Probabilistic Auto-Associative Models and Semi-Linear PCA

Serge Iovleff

Noname manuscript No.
(will be inserted by the editor)

Probabilistic Auto-Associative Models and Semi-Linear PCA

Serge Iovleff

Received: date / Accepted: date

Abstract Auto-Associative models cover a large class of methods used in data analysis, among them are for example the famous PCA and the auto-associative neural networks. In this paper, we describe the general properties of these models when the projection component is linear and we propose and test an easy to implement Probabilistic Semi-Linear Auto-Associative model in a Gaussian setting. We show that it is a generalization of the PCA model to the semi-linear case. Numerical experiments on simulated datasets and a real astronomical application highlight the interest of this approach.

Keywords Auto-Associative Models · Non-Linear PCA · Probabilistic Non-Linear PCA

1 Introduction

Principal component analysis (PCA) Pearson (1901); Hotelling (1933); Jolliffe (1986) is a well established tool for dimension reduction in multivariate data analysis. It benefits from a simple geometrical interpretation. Given a set of n points $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ with $\mathbf{y}_i \in \mathbb{R}^p$ and an integer $0 \leq d \leq p$, PCA builds the d-dimensional affine subspace minimizing the Euclidean distance to the scatter-plot Pearson (1901). The application of principal component analysis postulates implicitly some form of linearity. More precisely, one assumes that the data cloud is directed, and that the data points can be well approximated by their projections to the affine hyperplane corresponding to the first d principal components.

Starting from this point of view, many authors have proposed nonlinear extensions of this technique. Principal curves or principal surfaces methods Hastie and Stuetzle

S. Iovleff
Laboratoire Paul Painlevé – Université Lille 1 & CNRS
Modal Team – Inria Lille - Nord Europe
59655 Villeneuve d'Ascq Cedex – France
E-mail: serge.iovleff@inria.fr

(1989); Hastie et al (2001); Delicado (2001) as well as non-linear transformation of the original data set Durand (1993); Besse and Ferraty (1995) belong to this family of approaches. The auto-associative neural networks can also be viewed as a non-linear PCA model Baldi and Hornik (1989); Lu and Pandolfo1 (2010); Bishop (2006); Hinton et al (1997). In Girard and Iovleff (2005) we propose the auto-associative models (AAM) as candidates to the generalization of PCA using a projection pursuit regression algorithm Friedman and Stuetzle (1981); Klinke and Grassmann (2000) adapted to the auto-associative case. A common point of these approaches is that they have the intent to estimate an auto-associative model whose definition is given hereafter.

Definition 1 A function g is an auto-associative function of dimension d if it is a map from $\mathbb{R}^p$ to $\mathbb{R}^p$ that can be written as $g = R \circ P$ where P (the "Projection") is a map from $\mathbb{R}^p$ to $\mathbb{R}^d$ (generally $d < p$) and R (the "Restoration" or the "Regression") is a map from $\mathbb{R}^d$ to $\mathbb{R}^p$.

An auto-associative model (AAM) of dimension d is a manifold $\mathscr{M}_g$ of the form

$$\mathscr{M}_g = \{\mathbf{y} \in \mathbb{R}^p, \mathbf{y} - g(\mathbf{y}) = 0\}$$

where g is an auto-associative function of dimension d.

For example PCA constructs an auto-associative model using as auto-associative function an orthogonal projector on an affine subspace of dimension d. More precisely we have

$$g(\mathbf{y}) = \mathbf{m} + \sum_{i=1}^{d} \left\langle \mathbf{a}^i, \mathbf{y} - \mathbf{m} \right\rangle \mathbf{a}^i, \quad \mathbf{y} \in \mathbb{R}^p$$

$\mathbf{y}, \mathbf{m}, \mathbf{a}_i \in \mathbb{R}^p$. $\mathbf{m}$ is a position parameter and the vectors $\mathbf{a}_i$ are chosen in order to maximize the projected variance. g can be written $g = R \circ P$ with

$$P(\mathbf{y}) = \left(\left\langle \mathbf{a}^1, \mathbf{y} - \mathbf{m} \right\rangle, \ldots, \left\langle \mathbf{a}^d, \mathbf{y} - \mathbf{m} \right\rangle \right)$$

and

$$R(\mathbf{x}) = \mathbf{m} + x_1 \mathbf{a}^1 + \ldots + x_d \mathbf{a}^d$$

with $\mathbf{x} = (x_1, \ldots, x_d)'$. The AAM is then the affine subspace given by the following equation

$$\mathscr{M}_g = \left\{ \mathbf{y} \in \mathbb{R}^p; \mathbf{y} - \mathbf{m} - \sum_{i=1}^{d} \left\langle \mathbf{a}^i, \mathbf{y} - \mathbf{m} \right\rangle \mathbf{a}^i = 0 \right\}$$

Interested reader can check that principal curves, principal surfaces, auto-associative neural networks, kernel PCA Schölkopf et al (1999), ISOMAP Tenenbaum (2000) and local linear embedding Roweis and Saul (2000) have also the intent to estimate an AAM.

