Models for Multivariate Data Analysis

Philippe C. Besse

Laboratoire de Statistique et Probabilités, U.A. CNRS 745,
Université Paul Sabatier, 31062 Toulouse cedex, France

Abstract. This paper reviews some models for exploring multivariate data. If a fixed effect model is used to define a linear Principal Components Analysis (PCA), then risk functions can be defined and issues of metric and dimension optimality addressed. The model is then adapted to define a functional PCA which can be used to the study of smooth sampled curves. Finally, this model is generalised, giving a curvilinear PCA, which attempts to build smooth optimal transformations of data for the purpose of dimension reduction.

Keywords. principal component analysis, functional estimation, optimal transformations, spline smoothing

1 Introduction

A large number of statistical methods aim to analyze or to explore multivariate data. This paper is concerned with a class of them where the two main purposes are reduction of dimensionality and graphical presentation of data collected in a $(n \times p)$ matrix. Different frameworks have been proposed. They often lead to a common mathematical core as those 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 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, for example, the choice of metric, dimensionality, smoothing parameter.

This paper is mainly concerned with the analysis of quantitative data and focuses on recent developments of the Principal Component Analysis (PCA) and its extensions. It 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. The same kind of considerations occur when dealing with categorical data and contingency tables. Baccini et al. (1993) compare different models for Correspondence Analysis and their estimation with Association and Correlation models (Goodman, 1991 with discussion).

Section 2 defines the PCA as the estimation of a fixed effect model and discusses the optimal choices of dimensionality and metric by introducing quadratic risks in order to evaluate the estimation quality. Section 3 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 search for optimal transformations for dimension reduction. This leads to introduce a curvilinear fixed effect model including non-metric PCA's. A Smoothing Alternative Least Squares Algorithm (SALSA) is proposed to achieve the estimation. The lack of space, in this short review, does not allow to clarify all proofs or to give practical examples. These must be found in referenced papers or in a forthcoming book (Besse et al. 1995).

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

2.1 A 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; $\mathbf{M}$ is the metric which is introduced to compute distances between units or ultimately $\|\mathbf{z}_i - \mathbf{z}_{i'}\|_{\mathbf{M}}$.

To summarize,

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

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

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

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

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

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

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

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

$$
\begin{array}{rcl}
\widehat{\overline{\mathbf{z}}} & = & \overline{\mathbf{y}},\\
\widehat{E_q} & = & \mathrm{vect}\{\mathbf{v}^1,\ldots,\mathbf{v}^q\},\\
\widehat{\mathbf{P}_q} & = & \mathbf{V}_q\mathbf{V}'_q\mathbf{M} \text{ is the } \mathbf{M}\text{-orthogonal}\\
& & \text{projection matrix onto } \widehat{E_q},
\end{array}
$$

$$
\begin{aligned}
\widehat{\mathbf{Z}_q} &= \mathbf{X}\widehat{\mathbf{P}_q}', \\
\widehat{\mathbf{z}_i} &= \widehat{\mathbf{P}_q}\mathbf{x}_i + \overline{\mathbf{y}}.
\end{aligned}
$$

2.3 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 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 in order to optimize this metric choice. Consider the general loss function

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

where $\mathbf{A}$ denotes any Euclidean metric in $\mathbb{R}^p$.

The taking of the expectation defines a mean square error

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

which is to be optimized over $\mathbf{M}$ to derive an optimal metric.

Besse et al. (1988) proved, by means of Perturbation Theory, that if the variance error σ^2 is assumed to be small enough, $R_q(\mathbf{M}, \mathbf{A})$ can be expanded as

$$
R_q(\mathbf{M}, \mathbf{A}) = \sigma^2 \left[(q+1)\mathrm{tr}\mathbf{\Psi}\mathbf{A} + (n-q-1)\mathrm{tr}(\mathbf{P}_q\mathbf{\Psi}\mathbf{P}'_q\mathbf{A})\right] + 0(\sigma^3).
$$

Furthermore, for any metric $\mathbf{A}$, the leading term in $R_q(\mathbf{M}, \mathbf{A})$ 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 under some circumstances as is the case in the next section.

