Insight of a Dreamed PCA

Philippe C. Besse
Laboratoire de Statistique et Probabilités
CNRS – Université Paul Sabatier*

Abstract. This paper reviews some models and associated graphical displays for exploring multivariate data. Some new features developed around principal components analysis (PCA) are proposed. All can easily be driven from a SAS/INSIGHT$^{(\mathrm{TM})}$ like environment. Their aim is to help interpretation and dimensionality choice. They facilitate the investigation of multivariate outliers, clusters or non-linear relationships. They allow to take into account data specificities such as one extra categorical variable or the fact that the data are sampled curves.

1 Introduction

During the last Club (1993) in Toulouse, the SAS$^{(\mathrm{TM})}$ France manager asked the users to dream the future of SAS System. I did it. I dreamt about one unified system having only one syntax, only one logic.... It was a terrific nightmare featuring the imperialistic domination of only one statistical school over the enslaved scientific community. I thought of SAS System, I realized its diversity and its plurality. I was relieved by SAS necessary complexity and went back to sleep. My nights was then illuminated by browsing, brushing, rotating plots and other colored linked graphs. I realized that the future of computational statistics should mainly lay in dynamic graphics as they have been introduced in many softwares (*mac-spin, data-desk, Splus...*) and then SAS/INSIGHT$^{(\mathrm{TM})}$.

The purpose of this paper is to describe what could be the development of one statistical method, namely the principal components analysis (PCA), within a graphical environment as SAS/INSIGHT. Different frameworks have been proposed to define PCA. They often lead to a common mathematical core based on the spectral analysis of a symmetric positive matrix but the interpretation of results may depend on the adopted framework or the assumed model. On the one hand, observations are considered to come from a population, with certain distributional assumptions, and interest lies more in the population itself than in the statistical units. On the other hand, a descriptive approach aims to investigate possible patterns for the units and for the variables. As it has been

*31062 — Toulouse — France. email: besse@cict.fr.

noted by several authors (*e.g.* Gower, 1984), these two aims are somewhat contradictory; strong distributional assumptions seem inconsistent with a descriptive approach. Nevertheless, only the consideration of a model can clarify what kind of arbitrariness is unavoidable; *e.g.* our perception requires an Euclidean framework for graphical displays. Furthermore, a model allows us to discuss what guidelines can be put forward to help the choice of metric, dimensionality or smoothing parameter. This paper aims to emphasize advantages and consequences when introducing the PCA by means of a fixed effect model (Caussinus, 1986) which mainly deals with geometrical assumptions rather than probabilistic ones. It focuses on graphical help and extensions that could be easily introduced in a SAS/INSIGHT like environment in order to improve PCA and its interpretation.

Section 2 defines a general PCA framework as the estimation of a fixed effect model and proposes some simple graphs which aim at helping the PCA interpretation and the choice of dimensionality. Section 3 discusses the optimal choice of metric and gives two examples. The factorial discriminant analysis (FDA) and a kind of projection pursuit method are introduced as optimal PCAs. The later aims at revealing clusters or multivariate outliers. Section 4 makes the fixed effect model suitable for studying sampled curves. This leads to define a new kind of functional PCA featuring spline smoothing. Section 4 generalizes the fixed effect model in order to look for optimal transformations for dimension reduction. This leads to introduce a curvilinear fixed effect model including non-metric PCAs. The popular data of crime rates (SAS, 1989) are revisited to illustrate these different features but the lack of space, in this short review, does not allow to clarify all mathematical developments. These must be found in referenced papers or in a forthcoming book (Besse *et al.* 1995). This modeling approach could be extended to the correspondence analysis of categorical variables but it is another dream...

2 Linear PCA

Consider the following framework,

- $\mathbf{Y}$ denotes the original data matrix displaying observations of p real variables Y^j on n $(n > p)$ statistical units, each weighted by w_i,
- E is the Euclidean p-dimensional space of units where the metric is $\mathbf{M}$; it is introduced to compute distances between units or ultimately

$$\|\mathbf{z}_i - \mathbf{z}_j\|^2_{\mathbf{M}} = (\mathbf{z}_i - \mathbf{z}_j)'\mathbf{M}(\mathbf{z}_i - \mathbf{z}_j).$$

- F is the Euclidean n-dimensional space of variables where the metric is defined by the matrix $\mathbf{D} = \operatorname{diag}(w_1, \ldots, w_n)$.

2.1 A low modeling diet model

The data matrix rows are assumed to be the observations of n independent random vectors $\{\mathbf{y}_i\ ; i = 1, \ldots, n\}$ with the same covariance matrix $\sigma^2\mathbf{\Psi}$ but different expectations $\mathbf{z}_i$

belonging to a common q-dimensional affine subspace A_q of E. In this model, $E(\mathbf{y}_i) = \mathbf{z}_i$ is the so-called fixed effect attached to the ith unit.

To summarize,

$$\begin{array}{l} \{\mathbf{y}_i \; ; i = 1, \ldots, n\}, \; n \text{ independent random vectors of } E, \\ \mathbf{y}_i = \mathbf{z}_i + \mathbf{e}_i \; ; i = 1, \ldots, n \text{ with } \left\{ \begin{array}{l} E(\mathbf{e}_i) = 0, \; \mathrm{var}(\mathbf{e}_i) = \sigma^2 \boldsymbol{\Psi}, \\ \sigma > 0 \text{ unknown}, \\ \boldsymbol{\Psi} \text{ regular and known}, \end{array} \right. \\ \exists A_q, \; q \text{ dimensional affine subspace of } E \text{ such that } \forall i, \mathbf{z}_i \in A_q, (q < p). \end{array} \tag{1}$$

Let $\bar{\mathbf{z}} = \sum_{i=1}^n w_i \mathbf{z}_i$. The model assumptions implies $\bar{\mathbf{z}} \in A_q$. Let then E_q be the q dimensional linear subspace of E such that $A_q = \bar{\mathbf{z}} + E_q$. The parameters to be estimated are E_q and $\mathbf{z}_i, i = 1, \ldots, n$, eventually σ.

