Ordinal and Symbolic Data Analysis
E. Diday, Y. Lechevalier & O. Opitz Eds., Springer, 1996
pp. 359–368

A L_1-norm PCA and a heuristic approach

A. Baccini, Ph. Besse, A. de Falguerolles

Laboratoire de Statistique et Probabilités
URA-CNRS D0745, Université Paul Sabatier
31062 Toulouse cedex France

Abstract: A model for a L_1 principal component analysis (PCA) is considered and discussed. On the other hand, a PCA based on Gini's mean absolute differences is introduced in order to obtain heuristic estimates of the L_1 model. This PCA is connected to the canonical correlation analysis of the original data and their ranks. The method is illustrated on real and artificial data with the aim of achieving proper data reduction.

1 Introduction

Principal Component Analysis (PCA) is a widely used technique for dimension reduction of conventional data. For example, in pattern recognition, dimension reduction is an important stage as it often helps to reduce the cost of clustering and leads to the same classification results with a limited number of features instead of all. To increase the adaptability and flexibility of PCA, it is convenient to consider it as a combination of a model and of an estimation procedure under some probability distribution assumptions. Examples of this approach are the fixed effect PCA model (Caussinus (1986)) which focuses on metric and dimension choices and the generalized bilinear models (Falguerolles & Francis (1992)) which captures the flexibility of log-linear models (McCullagh & Nelder (1989)). Alternatively the PCA of ranks is a well known reduction method in the context of conventional multivariate ordinal data.
In this paper, we define a model which genuinely specifies a L_1-PCA and we investigate a heuristic method of estimation based on rank considerations. This PCA enlarges the scope of probability distribution assumptions which are necessary to model practical situations.
Thus, we consider a fixed effect model for a data matrix in which we assume Laplace error terms. In this framework, the maximum likelihood estimation of row scores, column scores and dispersion parameter of the errors amounts to a nonlinear L_1-norm minimisation problem under constraints (section 2). This problem can be solved numerically for small data sets with some suitable optimization software programme. However, for larger data sets, heuristic approximations with low computational costs are of interest.

As a proposal, we introduce in section 3 a method which takes into account Gini's mean absolute differences of the variables. This method is derived from the canonical correlation analysis of two matched pair sets of variables: the original variables and their respective ranks. In section 4, we consider real and artificial data to compare our approach to ordinary PCA and to the nonlinear L_1-biplots methodology.

2 A L_1 fixed effect model

2.1 Model and notation

Let $\{X_i, i = 1, \ldots, n\}$ be n independent random variables with values in $\mathbf{R}^p$. The fixed effect model assumes that each X_i can additively be decomposed as a fixed effect z_i, $z_i \in \mathbf{R}^p$, and a random error term ε_i also valued in $\mathbf{R}^p$

$$X_i = z_i + \varepsilon_i \tag{1}$$

with $E[\varepsilon_i] = 0$ and $\mathrm{Var}(\varepsilon_i) = 2\sigma^2 I_p$.
To consider a fixed effect model of PCA, we assume that the vector $z_i, i = 1, \ldots, n$, lie in a q-dimensional linear subspace E_q of $\mathbf{R}^p$ (Caussinus (1986)) and that $\sum_{i=1}^n z_i = 0$.
Assuming a normal distribution $\mathcal{N}(0, 2\sigma^2)$ for the error terms ε_{ij}, the fixed effect model leads to standard Gaussian PCA (Anderson (1958)). To introduce a L_1-norm PCA, we consider that the error components ε_{ij} are independent and identically distributed according to a centered Laplace (double-exponential) distribution whose density function is

$$f(x) = \frac{1}{2\sigma} \exp\,(-\frac{1}{\sigma}\,|x|). \tag{2}$$

2.2 Identification

Let us denote by Z the $n \times p$ matrix of fixed effects. Its rank is q and Z can consequently be written as $Z = MN^T$, where M and N are respectively $n \times q$ and $p \times q$ matrices of rank q. This factorisation is not unique, but particular product forms are conveniently derived from the singular value decomposition (SVD) of Z:

$$Z = ALB^T \Longleftrightarrow z_{ij} = \sum_{k=1}^{q} l_k a_{ik} b_{jk}. \tag{3}$$