2.3.1 PCA and Discriminant Analysis

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 k^{th} 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 $\overline{\mathbf{D}} = \text{diag}(\overline{w_1}, \ldots, \overline{w_m})$ 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 to the following model,

$$\begin{array}{l} \{\mathbf{y}_i \ ; i = 1, \ldots, n\}, \ n \text{ independent random vectors of } E, \\ \forall k, \forall i \in \Omega_k, \mathbf{y}_i = \mathbf{z}_k + \mathbf{e}_i \text{ with } \left\{ \begin{array}{l} E(\mathbf{e}_i) = 0, \ \text{var}(\mathbf{e}_i) = \mathbf{\Psi}, \\ \mathbf{\Psi} \text{ regular and unknown,} \end{array} \right. \\ \exists A_q, \ q\text{-dimensional affine subspace of } E \\ \text{such that } \forall k, \mathbf{z}_k \in A_q, (q < \min(p, m-1)). \end{array} \tag{4}$$

Let $\bar{\mathbf{z}} = \sum_{k=1}^{m} \overline{w_k}\mathbf{z}_k$. The model implies $\bar{\mathbf{z}} \in A_q$. Let 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 $\{\mathbf{z}_k \ ; k = 1, \ldots, m\}$; $\overline{w_k}$ are nuisance parameters that we will not consider now.

A least squares estimation for this problem can be formulated as follows.

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

We obtain

$$\sum_{k=1}^{m} \sum_{i \in \Omega_k} w_i \, \|\mathbf{y}_i - \mathbf{z}_k\|^2_{\mathbf{M}} = \sum_{k=1}^{m} \sum_{i \in \Omega_k} w_i \, \|\mathbf{y}_i - \overline{\mathbf{y}_k}\|^2_{\mathbf{M}} + \sum_{k=1}^{m} \overline{w_k} \, \|\overline{\mathbf{y}_k} - \mathbf{z}_k\|^2_{\mathbf{M}}$$

and we are led to minimize

$$\min_{E_q, \mathbf{z}_k} \left\{ \sum_{k=1}^{m} \overline{w_k} \, \|\overline{\mathbf{y}_k} - \mathbf{z}_k\|^2_{\mathbf{M}} \ ; \ \dim(E_q) = q, \mathbf{z}_k - \bar{\mathbf{z}} \in E_q \right\}.$$

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

The estimation of parameters E_q and $\mathbf{z}_k$ of the model (4) is given by the PCA of $(\overline{\mathbf{Y}}, \mathbf{W}^{-1}, \overline{\mathbf{D}})$. It leads to the computation of the same eigenvectors as the PCA of $(\overline{\mathbf{Y}}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$. It is called the Factorial Discriminant Analysis of $(\mathbf{Y}|\mathbf{T}, \mathbf{D})$.

2.3.2 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

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

where k is a decreasing positive real function (e.g. $k(t) = e^{-\gamma t}$ with $\gamma > 0$). A projection pursuit PCA, aiming to reveal clusters, is then defined as the PCA of $(\mathbf{X}, \widehat{\mathbf{\Psi}}^{-1}, \frac{1}{n}\mathbf{I}_n)$. Another robust estimate aiming to reveal multivariate outliers is defined by

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

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 2.3 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 least eigenvalues. The introduction of a model in order to define PCA allows to consider another framework for discussing this problem.

An optimal dimension could also be searched for by minimizing the risk function as introduced in Section 2. Unfortunately this approach raises a problem of indeterminacy if σ^2 is unknown; it is not possible to estimate simultaneously σ and the optimal dimension q, since the natural estimate for σ^2 is the average of the $(p-q)$ smallest eigenvalues.

This drawback can be overcome by considering the subspace $\widehat{E_q}$, which is an asymptotically consistent estimate (Fine and Pousse, 1992) rather than the vectors z_i. The loss function then becomes

$$\mathcal{L}_q = \frac{1}{2}\left\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\right\|_2^2 = q - \mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q}. \tag{5}$$

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. In that context, $\mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q}$ is also the sum of the squared canonical correlation coefficients between the component sets which respectively span E_q and $\widehat{E_q}$. 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,

$$R_q = E(\mathcal{L}_q). \tag{6}$$

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 to compute estimates for this risk function. Besse and Falguerolles (1993) compare these methods which are all computationally expensive. 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 consider that any row elimination introduces only a small perturbation in further computations. Perturbation theory enables us to write Taylor expansions of the eigenelements of $\mathbf{S}$ which lead to a Taylor expansion of the jackknife estimate and then to an analytic approximation