2.2 Estimation

Least squares estimation for the problem is formulated as follows

$$\min_{E_q, \mathbf{z}_i} \left\{ \sum_{i=1}^n w_i \left\| \mathbf{y}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 \; ; \; \dim(E_q) = q, \mathbf{z}_i - \bar{\mathbf{z}} \in E_q \right\}. \tag{2}$$

Denoting by $\mathbf{X} = \mathbf{Y} - \mathbf{1}_n \bar{\mathbf{y}}'$ the column centered matrix with $\bar{y} = \sum_{i=1}^n w_i \mathbf{y}_i$ and by $\mathbf{Z}$ the $(n \times p)$ matrix whose rows are the $(\mathbf{z}_i - \bar{\mathbf{z}})'$, as

$$\sum_{i=1}^n w_i \left\| \mathbf{y}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^n w_i \left\| \mathbf{y}_i - \overline{\mathbf{y}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{y}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2,$$

the problem (2) is equivalent to $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{y}}$ and

$$\min_{\mathbf{Z}} \left\{ \left\| \mathbf{X} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}}^2 \; ; \mathbf{Z} \in \mathcal{M}_{n,p}, \; \mathrm{rank}(\mathbf{Z}) = q \right\} \tag{3}$$

with

$$\left\| \mathbf{X} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}}^2 = \mathrm{tr}(\mathbf{X} - \mathbf{Z})' \mathbf{D} (\mathbf{X} - \mathbf{Z}) \mathbf{M}.$$

This is the approximation of a $(n \times p)$ matrix by another $(n \times p)$ matrix with a lower rank. Estimation of parameters E_q and $\{\mathbf{z}_i; i = 1, \ldots, n\}$ of the fixed effect model (1) is obtained by order q generalized PCA of $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$. This PCA is obtained by considering an order q generalized singular value decomposition (SVD) of $(\mathbf{X}, \mathbf{M}, \mathbf{D})$

$$\widehat{\mathbf{Z}}_q = \sum_{k=1}^q l_k^{1/2} \mathbf{u}^k \mathbf{v}^{k'} = \mathbf{U}_q \mathbf{L}^{1/2} \mathbf{V}_q'.$$

Finally the estimates are given by

$$\begin{aligned} \widehat{\overline{\mathbf{z}}} &= \overline{\mathbf{y}}, \\ \widehat{E}_q &= \mathrm{vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}, \end{aligned}$$

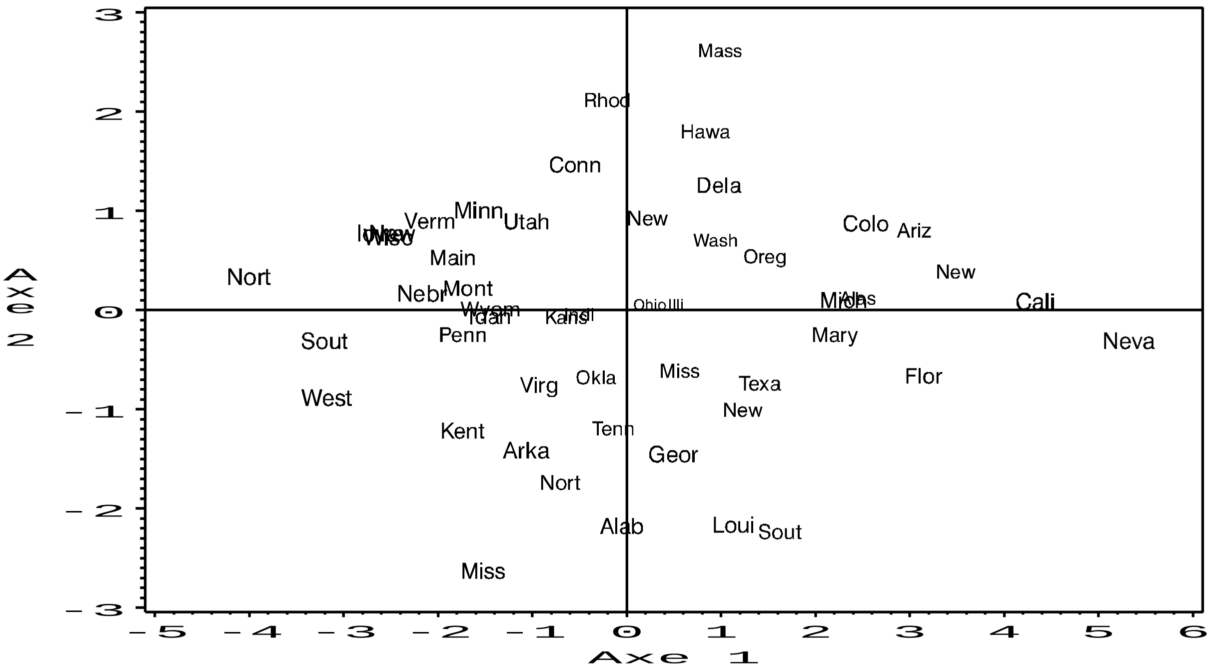

Figure 1: *Representation of statistical units.*

$$
\begin{aligned}
\widehat{\mathbf{P}}_q &= \mathbf{V}_q\mathbf{V}_q'\mathbf{M} \text{ is the } \mathbf{M}\text{-orthogonal} \\
&\quad \text{projection matrix onto } \widehat{E}_q, \\
\widehat{\mathbf{Z}}_q &= \mathbf{X}\widehat{\mathbf{P}}_q{}', \\
\hat{\mathbf{z}}_i &= \widehat{\mathbf{P}}_q\mathbf{x}_i + \overline{\mathbf{y}}.
\end{aligned}
$$

2.3 Graphical displays

All examples illustrating this paper are performed by studying the data (SAS, 1989) which give crime rates per 100,000 people in seven categories for each of the 50 Union states in 1977. All graphs are producted by SAS/GRAPH$^{(\mathrm{TM})}$ or SAS/INSIGHT, except figure 6 which is a S (Becker et al., 1988) result.