In the PCA approach the projection and the restoration function are both linear. It is thus natural to say that the PCA model is a *Linear Auto-Associative Model*. In the general case, the manifold $\mathscr{M}_g$ set can be empty (i.e. the auto-associative function g has no fixed point) or very complicated to describe. Our aim in this paper is to study from a theoretical and practical point of view the properties of some Auto-Associative

models in an intermediary situation between the PCA model and the general case: we will assume that the projection function is linear and let the regression function be arbitrary. We call the resulting AAM the *Semi-Linear Auto-Associative Models* (SLAAM).

Having restricted our study to the SLAAM, we have to give us some criterion to maximize. As we said previously, PCA tries to maximize the projected variance or, equivalently, to minimize the residual variance. Common AAM approaches also used the squared reconstruction error as criterion, or more recently a penalized criterion Hastie et al (2001). However as pointed out by M. E. Tipping and C. M. Bishop Tipping and Bishop (1999), one limiting disadvantage of this approach is the absence of a probability density model and associated likelihood measure. The presence of a probabilistic model is desirable as

- the definition of a likelihood measure permits comparison between concurrent models and facilitates statistical testing,
- A single AAM may be extended to a mixture of such models,
- if a probabilistic AAM is used to model the class conditional densities in a classification problem, the posterior probabilities of class membership may be computed.

We propose thus a Gaussian generative model for the SLAAM and try to estimate it using a maximum likelihood approach. In the general case we are faced with a difficult optimization problem and we cannot go further without additional assumptions. It will appear clearly that if the projection P, identified hereafter with a matrix $\mathbf{P}$, is *known* then the estimation problem of a SLAAM is very close to an estimation problem in a regression context. However, there are some differences we will underline. In particular it will appear that in order to get tractable maximum-likelihood estimates, we have to impose some restrictions to the noise. We call the resulting model of all these assumptions/simplifications a Semi-Linear Principal Component Analysis. It does not seem possible to add non-linearity to the PCA model and get tractable likelihood estimate for $\mathbf{P}$. But clearly, the assumption that $\mathbf{P}$ is known is too strong in practice. We thus propose to estimate it in a separate step using either PCA or a contiguity analysis Lebart (2000). Finally, even if $\mathbf{P}$ is assumed to be known, it remains to estimate the regression function R which is a non-linear function from $\mathbb{R}^d$ to $\mathbb{R}^p$. If $d > 1$ and p is moderately high the task becomes very complicated. Thus the implementation of the model proposed in this paper simplifies it once more. In this implementation, we assume that R is additive inspired by the Generalized Additive Model (GAM) approach Hastie and Tibshirani (1990).

In view of the experiments we have performed and we present here, it seems we obtain a practical and simple model which generalizes the PCA model to the non-linear case in an understandable way. The main advantages of the proposed model are:

1. the projection step being linear, it is easy to interpret the projected values $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ in term of the original variables $\mathbf{Y}$ and to represent them in a (kind of) bi-plot graph (see Gower and Hand (1995)),
2. the regression functions can be drawn individually, i.e. we can fix $p-1$ coordinates to a value, say for example $x_k = 0$ if $k \neq j$, and, using the additivity, represent

the maps $x_j \to R^j(x_j)$ for $j = 1,\dots,d$, in order to get a better understanding of the (eventually) non-linearity present in the data in term of x_j,
3. the probabilistic model allows to choose the intrinsic dimension of the projection and the regression function using a model selection criterion.

Some illustrative tables and graphs can be found in the examples given in this paper (section 5) and illustrate these facts. The main disadvantages of the method are

1. its use of intensive numerical computations, so that it cannot be applied to very big data sets (either in the number of individuals or number of variables),
2. the use of linear projection functions doesn't allow to modelize data presenting too strong non-linearity (see proposition 3).