In formula (3), L is diagonal with the singular values l_k arranged in decreasing order, while a_{ik} and b_{jk} are coordinates of the dual singular vectors a_k and b_k $(k = 1, \ldots, q)$ which are identified by

$$\sum_{i=1}^{n} a_{ik}a_{ik'} = \sum_{j=1}^{p} b_{jk}b_{jk'} = \delta_{kk'}, \tag{4}$$

where $\delta_{kk'}$ denotes the Kronecker symbol. Moreover, since $\sum_{i=1}^{n} z_i = 0$, the following constraints should be satisfied:

$$\sum_{i=1}^{n} a_{ik} = 0, \;\; k = 1, \ldots, q. \tag{5}$$

In the estimation, these constraints are imposed on the row coordinates. Generalized SVD of Z (see Greenacre (1984), appendix A) can be envisaged in some situations.

2.3 Plots and biplots

The singular elements l_k, a_k and b_k are combined to compute scores used in a range of diagrams: isometric plots for rows, isometric plots for columns and biplots for rows and columns (Gabriel (1971); Joliffe (1986), chapter 5).

2.4 Maximum likelihood estimation

Let X be the column centered $n \times p$ matrix of the observed values with general term x_{ij}. The log-likelihood of the model defined in subsection 2.1 is

$$l(X; Z, \sigma) = -np \log 2\sigma - \frac{1}{\sigma} \sum_{i=1}^{n} \|x_i - z_i\|, \tag{6}$$

$$\text{where } \; \|x_i - z_i\| = \sum_{j=1}^{p} |x_{ij} - z_{ij}|. \tag{7}$$

When q is known, the maximum likelihood estimates for the parameters z_{ij} in the proposed model are solutions to the L_1 problem

$$\min_{z_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{p} |x_{ij} - z_{ij}| \tag{8}$$

subject to identification constraints such as considered in subsection 2.2. The estimate for σ is then given by

$$\hat{\sigma} = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} |x_{ij} - \hat{z}_{ij}|. \tag{9}$$

For small data sets, the problem (8) can be directly solved with some suitable optimization software programme (for example, Gino©). However, for the sake of computational ease, we introduce in the next section a method aimed at producing fast heuristic solutions.

The assumption that q is known is quite strong. For a review of the estimation of q in the context of classical PCA the reader is refered to Besse & Falguerolles (1993).

2.5 Preprocessing of data

Usually a PCA problem involves nuisance parameters for the variables: location and possibly dispersion. In the classical form of PCA, these are estimated by the sample means and the standard deviations. In this context, robust estimates should be considered to preprocess the data (see, for example, Huber (1981)). This aspect will not be pursued further in this article.

2.6 Remark

Although this is not our primary goal, this L_1 model might be regarded as another example of a wide variety of methods aimed at "robustifying" PCA (see, for example, Jolliffe (1986)). Here we briefly review some of the works in this area. One strategy is to apply PCA to a robust estimate of the covariance or the correlation matrix; a review of this approach is given in Rivest & Plante (1988); an example of robust biplot is proposed in Daigle & Rivest (1992). An alternative strategy is to look for robust principal components as solutions of a variational problem in which a robust measure of dispersion (rather than the variance) is optimized; this approach is reviewed, with special attention to the L_1-norm, in Galpin & Hawkins (1987). Of related interest is the detection of influencial observations (Critchley (1985)). At last, another option is to consider nonlinear biplots, introduced in Gower & Harding (1988), which generalize the classical biplots for any Euclidean embeddable metric (for example the L_1-norm).

3 A heuristic approach

In this section we introduce a method which partly considers Gini's mean absolute differences of the variables rather than variances as it is the case with ordinary PCA. For this reason, the proposed method will be called Gini's PCA.