In a PCA, Graphical displays aim at a joint representation of units and variables. French practices have used to compute many tables for “helping” a PCA interpretation. They include squared cosines, contributions, correlations between factors and variables.... These become quite unreadable with a large number n of statistical units. This practice is somewhat contradictory with the aim of dimension reduction and thus of data compression. Some simple graphical features achieve the same purpose in a much more synthetic way. Figure 1 displays the statistical units whose coordinates are given by the principal

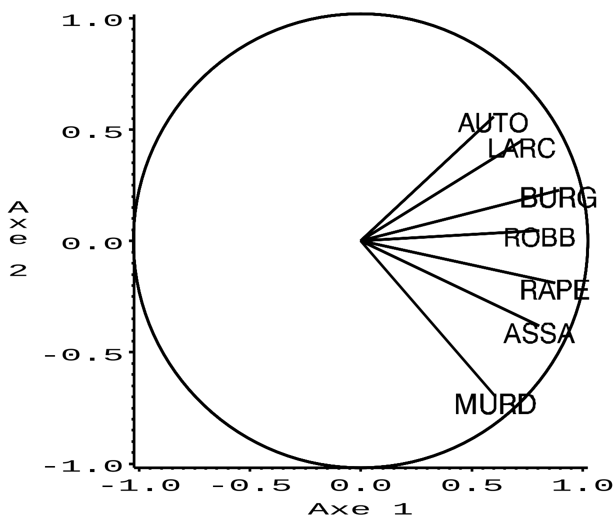

Figure 2: *Representation of variables.*

component matrix $\mathbf{C} = \mathbf{U}\mathbf{L}^{1/2}$. Figure 2 displays the standardized variables in the eigenvector basis. In these two scatter plots, the character size of each unit and of each variable depends on its representation quality namely the squared cosine of the angle between the vector and its projection on the plan. Furthermore, the vector length of a standardized variable is one and thus, the closer it is to the so-called "correlation" unit circle, the better is its representation. In figure 3, a parallel boxplot of principal components visualize their distribution. It is a simple way to point out possible outliers which have a major contribution to the direction of each eigenvector.

2.4 Optimal dimension

In this section, we set for the sake of clarity, $\mathbf{M} = \mathbf{\Psi} = \mathbf{I}_p$. In other cases of known error covariance matrix, raw data can be transformed by $\mathbf{\Psi}^{-1/2}$.

The quality of the estimates defined in section 2.2 directly depends on the choice of q; how many components need to be retained? Jolliffe (1986) reviews some rules frequently used to reach such a decision. He distinguishes between ad hoc rules that are intuitively plausible and rules based on more formal reasoning. Some of these rules further require strong distributional assumptions to test, for instance, the equality of last eigenvalues. Among ad hoc rules, it is usual to retain eigenvalues larger than the mean of all of them; it means larger than one in the case of PCA of standardized variables. An other rule consists to look for a possible shape of "knee" in the scree-graph (fig. 4) of eigenvalues and then to cut off just after it.

Furthermore, the introduction of a model in order to define PCA allows to consider