The paper is organized as follows. Section 2 introduces the Probabilistic Semi-Linear Auto-Associative Models and relate them to the PCA and Probabilistic PCA models. In section 3 we present the Semi-Linear PCA models and the estimation of theirs parameters conditionally to the knowledge of the projection matrix **P**. Section 4 focus on implementation details, in particular section 4.1 is devoted to the determination of the projection matrix **P** using either PCA or contiguity analysis, and section 4.2 presents the additive B-Spline regression. Data sets and experiments are detailed in Section 5 with a real astronomical data set. Some comparisons are done with the usual PCA method. Finally, some concluding remarks are proposed in Section 6.

2 Semi-Linear Auto-Associatif Models (SLAAM)

2.1 Geometrical properties of the SLAAM

Let us first consider a general auto-associative model $\mathscr{M}_g$ with $g = R \circ P$ as given in the definition 1. We have the following evident property

Proposition 1 *Let $H = \{P(\mathbf{y}); \mathbf{y} \in \mathscr{M}_g\} \subset \mathbb{R}^d$. On H the projection function and the regression function verify*

$$P \circ R = \mathrm{Id}_d \tag{1}$$

where Id_d denote the identity function of $\mathbb{R}^d$.

Proof Let $\mathbf{y} \in \mathscr{M}_g$ and let $\mathbf{x} = P(\mathbf{y})$, then

$$\mathbf{x} = P(\mathbf{y}) = P(g(\mathbf{y})) = P(R(P(\mathbf{y}))) = P(R(\mathbf{x})).$$

As a consequence, we have the following "orthogonality" property verified by an AAM when P is an additive function

Proposition 2 *Let $V = \{P(\mathbf{y}); \mathbf{y} \in \mathbb{R}^p\}$ and assume that the property (1) extends on V, let $\mathbf{y} \in \mathbb{R}^p$, $\bar{\mathbf{y}} = R(P(\mathbf{y}))$ and $\bar{\boldsymbol{\varepsilon}} = \mathbf{y} - \bar{\mathbf{y}}$. If P is additive, i.e. $P(\mathbf{y}+\mathbf{y}') = P(\mathbf{y}) + P(\mathbf{y}')$, then*

$$P(\bar{\boldsymbol{\varepsilon}}) = 0.$$

Proof Using the property (1), we have on one hand $P(\bar{\mathbf{y}}) = P(R(P(\mathbf{y}))) = P(\mathbf{y})$. While on the other hand $P(\bar{\mathbf{y}}) = P(\mathbf{y} - \bar{\varepsilon}) = P(\mathbf{y}) - P(\bar{\varepsilon})$ giving the announced result.

Clearly we have $H \subset V$ and the assumption given in this proposition seems quite natural. We focus now on the semi-linear case and we assume that

$$P(\mathbf{y}) = \left(\left\langle \mathbf{a}^1, \mathbf{y} \right\rangle, \ldots, \left\langle \mathbf{a}^d, \mathbf{y} \right\rangle \right) = \mathbf{P}\mathbf{y}. \tag{2}$$

with $\mathbf{P} = (\mathbf{a}^1, \ldots, \mathbf{a}^d)'$ a matrix of size (d, p).

Proposition 3 *Let $g = R \circ P$ be an auto-associative function, with P given in (2) and R verifying the property (1). Let $\mathscr{B} = \left(\mathbf{a}^1, \ldots, \mathbf{a}^d, \mathbf{a}^{d+1}, \ldots, \mathbf{a}^p\right)$ be an orthonormal basis of $\mathbb{R}^p$ with $(\mathbf{a}^{d+1}, \ldots, \mathbf{a}^p)$ chosen arbitrarily. Let $\mathbf{y} \in \mathscr{M}_g$, and let $\tilde{\mathbf{y}}$ and $\tilde{R}$ denote respectively the vector $\mathbf{y}$ and the auto-associative function R in the basis $\mathscr{B}$, then*

$$\begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_d \\ \tilde{y}_{d+1} \\ \vdots \\ \tilde{y}_p \end{pmatrix} = \begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_d \\ \tilde{R}_{d+1}(\tilde{y}_1, \ldots, \tilde{y}_d) \\ \vdots \\ \tilde{R}_p(\tilde{y}_1, \ldots, \tilde{y}_d) \end{pmatrix}. \tag{3}$$

Proof It is sufficient to notice that the change of basis matrix $\mathbf{Q}$ is given by

$$\mathbf{Q}' = \left(\mathbf{a}^1, \ldots, \mathbf{a}^d, \mathbf{a}^{(d+1)}, \ldots, \mathbf{a}^p \right),$$

thus the left multiplication of $\mathbf{y}$ and R by $\mathbf{Q}$, using (1), will give (3).