3.1 Principle

The idea of the proposed method is to consider the CCA (canonical correlation analysis) of the following two matched sets of variables: the variables of one set, those of interest, take the centered values of the observed variates, while the variables of the other set are derived from the ranks of the former (as usual, in presence of ties, the average rank for the tied values are considered). Let R be the $n \times p$ matrix of transformed ranks associated with matrix X (the original ranks, which range from 1 to n, are centered and divided by n). Let us denote by S_{XX} and S_{RR} respectively the empirical covariance matrices of X and R ($S_{XX} = \frac{1}{n}X^TX$ and $S_{RR} = \frac{1}{n}R^TR$) and by S_{XR} the empirical covariance matrix between X and R ($S_{XR} = \frac{1}{n}X^TR$). For the sake of simplicity we assume that S_{XX}, S_{RR} and S_{XR} are of full rank p. It is well known that the diagonal terms of S_{XR} are equal to one fourth of Gini's mean absolute differences of the original variables (Stuart (1954)). For the jth variable ($j = 1, \ldots, p$) this index is defined by

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}\left|x_i^j - x_l^j\right|. \tag{10}$$

The extra-diagonal terms have also some interesting properties (Schechtman & Yitzhaki (1987)) which are not considered in this article.

3.2 Estimation of the fixed effects

In ordinary PCA the matrix of fixed effects is estimated by the order q approximation of the data matrix. We propose here an alternative strategy. Let us consider the CCA of matrices X and R (see, for instance, Mardia *et al.* (1979)) which rests in practice upon the SVD of matrix

$$K = S_{XX}^{-1/2} S_{XR} S_{RR}^{-1/2}. \tag{11}$$

Then the SVD of K can be written in the form $K = U_{XR}\Lambda_{XR}V_{XR}^T$. In our heuristic approach, we select its first q singular elements and estimate the matrix of fixed effects by

$$Z_{XR}^q = XS_{XX}^{-1/2}U_{XR}^q(U_{XR}^q)^TS_{XX}^{1/2}. \tag{12}$$

Thus Z_{XR}^q is formed by linear combinations of the first q (standardized) canonical correlation variables $XS_{XX}^{-1/2}U_{XR}^q$. Note that $Z_{XR}^p = X$ and that the rank of Z_{XR}^q is q since in (12),

$$U_{XR}^q(U_{XR}^q)^T = U_{XR}^q[(U_{XR}^q)^TU_{XR}^q]^{-1}(U_{XR}^q)^T$$

is the projector onto the subspace spanned by the first q left singular vectors of K. Consequently Z_{XR}^q is a rank q approximation to X.

3.3 Plots

As noted in 2.3 all plots and bipots are based on the SVD of

$$Z^q_{XR} = U_Z \Lambda_Z V_Z^T$$

which can be easily performed. However note that the first q singular elements of Z^{q+1}_{XR} are not necessarily equal to those of the SVD of Z^q_{XR} (non-nested singular elements).

3.4 Variants of the method

Variants of the previous method are obtained by simply considering the SVD of

$$K' = S_{XR} \tag{13}$$

rather than K as defined in (11). Note that if S_{XR} is equal to S_{XX}, then the SVD of K' is in fact the eigen-decomposition of S_{XX} used in ordinary PCA of X.

Denoting the SVD of K' by $\widetilde{U}_{XR}\widetilde{\Lambda}_{XR}\widetilde{V}^T_{XR}$, possible estimated fixed effects are given either by

$$Z^q_U = X\widetilde{U}^q_{XR}(\widetilde{U}^q_{XR})^T$$

or by

$$Z^q_V = X\widetilde{V}^q_{XR}(\widetilde{V}^q_{XR})^T.$$

Plots can then be derived from the SVD of Z^q_U or Z^q_V. However simulations have shown that these variants did not provide better fits than the one considered in subsection 3.2.

4 Examples

4.1 The Rudeforth & Bradley (1972) data

These data have been used in Gower & Harding (1988) to illustrate the methodology of nonlinear biplots. They consist of centered log parts-per-million of 12 trace elements (columns) determined at each of 15 sites (rows) in Pembrokeshire (Wales).

Different methods for the choice of dimension indicated that $q = 2$, as considered in Gower & Harding (1988), was a reasonable value; therefore $q = 2$ was assumed. Figures 1 and 2 report the two-dimensionnal plots provided by the first option of our approach (Fig. 1) and Gower & Harding's nonlinear biplot (Fig. 2). The two plots show comparable "sketch" for the columns, but different patterns for the rows.