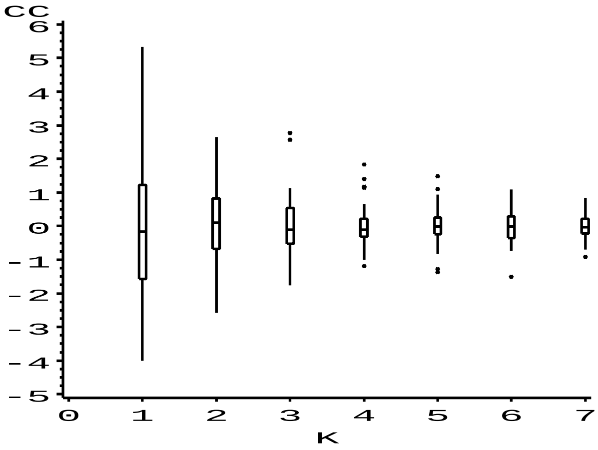

Figure 3: *Parallel boxplots of principal components.*

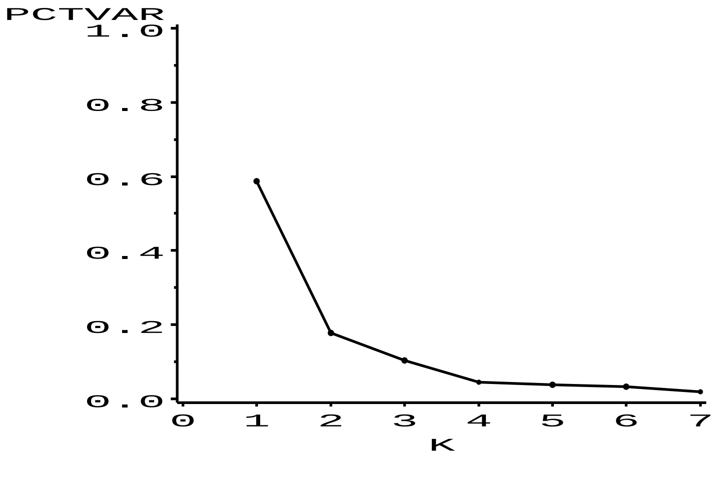

Figure 4: *Scree-graph of eigenvalues.*

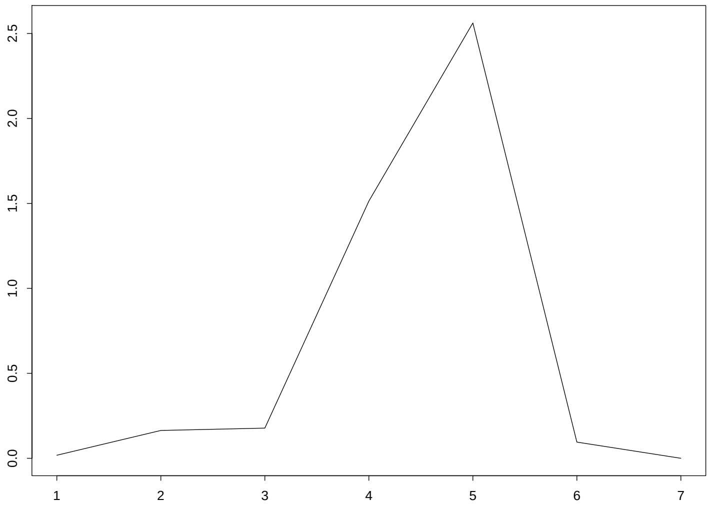

Figure 5: *Stability of the representation subspace versus its dimensionality q.*

another framework by defining a loss function

$$\mathcal{L}_q = \frac{1}{2}\left\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\right\|_2^2 = q - \mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q}. \tag{4}$$

It 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. This approach follows the principle that, in a PCA, the results are assumed to be reliable if the representation subspace is stable. A risk function is then defined by taking the expectation,

$$R_q = E(\mathcal{L}_q). \tag{5}$$

Besse and Falguerolles (1993) compare natural candidates to compute estimates for this risk function which are all computationally expensive such as the bootstrap and the jackknife. Furthermore Besse (1992) proposes an approximation of the jackknife which does not require much computational effort. If n is large enough, it is quite acceptable to write Taylor expansions of the eigenelements of $\mathbf{S}$ which lead to a Taylor expansion of the jackknife estimate

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

The analytic approximation of the jackknife estimate, is given by

$$\widehat{R_{Pq}} = \frac{1}{n-1}\sum_{k=1}^{q}\sum_{j=k+1}^{p}\frac{\frac{1}{n}\sum_{i=1}^{n} c_{ik}^2 c_{ij}^2}{(\lambda_j - \lambda_k)^2} \tag{6}$$

where c_{ij} denotes the general entry of the principal component matrix $\mathbf{C}$.

This shows 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 selected dimension and the first omitted one. It is very cheap and easy to plot $\widehat{R_{Pq}}$ versus q (fig. 5) and then to choose a convenient value for the dimensionality of the representation space $\widehat{E_q}$.

3 Optimal metric

Up until now, the choice of $\mathbf{M}$ seems rather arbitrary. However, under normality assumptions, it is easily seen that maximum likelihood estimation leads us to minimize (3) with, in this case, $\mathbf{M} = \mathbf{\Psi}^{-1}$. It is therefore natural to look for an optimal metric choice in the non-normal case.

Since PCA is defined in terms of model estimation, a quality criterion can be investigated to optimize this metric choice. Consider the general loss function

$$\mathcal{L}_q = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{z}_i - \widehat{\mathbf{z}}_i\|^2$$

whose the expectation defines a mean square error

$$R_q(\mathbf{M}) = E(\mathcal{L}_q)$$

which is to be optimized over $\mathbf{M}$ to derive an optimal metric. Assuming the variance error σ^2 to be small enough, Besse *et al.* (1988) proved that the leading term in an expansion of $R_q(\mathbf{M})$ is minimized for $\mathbf{M} = \mathbf{\Psi}^{-1}$.

This gives a property similar to Gauss Markov for PCA; the optimal metric is given by the inverse of error covariance matrix. Of course, this matrix is generally unknown but it can be estimated in the following cases.

3.1 PCA and discriminant analysis

Let us see factorial discriminant analysis (FDA) is an optimal PCA which displays clusters of units when groups are known. Data consist of p real random variables $Y^1, \ldots, Y^p$ and a *qualitative* variable T which are jointly observed on the same units. It is assumed that T has m categories $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$ and thus defines a partition $\{\Omega_k; k = 1, \ldots, m\}$ of the set Ω of statistical units ω_i. Let $\mathbf{T}$ be the $(n \times m)$ matrix of indicator functions,

$$t_i^k = t^k(\omega_i) = \begin{cases} 1 & \text{if } T(\omega_i) = \mathcal{T}_k, \\ 0 & \text{if not.} \end{cases}$$

It is natural to consider that all units, which hold the same kth category of variable T, are n_k repeated measures of a same fixed effect $\mathbf{z}_k$ weighted by $\overline{w_k} = \sum_{i \in \Omega_k} w_i$. We also define $\overline{\mathbf{D}} = \mathrm{diag}(\overline{w_1}, \ldots, \overline{w_m})$ and

$$\overline{\mathbf{Y}} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{Y} = \begin{bmatrix} \overline{y_1}' \\ \vdots \\ \overline{y_m}' \end{bmatrix} \text{ where } \overline{y_k} = \frac{1}{\overline{w_k}} \sum_{i \in \Omega_k} w_i y_i.$$