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

This analytic approximation of the jackknife estimate, is given by

$$\widehat{R_{Pq}} = \frac{1}{n-1}\sum_{k=1}^{q}\sum_{j=k+1}^{p}\frac{\tau_{jk}}{(\lambda_j-\lambda_k)^2} \text{ with } \tau_{jk} = \frac{1}{n}\sum_{i=1}^{n} c_{ik}^2 c_{ij}^2 \tag{7}$$

where c_{ij} denotes the general entry of the principal component matrix

$$\mathbf{C} = \mathbf{U}\mathbf{L}^{1/2}.$$

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. This is consistent 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 $\mathbf{v}^q$ and $\mathbf{v}^{q+1}$. It is very cheap and easy to plot $\widehat{R_{Pq}}$ versus q and then to choose a convenient value for the dimensionality of the representation space $\widehat{E_q}$.

3 Functional PCA

3.1 Spline regression

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

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

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 and smooth. One way to ensure the regularity of f is to impose a condition such as the existence of $(m-1)$ absolutely continuous derivatives and a square integrable mth integrable derivative, $(m \geq 1)$. This means that each f_i belongs to a Sobolev space $W^m(T)$. Furthermore, the smoothness is imposed by a constraint

$$\|f_i\|_m^2 = \int_T (f^{(m)})^2(t) dt < c, (c > 0), \tag{8}$$

where $\|.\|_m$ defines a semi-norm on $W^m(T)$. One could generalize this to any equivalent semi-norm of $W^m(T)$.

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

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

ρ is the smoothing parameter; it is also the Lagrange multiplier 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. The symmetric semi-definite and positive matrix $\mathbf{N}$ is the representation in the canonical basis of the semi-norm which is induced by $\|.\|_m$ on a p-dimensional spline subspace of W^m. With matrix notation, it simply amounts to

$$\left\|\widehat{f}\right\|_m^2 = \widehat{\mathbf{z}}'\mathbf{N}\widehat{\mathbf{z}} = \|\widehat{\mathbf{z}}\|_{\mathbf{N}}^2 .$$

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. This could be achieved by computing n classical non-parametric (kernel or spline) regressions where each smoothing parameter is independently optimized by Generalized Cross Validation (GCV). But this approach does not take any special features of the data into account and simulations show (Besse, 1994) that it induces a wide information loss. This kind of data has been previously studied by Besse and Ramsay (1986), Ramsay and Dalzell (1991), Besse and Pousse (1992) who deal with Principal Components Analysis of curves. These are closely related to works by Kneip and Gasser (1992) and to other papers which aim at approximating a covariance function by smooth eigenfunctions (Rice and Silverman, 1991, Pezzuli and Silverman, 1993).

3.2 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 $W^m(T)$. 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

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

As above, the smoothness constraint is taken into account by introducing the Lagrange multiplier $\rho, (\rho > 0)$ in the least squares criterion; it leads to consider the following problem

$$\min_{\mathbf{z}_i, A_q} \left\{ \sum_{i=1}^{n} w_i \left(\|\mathbf{y}_i - \mathbf{z}_i\|_{\mathbf{I}}^2 + \rho \|\mathbf{z}_i\|_{\mathbf{N}}^2 \right) ; \forall i, \mathbf{z}_i \in A_q,\ \dim(A_q) = q \right\}. \tag{12}$$

Let $\bar{\mathbf{z}} = \sum_{i=1}^n w_i \mathbf{z}_i$, $\bar{\mathbf{y}} = \sum_{i=1}^n w_i \mathbf{y}_i$ and $\mathbf{X}$ $(n \times p)$ be the column centered matrix with rows $(\mathbf{y}_i - \bar{\mathbf{y}})'$; $\mathbf{Z}$ $(n \times p)$ denotes the matrix with rows $(\mathbf{z}_i - \bar{\mathbf{z}})'$. It follows that the vectors $\mathbf{z}_i - \bar{\mathbf{z}}$ belong to the linear subspace E_q $(A_q = \bar{\mathbf{z}} + E_q)$ and thus the rank of $\mathbf{Z}$ is q.