From this last proposition we can see that the Semi-Linear Auto-Associative models have a relatively simple geometrical structure and that we cannot expect to model highly non-linear models with them.

2.2 Probabilistic Semi-Linear Auto-Associative Models

In the sequel, we will denote by V the subspace spanned by the set of vectors $(\mathbf{a}^1, \ldots, \mathbf{a}^d)$, and give us an arbitrary orthonormal basis of $V^\perp$ denoted by $(\mathbf{a}^{d+1}, \ldots, \mathbf{a}^p)$. We will denote by $\mathbf{P}$ the matrix $(\mathbf{a}^1, \ldots, \mathbf{a}^d)'$ and by $\bar{\mathbf{P}}$ the matrix $(\mathbf{a}^{d+1}, \ldots, \mathbf{a}^p)'$. As in proposition 3, $\mathbf{Q}$ represents the unitary matrix $(\mathbf{P}|\bar{\mathbf{P}})' = (\mathbf{a}^1, \ldots, \mathbf{a}^p)'$.

2.2.1 General Gaussian Setting

Definition 2 Let $\mathbf{x}$ be a d-dimensional Gaussian random vector:

$$\mathbf{x} \sim \mathscr{N}(\mu_x, \Sigma_x) \tag{4}$$

and let $\tilde{\varepsilon}$ be a p-dimensional centered Gaussian random vector, independent of $\mathbf{x}$, with a diagonal covariance matrix $\Sigma_{\tilde{\varepsilon}} = \mathrm{Diag}(\sigma_1, \ldots, \sigma_p)$. The p-dimensional vector

$\mathbf{y}$ is a Probabilistic Semi-Linear Auto-Associative Model (semi-linear PCA) if it can be written as

$$\mathbf{y} = \mathbf{Q}' \left(\begin{pmatrix} x_1 \\ \vdots \\ x_d \\ \tilde{r}_{d+1}(\mathbf{x}) \\ \vdots \\ \tilde{r}_p(\mathbf{x}) \end{pmatrix} + \tilde{\varepsilon} \right) = R(\mathbf{x}) + \varepsilon, \tag{5}$$

where the $\tilde{r}_j(\mathbf{x})$, $d+1 \leq j \leq p$, are arbitrary real functions from $\mathbb{R}^d$ to $\mathbb{R}$.

2.2.2 Link with the Principal Component Analysis

Assume that:

1. $\tilde{r}_j(\mathbf{x}) = \tilde{\mu}_j$ for all $j \in \{d+1, \dots p\}$,
2. the covariance matrix of $\mathbf{x}$, $\Sigma_x = \mathrm{Diag}(\sigma_1^2, \dots, \sigma_d^2)$ is diagonal with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_d$,
3. the Gaussian noise $\tilde{\varepsilon}$ has the following covariance matrix $\Sigma_\varepsilon = \mathrm{Diag}(0, \dots, 0, \sigma^2, \dots, \sigma^2)$ (d zeros and $p-d$ sigmas on the diagonal) with $\sigma < \sigma_d$

then the vector $\mathbf{y}$ is a Gaussian random vector

$$\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$$

with

$$\mu = \mathbf{Q}' \begin{pmatrix} \tilde{\mu}_1 \\ \vdots \\ \tilde{\mu}_d \\ \tilde{\mu}_{d+1} \\ \vdots \\ \tilde{\mu}_p \end{pmatrix} \quad \text{and} \quad \Sigma = \mathbf{Q} \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & 0 & \\ & & \sigma_d & & & \\ & & & \sigma & & \\ & 0 & & & \ddots & \\ & & & & & \sigma \end{pmatrix} \mathbf{Q}'$$

and $\mathbf{a}^1, \dots, \mathbf{a}^d$ are the d first eigenvectors given by PCA.

2.2.3 Link with the Probabilistic Principal Component Analysis

The probabilistic PCA Tipping and Bishop (1999) is a model of the form

$$\mathbf{y} = \mu + \mathbf{W}\mathbf{x} + \varepsilon, \tag{6}$$

with $\mathbf{W}$ a (p,d) matrix, $\mathbf{x}$ a d-dimensional isotropic Gaussian vector, i.e. $\mathbf{x} \sim \mathcal{N}(0, I_d)$, and ε a p-dimensional centered Gaussian random vector with covariance matrix $\sigma^2 I_p$. The law of $\mathbf{y}$ is not modified if $\mathbf{W}$ is right multiplied by a (d,d) unitary matrix, it is thus possible to impose to the rows of $\mathbf{W}$ to be orthogonal (assuming that $\mathbf{W}$ is of full rank).