This leads us to consider the following model,

$$\begin{array}{l} \{\mathbf{y}_i \,; i = 1, \ldots, n\}, \ n \text{ independent random vectors of } E, \\ \forall k, \forall i \in \Omega_k, \mathbf{y}_i = \mathbf{z}_k + \mathbf{e}_i \text{ with } \left\{ \begin{array}{l} E(\mathbf{e}_i) = 0, \ \mathrm{var}(\mathbf{e}_i) = \boldsymbol{\Psi}, \\ \boldsymbol{\Psi} \text{ regular and unknown,} \end{array} \right. \\ \exists A_q, \ q\text{-dimensional affine subspace of } E \\ \text{such that } \forall k, \mathbf{z}_k \in A_q, (q < \min(p, m-1)). \end{array} \qquad (7)$$

As seen above, an optimal choice for $\mathbf{M}$ is $\boldsymbol{\Psi}^{-1}$. Since this matrix is unknown it has to be estimated. In this context $\boldsymbol{\Psi}$ may be estimated by the sample within group covariance matrix $\mathbf{W}$.

The least squares estimation of parameters A_q and $\mathbf{z}_k$ is given by PCA of $(\overline{\mathbf{Y}}, \mathbf{W}^{-1}, \overline{\mathbf{D}})$. It leads us to the computation of the same eigenvectors as PCA of $(\overline{\mathbf{Y}}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$ where $\mathbf{S}$ is the empirical covariance matrix. It is called the factorial discriminant analysis of $(\mathbf{Y}|\mathbf{T}, \mathbf{D})$.

3.2 Projection pursuit with PCA

In the case of unknown groups and when the within group covariance is "larger" than the between group covariance, a classical PCA, which focuses on maximal dispersion, can be very misleading. Exploratory projection pursuit (Friedman, 1987) may overcome such problems. It aims at producing interesting low-dimensional projections by optimizing criteria based on indices of non-normality (entropy) rather than the variance of the projected distributions. Usually, such methods require a very large amount of data which rapidly increases with p and q; this is the "curse of dimensionality". Moreover they are time consuming. In the same context, various kinds of suitably generalized PCA have been proposed by Caussinus and Ruiz (1990) in order to reveal data structures such as outliers or clusters at a very low computational cost. The basic idea is to draw from the data a convenient metric $\mathbf{M}$ on the space E of units. For instance, a robust estimate of the unknown within group covariance is given by

$$\widehat{\boldsymbol{\Psi}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k\left(\|\mathbf{y}_i - \mathbf{y}_j\|^2_{\mathbf{S}^{-1}}\right)(\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)'}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} k\left(\|\mathbf{y}_i - \mathbf{y}_j\|^2_{\mathbf{S}^{-1}}\right)};$$

where k is a decreasing positive real function (e.g. $k(t) = e^{-\gamma t}$ with $\gamma > 0$). A projection pursuit PCA, aiming at revealing clusters, is then defined as PCA of $(\mathbf{X}, \widehat{\boldsymbol{\Psi}}^{-1}, \frac{1}{n}\mathbf{I}_n)$. It gives similar results as FDA but without a prior knowledge of the classes.

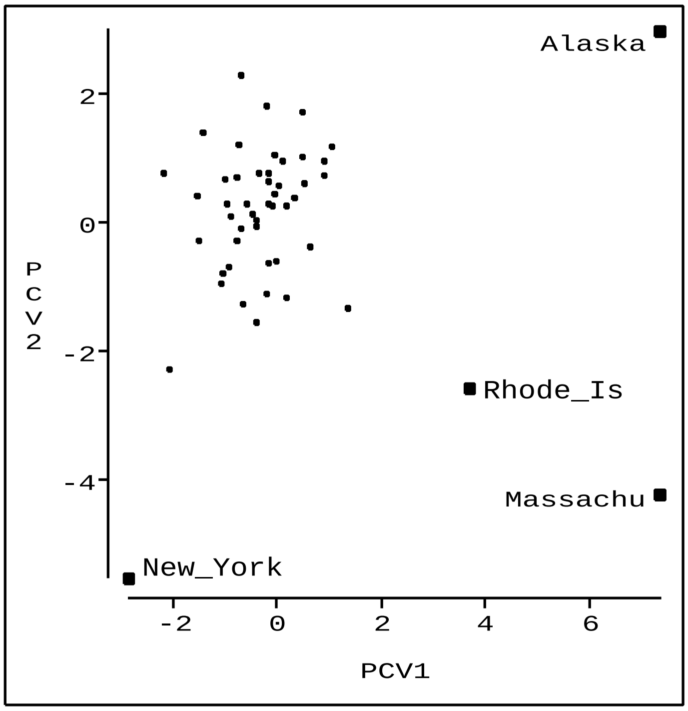

Figure 6: *Mapping of multivariate outliers.*

Another robust estimate of the unknown within group covariance aiming at pointing out multivariate outliers is defined by

$$\widehat{\mathbf{\Psi}} = \frac{\sum_{i=1}^{n-1} k\left(\|\mathbf{y}_i - \bar{\mathbf{y}}\|^2_{\mathbf{S}^{-1}}\right)(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'}{\sum_{i=1}^{n-1} k\left(\|\mathbf{y}_i - \bar{\mathbf{y}}\|^2_{\mathbf{S}^{-1}}\right)}.$$

The scatter plot of the states is then given by figure 6 when using this specific metric to compute PCA.