Besse (1994), in parallel with Denby and Mallows (1993) proves that the solution of (12) is given by

$$\widehat{\mathbf{Z}_q} = \mathbf{U}_q \mathbf{L}_q^{1/2} \mathbf{V}_q' \mathbf{A}(\rho)^{1/2}, \tag{13}$$

$$\widehat{\mathbf{z}_i} = \left[\widehat{\mathbf{Z}_q}\right]_i' + \mathbf{A}(\rho)\bar{\mathbf{y}}; \tag{14}$$

where matrices $\mathbf{U}_q(n \times q), \mathbf{L}_q(q \times q)$ and $\mathbf{V}_q(p \times q)$ define the first q leading terms of the generalized SVD of $(\mathbf{XA}(\rho)^{1/2}, \mathbf{I}, \mathbf{D})$; $\mathbf{A}(\rho)$ denotes the hat matrix. $\mathbf{V}$ contains the eigenvectors of $\mathbf{A}(\rho)^{1/2}\mathbf{SA}(\rho)^{1/2}$ where $\mathbf{S} = \mathbf{X}'\mathbf{DX}$ denotes the empirical covariance matrix. They are stored in the decreasing order of their associated eigenvalues. $\widehat{E_q}$ is spanned by the first q vectors $\widetilde{\mathbf{v}}^k$ of $\widetilde{\mathbf{V}}$.

As in the case $n = 1$, the estimates of functions f are then obtained by the spline interpolation functions $\widehat{f}_i \in W^m$ that exactly fit the values of the vectors

$$\widehat{\mathbf{z}_i} = (\widehat{f}_i(t_j); j = 1, \ldots, p)$$

and which minimize $\left\| \widehat{f}_i \right\|_m$. The estimates are derived from 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.

3.3 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. GCV is commonly used to optimize the spline smoothing parameter. Unfortunately, its application to the dimension choice in PCA (Krzanowski, 1987) is not convincing (Besse and Ferré, 1993) and never used in practice.

Besse and Pousse (1992) suggest another 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 easy but this heuristic approach leads to fairly good working choices of ρ and q values. An asymptotic study still needs to be done.

4 Curvilinear PCA

Different kinds of generalized additive models were studied 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), Rijckvorsel (1988), Gifi (1990) define different kinds of "Non metric PCAs". Winsberg and Ramsay (1983), Ramsay (1988) search optimal transformations for dimension reduction. A generalization of the fixed effect model is proposed in order to take into account these kind of problems. This leads to a new method of dealing with spline smoothing.

4.1 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 functions f belonging to a Sobolev space $W^m(T)$ to define data transformations

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

These are built by using a spline smoother in order to easily control the smoothness. The spline smoothing principle is briefly recalled in section 3.1. There, the smoothing criterion becomes

$$\|f\|^2_m = \widetilde{\mathbf{y}}'\mathbf{N}\widetilde{\mathbf{y}} = \|\widetilde{\mathbf{y}}\|^2_{\mathbf{N}}$$

and $\mathbf{N}$ depends on $\mathbf{y}$ and m.

We denote by $(\mathcal{C})$ the following set of constraints,

$$(\mathcal{C}) \Longleftrightarrow \left\{ \begin{array}{c} f \in W^m(T), \\ \|f\|^2_m \le c, \\ \mathrm{var}(f(Y)) = 1. \end{array} \right. \tag{15}$$

4.2 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 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 $W^m(T_j)$ 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 $\{\mathbf{y}_i;\ i = 1, \ldots, n\}$ 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

$$\begin{array}{l} \{\mathbf{y}_i; i = 1, \ldots, n\}\ n \text{ independent random variables in } \mathbb{R}^p \\ f(\mathbf{y}_i) = \mathbf{z}_i + \mathbf{e}_i\ ; i = 1, \ldots, n \text{ with } E(\mathbf{e}_i) = 0,\ \mathrm{var}(\mathbf{e}_i) = \sigma^2\mathbf{I}, \\ \left.\begin{array}{l} \exists E_q, dim(E_q) = q \\ \exists f, f_j \text{ obeying } (\mathcal{C}_j) \end{array}\right\} \text{ such that } \forall i, \mathbf{z}_i \in E_q. \end{array} \quad (16)$$

4.3 Estimation

$\mathbf{Z}$ denotes the $(n \times p)$ matrix whose rows are the $\mathbf{z}_i$ and columns the $\mathbf{z}^j$. A least squares estimation of the transformations and the fixed effects leads to the solution of the following problem

$$\min_{\mathbf{z}^j, f^j} \left\{ \sum_{j=1}^{p} \left\| f^j(\mathbf{y}^j) - \mathbf{z}^j \right\|_{\mathbf{D}}^2\ ; f_j \text{ holding } (\mathcal{C}_j) \text{ and } \mathrm{rank}(\mathbf{Z}) = q \right\}. \quad (17)$$

Other sets of constraints than (15) 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).