The following proposition is then straightforward

Proposition 4 *Assume that $\tilde{\varepsilon}$ (and thus ε) is an isotropic Gaussian noise, i.e. $\Sigma_{\tilde{\varepsilon}} = \sigma^2 I_p$, take $\tilde{r}_j = \tilde{\mu}_j$ for all $d+1 \leq j \leq p$ and set*

$$\mathbf{W} = \mathbf{P}' \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_d \end{pmatrix}.$$

The resulting Probabilistic Semi-Linear Auto-Associative Model is a Probabilistic Principal Component Analysis.

For this simple model there exists a close form of the posterior probability of $\mathbf{y}$ and for the maximum likelihood of the parameters of the model. In particular, the matrix $\mathbf{W}$ can be estimated up to a rotation and spans the principal subset of the data.

3 Semi-Linear PCA

Our aim is now to generalize the PCA model we present in part 2.2.2 to the semi-linear case and we assume in the rest of this article that
[N] *the Gaussian noise $\tilde{\varepsilon}$ have the following covariance matrix*

$$\Sigma_{\tilde{\varepsilon}} = \mathrm{Diag}(0, \dots, 0, \sigma^2, \dots, \sigma^2).$$

$\Sigma_{\tilde{\varepsilon}}$ is a diagonal matrix with d zeros and $p-d$ sigmas on the diagonal.

As we said in the introduction, adding non-linearity make everything more difficult and a practicable/tractable implementation for a real usage forces us to make some simplifications. The model is estimated using a two stages strategy: we first determine the projection matrix $\mathbf{P}$ using either the first components of a PCA or using a contiguity analysis (section 4.1), and in a second step we estimate the parameters and the regression function R (section 4.2).

Using **[N]** and expressing $\mathbf{y}$ in the basis $\mathscr{B}$ (definition 2) by rotation we get the following expression for $\tilde{\mathbf{y}}$:

$$\begin{pmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_d \\ \tilde{y}_{d+1} \\ \vdots \\ \tilde{y}_p \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \\ \tilde{r}_{d+1}(\mathbf{x}) \\ \vdots \\ \tilde{r}_p(\mathbf{x}) \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\varepsilon}_{d+1} \\ \vdots \\ \tilde{\varepsilon}_p \end{pmatrix}. \tag{7}$$

In other words, the coordinates of $\tilde{\mathbf{y}}$ can be split in two sets. The first d coordinates are the Gaussian random vector $\mathbf{x}$, while the remaining $p-d$ coordinates are a random vector $\mathbf{z}$ which is conditionally to $\mathbf{x}$ a Gaussian random vector $\mathscr{N}\left(\tilde{r}(\mathbf{x}), \sigma^2 I_{p-d}\right)$. Observe that the regression functions are dependents of the choice of the vectors $\mathbf{a}_{d+1}, \dots, \mathbf{a}_p$ and that, as the noise ε lives in the orthogonal of V, we have $\mathbf{x} = \mathbf{P}\mathbf{y}$.

The parameters we have to estimate are the position and correlation parameters μ_x and Σ_x for the $\mathbf{x}$ part and the noise and the regression functions $(\sigma^2, \tilde{r})$ for the non-linear part. Given a set of n points $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ in $\mathbb{R}^p$, we get by projection two sets of n points $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' = \mathbf{YP}'$ in $\mathbb{R}^d$, and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' = \mathbf{Y\bar{P}}'$ in $\mathbb{R}^{p-d}$.

Standard calculation give the Gaussian maximum likelihood for μ_x and Σ_x

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i. \tag{8}$$

and

$$\hat{\Sigma}_x = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mu}_x)(\mathbf{x}_i - \hat{\mu}_x)'. \tag{9}$$

The maximum likelihood of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n(p-d)}\sum_{i=1}^{n} \left\| \mathbf{z}_i - \hat{\tilde{\mathbf{r}}}(\mathbf{x}_i) \right\|^2. \tag{10}$$

It remains to estimate $\tilde{\mathbf{r}}$. This will be obtained by minimizing $\hat{\sigma}^2$. Many parametric or non-parametric regression tools can be used for this purpose. A practical method based on B-Spline is presented in the next section.