4 Functional PCA

4.1 Spline smoothing

In this section, we consider data generated from the observation of sampled curves. The relevant model is then

$$\mathbf{y}_i = \mathbf{z}_i + \mathbf{e}_i, \text{ with } z_i^j = f_i(t_j) \begin{cases} i = 1, \ldots, n \\ j = 1, \ldots, p \end{cases}$$

and where the vectors $\mathbf{y}_i$ are noisy observations of n functions f_i for

$$t_1 > \cdots > t_p; t_j \in T = [a, b] \subset \mathbb{R}.$$

Each observation is weighted by $w_i > 0$ with $\sum_{i=1}^n w_i = 1$. The error $\mathbf{e}_i$ is considered as a “white noise” with zero mean and whose covariance matrix is $\sigma^2\mathbf{I}$. The functions f_i are assumed to be "regular", it owns a second order derivative, and smooth. The smoothness is imposed by a constraint

$$\|f_i\|_2^2 = \int_T (f'')^2(t)dt < c, (c > 0). \tag{8}$$

In the case $n = 1$, this problem is classically solved by cubic spline smoothing or spline non-parametric regression (see Wahba (1990) for a review),

$$\widehat{f} = \arg \min_{f \in W^m} \left\{ \sum_{j=1}^{p} (f(t_j) - y^j)^2 + \rho \|f_i\|_2^2 \right\}; \tag{9}$$

ρ is the smoothing parameter depending on c. It controls the tradeoff between the roughness of the solution and the lack of fit to the data. The solution is obtained by a polynomial spline interpolation which exactly fits the values that are stored in the vector $\hat{\mathbf{z}} = (\widehat{f}(t_j) \; ; j = 1, \ldots, p)$ given by

$$\hat{\mathbf{z}} = \mathbf{A}(\rho)\mathbf{y} \text{ with } \mathbf{A}(\rho) = (\mathbf{I} + \rho\mathbf{N})^{-1}; \tag{10}$$

$\mathbf{A}$ is the so-called hat matrix and the matrix $\mathbf{N}$ depends on the sequence $\{t_1, \ldots, t_p\}$. It can be built by means of reproducing kernels (Wahba, 1990) or, computationally faster, by using tridiagonal matrices as described for instance by Buja *et al.* (1989).

For large n, we assume that a data set, consists of sampled curves, issued from noisy observations of n regular trajectories $f_i(t)$. The problem then is one of functional estimations of several curves, and means simultaneous non-parametric regressions.

4.2 Model and estimation

The nature of the data introduces two kinds of constraints in a fixed effect model. On one hand, a dimension constraint assumes that curves only span a subspace of a suitable functional space. In finite dimension, this means that sampled curves lie in a q dimensional subspace of $\mathbb{R}^p$. On the other hand, each real curve is assumed to be smooth enough and then they obey the constraint (8) with a common constant c.

The fixed effect model becomes

$$\begin{array}{l} \{\mathbf{y}_i \; ; i = 1, \ldots, n\}, \; n \text{ independent random vectors of } \mathbb{R}^p, \\ \mathbf{y}_i = \mathbf{z}_i + \mathbf{e}_i \; ; i = 1, \ldots, n \text{ with } \left\{ \begin{array}{l} E(\mathbf{e}_i) = 0, \; \mathrm{var}(\mathbf{e}_i) = \sigma^2\mathbf{I}, \\ \sigma > 0 \text{ unknown }, \end{array} \right. \\ \exists A_q, \; q \text{ dimensional affine subspace of } \mathbb{R}^p \text{ such that} \\ \forall i, \mathbf{z}_i \in A_q, (q < p), \|\mathbf{z}_i\|_{\mathbf{N}} < c, (c > 0). \end{array} \tag{11}$$

Besse (1994) proves that the least squares estimates are derived from those of the PCA of $\mathbf{XA}(\rho)^{1/2}$ that leads to simultaneously perform spline smoothing and a singular value decomposition. This produces a kind of smooth Karhunen-Loeve expansion of the observed process in that functional context.

4.3 Smoothing and dimension optimality

Both splines and SVD act as smoothing tools. The first is driven by a smoothing parameter ρ whose optimal value depends on data regularity. In contrast, SVD of longitudinal data gives approximations of trajectories whose smoothness usually depends on the number q of selected components; neglected components are often essentially noisy. These parameter values must be jointly optimized in order to obtain better fits of the original functions. We suggest a strategy based on the stability criterion which is described in section 2.4. PCA results are assumed to be reliable if the estimated subspace $\widehat{E_q}$ is stable with respect to data perturbations. It is then proposed to search for a dimension and a smoothing parameter which induce the greatest stability.

Optimality is achieved by minimizing $\widehat{R_{Pq}}$ on both ρ and q. Simulations (Besse, 1994) show that this is not too easy but this heuristic approach leads to fairly good choices of ρ and q values.

5 Curvilinear PCA

Different kinds of generalized additive models were reviewed by Hastie and Tibshirani (1990) in the framework of regression. In the context of PCA, different generalizations, which introduce non linear transformations for each variable, were previously developed. De Leeuw *et al.* (1981), De Leeuw (1982) define different kinds of "Non metric PCAs". Winsberg and Ramsay (1983) search optimal transformations for dimension reduction. A generalization of the fixed effect model is proposed in order to take into account these kinds of problems. This leads to a new method by dealing with spline smoothing.