With matrix notation and when denoting the transformed data columns by $\widetilde{\mathbf{y}}^j$, Besse and Ferraty (1994) prove that problem (17) is equivalent to

$$\min_{\mathbf{Q}_q, \widetilde{\mathbf{Y}}} \left\{ \left\| \widetilde{\mathbf{Y}} - \mathbf{Q}_q \widetilde{\mathbf{Y}} \right\|_{\mathbf{I},\mathbf{D}}^2 + \sum_{j=1}^{p} \rho_j \left\| \widetilde{\mathbf{y}}^j \right\|_{\mathbf{N_j}}^2\ ; \|\widetilde{\mathbf{y}}_j\|_{\mathbf{D}} = 1 \right\} \quad (18)$$

where $\mathbf{Q}_q$ is a rank q $\mathbf{D}$-orthogonal projection matrix.

4.4 Algorithm

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 a SAS (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 new algorithm to estimate the curvilinear fixed effect model. For this algorithm, only the dimension q and the p smoothing parameters must be set.

SALSA Smoothing Alternative Least Square Algorithm

1. *Let* $\widetilde{\mathbf{Y}}_{(0)} = \left[\mathbf{y}^1|\cdots|\mathbf{y}^p\right]$ *be the original data set. A matrix* $\mathbf{N}_j$ *is associated to each vector* $\widetilde{\mathbf{y}}^j$.
2. k^{th} *iteration*
   - *Set* $\widetilde{\mathbf{Y}}_{(k)}$ *and solve (18) by the first* q *terms of the singular value decomposition of* $\widetilde{\mathbf{Y}}_{(k)}$*;*

$$\widehat{\widetilde{\mathbf{Y}}}^{q}_{(k)} = \mathbf{U}_q\mathbf{L}_q^{1/2}\mathbf{V}'_q \text{ and } \mathbf{Q}_{(k)} = \mathbf{U}_q\mathbf{U}'_q\mathbf{D}.$$

   - *Set* $\mathbf{Q}_{(k)}$ *and solve for* $j = 1, \ldots, p$

$$\min_{\widetilde{\mathbf{y}}} \left\{ \left\|\mathbf{Q}_{(k)}\widetilde{\mathbf{y}} - \widetilde{\mathbf{y}}\right\|^2_{\mathbf{D}} + \rho_j \left\|\widetilde{\mathbf{y}}\right\|^2_{\mathbf{N_j}} ; \left\|\widetilde{\mathbf{y}}_j\right\|_{\mathbf{D}} = 1 \right\}. \qquad (19)$$

   The solution $\widetilde{\mathbf{y}}^j_{(k+1)}$ *is the eigenvector of* $\mathbf{D}\left(\mathbf{I} - \mathbf{Q}_{(k)}\right) + \rho_j\mathbf{N}_j$ *corresponding to the smallest eigenvalue.*

As the least squares criterion is positive and monotonously decreasing, SALSA reaches a minimum but this could be a local one. When a minimum is reached, the last SVD provides a rank q estimate of the fixed effect and also an estimate of the graphical representation subspace $\widehat{E_q}$. It is spanned by the first q eigenvectors $\mathbf{v}^1, \ldots, \mathbf{v}^q$ of the correlation matrix of the transformed data $\mathbf{R} = \widetilde{\mathbf{Y}}'\mathbf{D}\widetilde{\mathbf{Y}}$.

References

BACCINI, A., CAUSSINUS, H., and DE FALGUEROLLES, A. (1993). Analyzing Dependence in large Contingency Tables: Dimensionality and Patterns in Scatter-plots, in *Multivariate analysis: future directions 2*, Cuadras, C.M. and Rao, C.R. eds., Elsivier Science Publishers.

BESSE, P.C. (1992). PCA Stability and Choice of Dimensionality, *Statistics & Probability Letters*, North-Holland **13**, 405-410.

BESSE, P.C. (1994). Simultaneous Non-Parametric Regressions, submitted for publication.

BESSE, P.C., CAUSSINUS, H., FERRÉ, L. and FINE, J. (1988). Principal Components Analysis and Optimization of Graphical Displays. *Statistics*, **19**, 301-312.

BESSE, P.C. and DE FALGUEROLLES, A. (1993). Application of Resampling Methods to the choice of Dimension in Principal Components Analysis, in *Computer Intensive Methods in Statistics*, Härdle W. and Simar L. eds., Physica-Verlag.

