derived from principal axes techniques (correspondence analysis (CA), principal component analysis (PCA)). Gifi (1981), Meulman (1982), Greenacre (1984) have done a pionneering work in the context of two-way or multiple correspondence analysis. In the case of principal component analysis, Diaconis and Efron (1983), Holmes (1985, 1989), Stauffer et al. (1985), Daudin et al. (1988) have adressed the problem of the choice of the relevant number of axes, and have proposed confidence intervals for points in the subspace spanned by the principal axes. These parameters are computed after the realization of each replicated samples, and involve constraints that depend on these samples. Several procedures have been proposed to overcome these difficulties: partial replications using supplementary elements (Greenacre), use of a three-way analysis to process simultaneously the whole set of replications (Holmes), filtering techniques involving reordering of axes and procrustean rotations (Milan and Whittaker, 1995).

We focus on a discussion about advantages and limitations of the partial bootstrap in PCA; the resampling context of CA is markedly different, due to the non-parametric setting of the contingency table analysis. However, for some applications, bootstrap may produce unrealistic replications.

2. About bootstrap in the framework of PCA

Let X be a (n,p) data table. It is usual to draw with replacement observations from the initial sample, observation i being characterized by its whole pattern of responses (i-th row of X). The appearance of twice or three times the same pattern

is a zero-probability event that is much more influential in the multidimensional case. This is all the more evident in the case of Multiple Correspondence Analysis (MCA, or Homogeneity Analysis). When dealing with p nominal variables (variable s having p_S categories) the number of possible different patterns is $m = \Pi p_s$; in a frequently occurring case of 20 questions having each *4* categories, $m = 4^{20}$. An alternative procedure consists of resampling by generating row vectors consistent with the observed covariance structure, but allowing for new patterns. This induces to perform a classical parametric simulation using the multivariate normal distribution based on the observed covariance matrix. Repeated patterns are thus rather improbable

Various generalization of the original bootstrap have been proposed, leading to several smoothing or weighting schemes (a review is included in: Barbe and Bertail, 1995). In the context of the assessment of eigen-elements, a specific procedures can be used together with the classical ones: the damped bootstrap. Each observation is left unchanged with probability π, or replaced by any other observation with probability $(1-\pi)$, leading to a continuous scale of resampling, from the unchanged sample $(\pi = 1)$ up to the bootstrap $(\pi = 1/n)$. It can lead to a perturbated set of replications in various contexts; for $\pi<0.3$, the damped bootstrap remains very close to the original bootstrap. In such a case, the probability for an observation to be absent from a replicate is, asymptotically with n, $(1-\pi)e^{-(1-\pi)}$ instead of e^{-1} in the classical bootstrap. If π is chosen close to *1*, the columns of *X* can be resampled independently, producing non parametric perturbation of the data.

Regardless qualities of replications, it has been stressed by several authors that the user interested in the bootstrap variability of eigenvalues and eigenvectors is dealing with a non-standard application of bootstrap. Whereas replication of the covariance matrix is straightforward, identification and comparisons of the eigen-elements resulting from PCA of replicated matrices leads to difficulties.

Suppose that observation vector *i* (*i-th* row of *X*) has a contribution $c(i, \alpha)$ to the variance along axis α resulting from PCA of the observed covariance matrix. If the difference between two consecutive eigenvalues is such that $(\lambda_\alpha - \lambda_{\alpha+1}) \leq c(i, \alpha)$, we may expect rotations (or exchanges) of axes depending upon the bootstrap weight assigned to i (for applications of perturbation theory to PCA, see for instance Escofier and Leroux, 1972; Benasseni, 1986). We may also expect reflections of axes due to the arbitrary sign of the eigenvectors. We may try to identify as well as possible homologous axes within the set of replicates (Milan and Whittaker, 1995), or choose a common reference space to position the whole set of replicates (partial bootstraping). The two approaches provide the user with distinct tolerance regions for points location on the obtained maps.