5.1 Smooth transformation

Let Y be a real statistical variable valued within the range T; n realizations of this variable are collected in the vector $\mathbf{y}$ of $\mathbb{R}^n$. Several approaches, in different frameworks have been proposed to define transformations of a statistical variable as power functions, polynomial piecewise functions based on B-splines, with monotony constraint or not.... In this section, we look for smooth functions f to define data transformations

$$\tilde{\mathbf{y}} = f(\mathbf{y}) = [f(y_1), \ldots, f(y_n)]'.$$

These are built by using a cubic spline smoother in order to easily control the smoothness. There, the smoothing criterion becomes

$$\|f\|_2^2 = \tilde{\mathbf{y}}'\mathbf{N}\tilde{\mathbf{y}} = \|\tilde{\mathbf{y}}\|_{\mathbf{N}}^2$$

and $\mathbf{N}$ depends on $\mathbf{y}$. We denote by $(\mathcal{C})$ the following set of constraints,

$$(\mathcal{C}) \Longleftrightarrow \begin{cases} f \text{ is "regular"}, \\ \|f\|_2^2 \leq c, \\ \operatorname{var}(f(Y)) = 1. \end{cases} \tag{12}$$

5.2 The fixed effect curvilinear model

Consider now $\{\mathbf{y}_i;\ i = 1, \ldots, n\}$ n independent random variables in $\mathbb{R}^p$. Their realization is collected in a $(n \times p)$ matrix $\mathbf{Y}$ which is assumed, without loss of generality, to be column centered when each row is weighted by w_i.

This framework leads us to consider a set f of p smooth transformation functions f_j operating on the p statistical variables Y^j which are each valued in a range T_j. The columns $\mathbf{y}^j$ of the data matrix become

$$\widetilde{\mathbf{Y}} = f(\mathbf{Y}) = [f_1(\mathbf{y}^1), \ldots, f_p(\mathbf{y}^p)],$$

Each function f_j much obey $\mathcal{C}_j$. It belongs to a subspace of "regular" functions whose the semi-norm matrix is $\mathbf{N}_j$; the smoothing parameter is c_j; $\mathbf{I}$ denotes the identity matrix in $\mathbb{R}^p$.

The data matrix rows are assumed to be the observations of n random vectors $\mathbf{y}_i$ whose the transformations $\{f(\mathbf{y}_i);\ i = 1, \ldots, n\}$ have the same covariance matrix $\sigma^2\mathbf{I}$ but different expectations $\mathbf{z}_i$ belonging to a common q-dimensional $(q < p)$ subspace of $\mathbb{R}^p$.

To summarize, the curvilinear fixed effect model is defined by

$$\begin{array}{l} \{\mathbf{y}_i; i = 1, \ldots, n\}\ n \text{ independent random variables in } \mathbb{R}^p \\ f(\mathbf{y}_i) = \mathbf{z}_i + \mathbf{e}_i\ ; i = 1, \ldots, n \text{ with } E(\mathbf{e}_i) = 0,\ \operatorname{var}(\mathbf{e}_i) = \sigma^2\mathbf{I}, \\ \left.\begin{array}{l} \exists E_q, dim(E_q) = q \\ \exists f, f_j \text{ obeying } (\mathcal{C}_j) \end{array}\right\} \text{ such that } \forall i, \mathbf{z}_i \in E_q. \end{array} \tag{13}$$

Other sets of constraints than (12) lead to equivalent definitions of other methods.

- If f_j is assumed to be a step function depending on a sequence of knots belonging to T_j, it is the non-metric PCA as defined by De Leeuw (1982).
- If f_j is assumed to belong to a B-spline subspace of $W^m(T)$, it gives the approach by De Leeuw *et al.* (1981).

In this class of problem, the solution is achieved through an alternative least squares algorithm (Gifi, 1990). An implementation of the "classical" non metric PCA can be found in the SAS/STAT(TM) (1989) procedure `prinqual` in which the user must set several prior parameter values;

- number and positions of knots which define the B-spline basis for each variable transformation,
- shape and properties of transformations (degree of B or I-splines),
- the dimension q (note that solutions are not nested).

The introduction of a spline smoother rather than B or I-splines led Besse and Ferraty (1994) to propose a smoothing alternative least square algorithm (SALSA) to estimate the curvilinear fixed effect model. For this algorithm, only the dimension q and the p