BESSE, P.C., DE FALGUEROLLES, A. and FINE, J. (1995). *A Modeling Approach to Principal Component Analysis*, to appear, Springer-Verlag.

BESSE, P.C. and FERRATY, F. (1994). A Fixed Effect Curvilinear Model, submitted for publication.

BESSE, P.C. and FERRÉ, L. (1993). Sur l'usage de la Validation Croisée en Analyse en Composantes Principales, *Rev. Statistique Appliquée*, **XLI**(1), 71-76.

BESSE, P.C. and POUSSE, A. (1992). Extension des Analyses Factorielles, in *Modèles pour l'Analyse des Données Multidimensionnelles*, J.J. Droesbeke et al. eds., Economica, 129-158.

BESSE, P.C. and RAMSAY, J.O. (1986). Principal Component Analysis of Sampled Curves, *Psychometrika*, **51**, 285-311.

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 *Multidimensional Data Analysis*, J. de Leeuw et al. eds., DSWO Press, Leiden, 149-170.

CAUSSINUS, H. and RUIZ, A. (1990). Interesting Projections of Multidimensional Data by Means of Generalized Principal Components Analysis, *COMPSTAT 90*, Physica-Verlag, Heidelberg, 121-126.

DENBY, L. and MALLOWS, C.L. (1993). Smooth Reduced-Rank Approximations, *I.S.I.*, 49th session, contributed papers, book 1, 355-356.

FINE, J. and POUSSE, A. (1992). Asymptotic study of the Multivariate Functional Model; application to the Metric choice in PCA, *Statistics*, **23**, 63-83.

FRIEDMAN, J.H. (1987). Exploratory Projection Pursuit, *J. Amer. Statist. Assoc.*, **82**, 249-266.

GIFI, A. (1990). *Non Linear Multivariate Analysis*, Wiley & Sons.

GOODMAN, L.A. (1991). Measures, Models, and Graphical Displays in the Analysis of Cross-classified data, with discussion, Journal of the American Statistical Society, **86**, 1085-1123.

GOWER, J.C. (1984). Multivariate Analysis: Ordination, Multidimensional Scaling and Allied Topics. in *Handbook for Applicable Mathematics*: vol. VI. Statistics, (ed. E.H., Lloyd), 727-781, Wiley.

HASTIE, T.J. and TIBSHIRANI, R.J. (1990). *Generalized Additive Models*, Chapman and Hall.

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

KNEIP, A. and GASSER, T. (1992). Statistical Tools to Analyse Data Representing a Sample of Curves, *The Annals of Statistics*, **20**, 1266-1305.

KRZANOWSKI, W.J. (1987). Cross Validation Choice in Principal Components Analysis, *Biometrics*, **43**, 575-584.

DE LEEUW, J. (1982). Non-linear Principal Component Analysis, COMPSTAT 82, H. Caussinus et al. eds., Physica-Verlag, 77-85.

DE LEEUW, J., VAN RIJCKVORSEL, J.L.A. and WONDER H. (1981). Non-linear Principal Component Analysis with B-spline, *Methods of Operations Research* **33**, 379-393.

PEZZULLI, S. and SILVERMAN, B.W. (1993). On Smoothed Principal Components Analysis, *Computational Statistics*, **8**, 1-16.

RAMSAY, J. (1988). Monotone Regression Splines in Action, *Statistical Sciences* (With Discussion) **3**, 425-461.

RAMSAY, J.O. and DALZELL, C.J. (1991). Some Tools for Functional Data Analysis (with discussion), *J.R. Statist. Soc. B*, **53**, 539-572.

RICE, J.A. and SILVERMAN, B.W. (1991). Estimating the Mean and Covariance Structure Nonparametrically when the Data are Curves, *J. R. Statist. Soc. B*, **53**, 233-243.

RIJCKVORSEL, J. (1988). Fuzzy Coding and B-spline, in *Component and Correspondence Analysis*, J.L.A. van Rijckvorsel and J. de Leeuw eds., Wiley & Sons.

SAS INSTITUTE INC. (1989). *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*, Cary, NC : SAS Institute Inc., 846 pp.

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

WINSBERG, S. and RAMSAY J.O. (1983). Monotone Spline Transformations for Dimension Reduction, *Psychometrika* **48**, 575-599.