3. Partial Replications

Partial bootstrap making use of projections of replicated elements on the reference subspace provided by Singular Value Decomposition of the observed covariance matrix has several advantages for data analysts. From a descriptive standpoint, this initial subspace is better than any subspace perturbated by a random noise. In fact, it is the expectation of all the pertubated subspaces (replicates). The plane spanned by the first two axes, for instance, provides nothing but a point of view on the data set. In this context, to apply the classical non-parametric bootstrap to PCA, one may project variable-points in the reference common subspace according to two procedures :

- (1) Projection using a stacking of the covariance (or correlation) matrices.

We use here the property that SVD of a covariance matrix C (considered as a data matrix) leads to diagonalization of the matrix C^2, and produces the same unit α-th eigenvectors than the PCA of the original data matrix (with eigenvalues λ_α^2 instead of λ_α). We can then stack the k replicates C_k of C, and project the rows of the stacked matrices as supplementary elements (variables) on the reference subspace. The analysis of C is by no means necessary since we have already obtained the eigen-vector from the PCA of the initial sample. Note that this situation is very similar to that of MCA, where the projections of replicates categories are obtained from the Burt contingency table which plays the same role as the correlation matrix.

- (2) Scalar products with unit-individual eigenvectors.

From the SVD equation: $X=\Sigma \lambda_\alpha v_\alpha, u'_\alpha$, where v_α and u_α are respectively the α^{th} unit eigenvectors of XX' and $X'X$, we obtain the so-called transition relationships: $u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X' v_\alpha$ (A' denotes the transpose of A). If D_k designates the (n, n) diagonal matrix whose diagonal elements are the bootstrap weights of the replicate k, the projection of the k th replicate of the p variables is given by the p-vector $u_\alpha(k)$ such that : : $u_\alpha(k) = \frac{1}{\sqrt{\lambda_\alpha}} X' D_k v_\alpha$

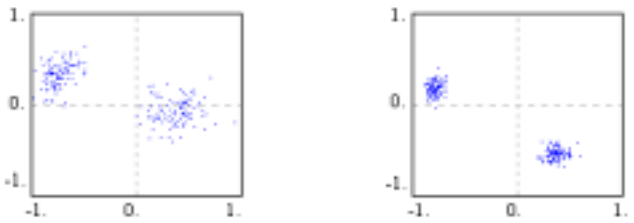
Approaches (1) and (2) coincide in the case of PCA on covariance matrices, but approach (2) is rather unsuitable in the case of PCA on correlation matrices. In both situations, it is easy to compute replicated variances along the reference axes (which evidently are not replicates of the eigenvalues).

4. Results

In order to compare these replication schemes, we generate a series of samples S_m (size *m=30*, to *m=200*) from a multinormal distribution F defined by its (6x6) covariance matrix *C*. We simulate the sampling variability of S_m through two different procedures : (i) generation of other samples from *C*, analogous to S_m , (ii) bootstrap replications of S_m. How relevant for *F* are the statements built from S_m ? In both cases, PCA of S_m provides a unique and common reference space.

(i) The sampling distribution of the j[th] column coordinates $\varphi_{mj\alpha}$ is computed as follows : K independent samples S_{mk} are drawn from matrix *C*; matrices C_{mk} are stacked as supplementary elements in the PCA of S_m.

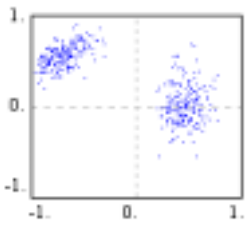
Figures (1) and (2) represent the sample variations of φ_{m1} and φ_{m6} in the principal plane (α=*1,2*) , for samples S_{30} and S_{150}, respectively.

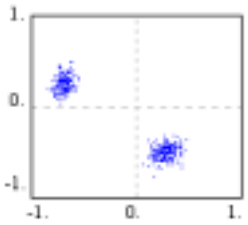


Figures (1) and (2): sampling distribution of col. 1 and 6 on first 2 principal components, for original samples S_{30} and S_{150}, respectively

(ii) Three replication schemes are then carried out in order to assess the variability of the column position in the PCA of S_m : partial bootstrap, damped bootstrap (π varying from 0.1 to 0.9), and damped bootstrap with fixed rows.