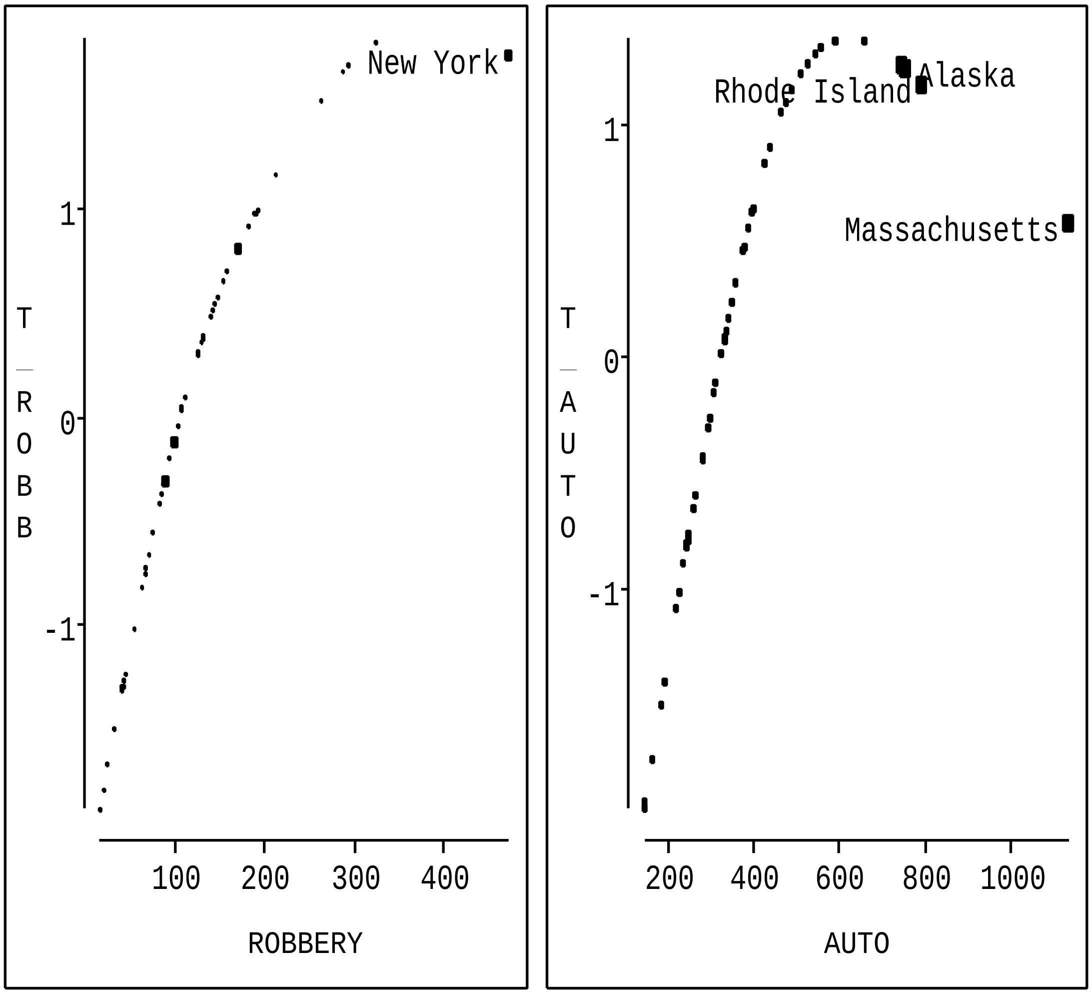

Figure 7: *Optimal transformations of* `robbery` *and* `auto-theft`.

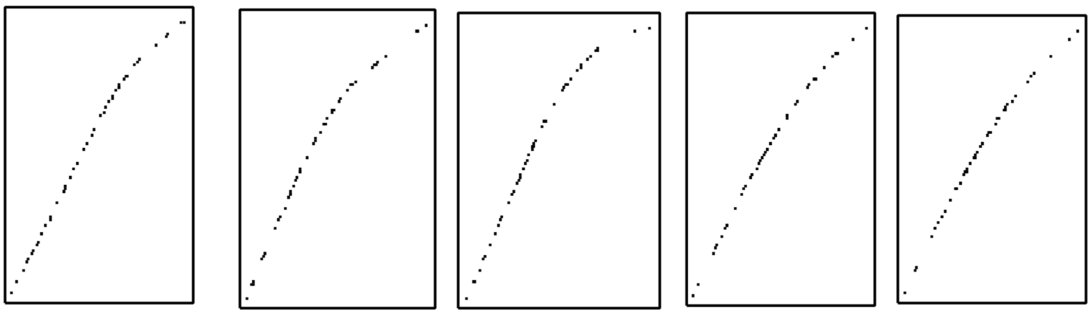

Figure 8: *Optimal transformations of* `murder`, `rape`, `assault`, `larceny` *and* `burglary`.

smoothing parameters must be set. They can be optimized through a bootstrap estimation of the above quadratic risk (4) concerning the representation space stability. This has been done for the data of crime rates whose figures 7 and 8 display the transformations which optimize the PCA. In this example, only two variables require a strong transformation to force a better fit in a two dimensional model. It acts as to make PCA robust by reducing some crime rates of "outliers" states.

References

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

Besse, P. (1992). Pca stability and choice of dimensionality. *Statistics & Probability Letters*, 13:405–410.

Besse, P. (1994). Simultaneous non-parametric regressions. *Submitted for publication.*

Besse, P., Caussinus, H., Ferré, L., and Fine, J. (1988). Principal component analysis and optimization of graphical displays. *Statistics*, 19:301–312.

Besse, P. and Falguerolles, A. d. (1993). Application of resampling methods to the choice of dimension in principal component analysis. In Härdle, W. and Simar, L., editors, *Computer Intensive Methods in Statistics*, pages 167–176. Physica-Verlag.

Besse, P., Falguerolles, A. d., and Fine, J. (1995). *A Modeling Approach to Principal Component Analysis*. Springer-Verlag.

Besse, P. and Ferraty, F. (1995). A curvilinear fixed effect model. *submitted for publication.*

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17:453–555.

Caussinus, H. (1986). Models and uses of principal component analysis. In de Leeuw, J. et al., editors, *Multidimensional Data Analysis*, pages 149–170. DSWO Press.

Caussinus, H. and Ruiz, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analysis. In XX, X., editor, *COMPSTAT 90*, pages 121–126. Physica-Verlag.

De Leeuw, J. (1982). Non-linear principal component analysis. In Caussinus, H. et al., editors, *Compstat 82, Proceedings in Comp. Stat.*, pages 77–85. Physica-Verlag.

De Leeuw, J., Rijckvorsel, J. v., and Wonder, H. (1981). Non-linear principal components analysis with b-spline. *Methods of Operations Research*, 33:379–393.

Friedman, J. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.

Gifi, A. (1990). *Non-linear Multivariate Analysis*. Wiley.

Gower, J. (1984). Multivariate analysis: ordination, multidimensional scaling and allied topics. In Lloyd, E., editor, *Handbook for Applicable Mathematics*, volume VI. Statistics, pages 727–781. Wiley.

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

SAS (1989). *SAS/STAT User's Guide*, fourth edition. version 6.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Winsberg, S. and Ramsay, J. (1983). Monotone spline transformations for dimension reduction. *Psychometrika*, 48:575–599.