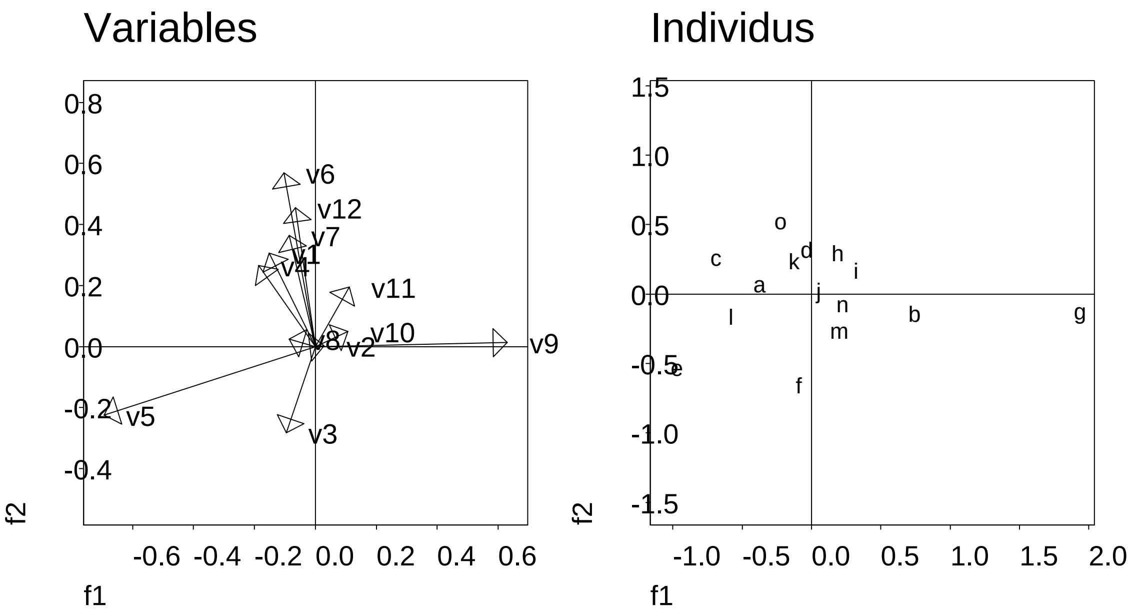

Figure 1: Gini's PCA

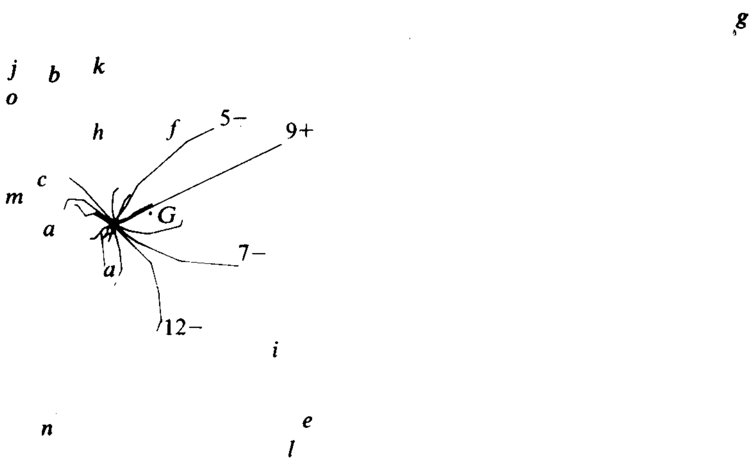

Figure 2: Nonlinear biplot (from Gower & Harding)

4.2 Artificial data

These data were constructed to simulate heavy tailed noise and known signal-to-noise ratio. We computed:

- a 100×10 matrix for the fixed effects with rank $q = 4$ and singular values
$$l_1 = 64,\ l_2 = 49,\ l_3 = 36,\ l_4 = 25,$$
in formula (3);

- a 100×10 noise matrix of i.i.d. Laplace error terms with mean 0 and standard deviation $\sqrt{2}$.