Figures (3) and (4), show that the partial bootstrap variability is roughly equivalent to the sampling variability represented above However the distributions of columns points are centered on the original $\varphi_{mj\alpha}$. instead of the projected columns of *C* as in figures (1) and (2). The latter are included in the convex hull of the replicated columns points.





Figures (3) and (4) : partial bootstrap distribution of replicated col. 1 and 6 on the first 2 principal components, original samples S_{30} and S_{150}, respectively

In fact, one can easily see in table (5) that the partial bootstrap total variance of the $\varphi_{mj\alpha}$ is almost the same as the sampling total variance in all cases we report, although slightly optimistic.

Sample size	30	50	75	100	125	150	200
Simulation	0.3871	0.2184	0.1374	0.1090	0.0929	0.0787	0.0569
Part. Bootstrap	0.3726	0.2054	0.1297	0.1048	0.0895	0.0679	0.0549

Table (5) : total sampling and bootstrap variance of column coordinates (axis 1,2,3)

Figure (6) shows how partial damped bootstrap behaves (*m=100*). The fixed-row scheme total variance tends to that of the bootstrap when values of π decrease, as stated above. Column-independent replication scheme gives estimates of the total variance whose range contains the classical partial bootstrap value.

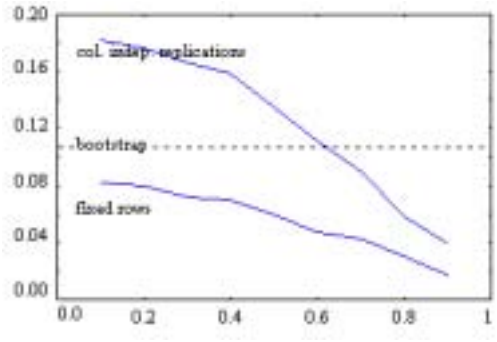


Figure (6) : total damped bootstrap variance of col. coordinates, acc. to π

Conclusion : partial bootstrap gives a fair estimate of the sample variation of eigen-elements. Damped bootstrap leads to a representation of this variability which is coherent with the preceding ones, and may, in other sampling schemes, be more accurate. It may also be used to overcome the specific penalty that bootstrap brings in multiple correspondence analysis, due to the irrealistic replication of individuals.

References

Barbe P., Bertail P. (1995) - The weighted Bootstrap. Springer Verlag.

Benasseni J. (1986) - Stabilité de l'analyse en composantes principales par rapport à une perturbation des données. Revue Statist. Appl., 35, 3, p 49-64.

Daudin J.-J., Duby C., Trécourt P. (1988) - Stability of principal components studied by the bootstrap method. Statistics, 19, p 241-258.

Diaconis P., Efron B. (1983) - Computer intensive methods in statistics. Scientific American, 248, (May), p 116-130.

Efron B. (1979) - Bootstraps methods : another look at the Jackknife. Ann. Statist., 7, p 1-26.

Escofier B., Leroux B. (1972) - Etude de trois problèmes de stabilité en analyse factorielle. Publication de l'Institut Statistique de l'Université de Paris, 11, p 1-48

Gifi A. (1981) - Non Linear Multivariate Analysis, Department of Data theory, University of Leiden. (Updated version : 1990, same title, J. Wiley, Chichester.)

Greenacre M. (1984) - Theory and Applications of Correspondence Analysis. Academic press, London.

Holmes S. (1985) - Outils Informatiques pour l'évaluation de la pertinence d'un résultat en analyse des données. Thèse USTL, Montpellier.

Holmes S. (1989) - Using the bootstrap and the RV coefficient in the multivariate context. in : Data Analysis, Learning Symbolic and Numeric Knowledge, E. Diday (ed.), Nova Science, New York, p 119-132.

Lebart L., Morineau A., Warwick K. (1984) - Multivariate Descriptive Statistical Analysis. J. Wiley, New York.

Meulman J. (1982) - Homogeneity Analysis of Incomplete Data. DSWO Press, Leiden.

Milan L., Whittaker J. (1995) - Application of the parametric bootstrap to models that incorporate a singular value decomposition. Appl. Statist. 44, 1, p 31-49.

Stauffer D. F., Garton E. O., Steinhorst R. K. (1985) - A comparison of principal component from real and random data. Ecology, 66, p 1693-1698