Four data sets were then derived by adding to the matrix of fixed effects the matrix of noise multiplied by a constant σ chosen in order to tune the signal-to-noise ratio. We considered the following values:

$$\sigma_1 = \sqrt{l_1}\ ,\ \sigma_2 = \sqrt{\frac{l_2 + l_3}{2}}\ ,\ \sigma_3 = \sqrt{l_4}\ ,\ \sigma_4 = \sqrt{l_4 - \frac{l_2 - l_3}{2}}\ .$$

Ordinary PCA and our heuristic approach were applied to these data with $q = 4$. The comparison between these approaches was done by computing two overall measures of discrepancy between the real fixed effect and its estimate:

$$D_1(Z, Z^q_{XR}) = \frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left|z_i^j - \hat{z}_i^j\right|,$$
$$D_2(Z, Z^q_{XR}) = [\frac{1}{2np}\sum_{i=1}^{n}\sum_{j=1}^{p}(z_i^j - \hat{z}_i^j)^2]^{1/2}.$$

Note that $D_1(X, Z^q_{XR})$ and $D_2(X, Z^q_{XR})$ estimate σ in (2). Table 1 reports the results and shows that our heuristic approach gives better estimates for the fixed effects than ordinary PCA.

σ values	σ_1		σ_2		σ_3		σ_4	
index of fit	D_1	D_2	D_1	D_2	D_1	D_2	D_1	D_2
PCA	6.29	6.41	5.22	5.27	4.85	4.87	3.30	3.36
heuristic approach	5.50	5.27	4.58	4.35	4.26	4.02	3.23	3.01

Table 1: Comparative values of indexes $D_1(Z, Z^4_{XR})$ and $D_2(Z, Z^4_{XR})$

5 Conclusion

In this article we have introduced a PCA based on Gini's mean absolute differences. This PCA, which can be obtained at very low computational costs, provides heuristic estimates for the general L_1 problem. However, further work need to be done to compare more precisely the biplot obtained by the proposed method to the interesting nonlinear biplot introduced in Gower & Harding (1988). Nevertheless, this work is part of a larger study, which aims at integrating more flexibility in PCA in order to cope with the modeling of more sophisticated probabilistic context (e.g. Diday & Emilion (1996)).

References

Anderson, T.W. (1958): An introduction to multivariate statistical analysis. John Wiley & Sons, New York.

Besse, Ph., & Falguerolles, A. de (1993): Application of resampling methods to the choice of dimension in principal component analysis. In: W. Härdle and L. Simar (eds.): *Computer intensive methods in statistics*. Physica-Verlag, Heidelberg, 167–174.

Caussinus, H. (1986): Models and uses of principal component analysis. In: J. de Leeuw *et al.* (eds.): *Multidimensional Data Analysis*. DSWO Press, Leiden, 149–309.

Critchley, F. (1985): Influence in principal components analysis. *Biometrika*, 72, 627–636.

Daigle, G. & Rivest, L.P. (1992): A robust biplot. *The Canadian Journal of Statistics*, 20, 241–255.

Diday, E. and Emilion, R. (1996): Lattices and capacities in analysis of probabilistic objects. In: *Proceedings of the International Conference on Ordinal and Symbolic Data Analysis*. To appear.

Falguerolles, A. de & Francis, B. (1992): Algorithmic approaches for fitting bilinear models. In: Y. Dodge and J. Whittaker (eds.): *Computational statistics*. Physica-Verlag, Heidelberg, 77–82.

Gabriel, K.R. (1971): The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58, 453–467.

Galpin, J.S., & Hawkins, D.M. (1987): Methods of L_1 estimation of a covariance matrix. *Computational Statistics & Data Analysis*, 5, 305–309.

Gower, J.C., & Harding, S.A. (1988): Non linear biplots. *Biometrika*, 75, 445–455.

Huber, P.J. (1981): Robust statistics. John Wiley & Sons, New York.

Greenacre, M.J. (1984): Theory and applications of correspondence analysis. Academis Press, London.

Jolliffe, I.T. (1986) Principal component analysis. Springer Verlag, New York.

McCullagh, P. & Nelder, J.A. (1989) Generalized linear models. Chapman & Hall, London.

Mardia, K.V., Kent, J.T., & Bibby, J.M. (1979): Multivariate analysis. Academic Press, London.

Rivest, E., & Plante, N. (1988): L'analyse en composantes principales robuste. *Revue de Statistique Appliquée*, 36, 1, 54–66.

Schechtman, E., & Yitzaki, S. (1987): A measure of association based on Gini's mean difference. *Communications in Statistics, Part A–Theory and Methods*, 16, 207–231.

Stuart, A. (1954): The Correlation between variate values and ranks in sample from a continuous distribution. *British Journal of Statistical Psychology*, 7, 37–44.