4 Computational aspects

This section is devoted to the description of two main computational steps: the determination of a "good" projection matrix (section 4.1), and the estimation of the regression function (section 4.2). We also present briefly the model selection criteria AIC and BIC (section 4.3) and we terminate with some practical implementation details (4.4). All the methods described are implemented in a C++ library called aam.

4.1 Computing the Projection matrix : PCA and Contiguity Analysis

Given $(\mathbf{a}^1, \ldots, \mathbf{a}^d)$ an orthonormal set of vector in $\mathbb{R}^p$, an index I: $\mathbb{R}^{p\times d} \to \mathbb{R}_+$ is a functional measuring the interest of the projection of the vector $\mathbf{y}$ on $\mathrm{Vec}(\mathbf{a}^1, \ldots, \mathbf{a}^d)$ with a non negative real number. The choice of the index I is crucial in order to find "good" parametrization directions for the manifold to be estimated. We refer to Huber (1985) and Jones and Sibson (1987) for a review on this topic in a projection pursuit setting. The meaning of the word "good" depends on the considered data analysis problem. For instance, Friedman *et al* Friedman and Tukey (1974); Friedman (1987), and more recently Hall Hall (1990), have proposed an index which measure the deviation from the normality in order to reveal more complex structures of the scatter plot. An alternative approach can be found in Caussinus and Ruiz-Gazen (1995) where a particular metric is introduced in PCA in order to detect clusters. We can also mention the index dedicated to outliers detection Pan et al (2000).

A widely used choice of I is $I(\langle \mathbf{a}^1, \mathbf{y} \rangle, \ldots, \langle \mathbf{a}^d, \mathbf{y} \rangle) = tr(\mathrm{Var}[\mathbf{Py}])$, the projected variance. This is the criterion maximized in the usual PCA method Jolliffe (1986). It can be computed using the total variance matrix

$$\mathbf{V} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\mu})(\mathbf{y}_i - \hat{\mu})'. \tag{11}$$

and by maximizing

$$I(\mathbf{a}^1, \ldots, \mathbf{a}^d) = \sum_{i=1}^{d} \mathbf{a}^{i\prime} \mathbf{V} \mathbf{a}^i, \qquad \langle \mathbf{a}^i, \mathbf{a}^j \rangle = \delta_{ij}, \tag{12}$$

where δ_{ij} represent the Kronecker's delta. The resulting vectors are then the eigenvectors associated to the d largest eigenvalues of $\mathbf{V}$.

An other approach generalizes the one presented in Girard and Iovleff (2005) and consists in defining a contiguity coefficient similar to Labart one's Lebart (2000) whose maximization allows to unfold nonlinear structures. A *contiguity* matrix is a $n \times n$ boolean matrix M whose entry is $m_{ij} = 1$ if data points i and j are "neighbors" and $m_{ij} = 0$ otherwise. Lebart proposes to use a threshold r_0 to the set of $n(n-1)$ distances in order to construct this matrix (but the choice of r_0 could be delicate if the data scale is not homogeneous) or to use a k-contiguity matrix, i.e. to set $m_{ij} = 1$ iff j is one of the k-nearest neighbor of i. This is the approach we have implemented.

The contiguity matrix being determinate, we compute the *local covariance* matrix

$$\mathbf{V}^* = \frac{1}{2kn} \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} (\mathbf{y}_i - \mathbf{y}_j)(\mathbf{y}_i - \mathbf{y}_j)'. \tag{13}$$

The axis of projection are then estimated by maximizing the contiguity index

$$I(\mathbf{a}^1, \ldots, \mathbf{a}^d) = \sum_{i=1}^{d} \frac{\mathbf{a}^{i\prime} \mathbf{V}^* \mathbf{a}^{\mathbf{i}}}{\mathbf{a}^{i\prime} \mathbf{V} \mathbf{a}^i}. \tag{14}$$

which can be expected to unfold the scatter-plot since distant observations are ignored in computing the local covariance matrix. Using standard optimization techniques, it can be shown that the resulting axis are the d eigenvectors associated with the largest eigenvalues of the matrix $V^{*-1}V$.

The behavior of the PCA and contiguity indexes is illustrated on the next two figures (figure 1). Three hundred data have been simulated uniformly on the intervals $[-1,1]^2$ and $[-2,2]^2$ and the third coordinate have been computed on the parabola $z = x^2 + y^2$. In the first case PCA and contiguity methods give almost the same results, while in the second case the contiguity method preserve the local structure of the scatter-plot.

Remark: In all the examples k has been fixed to 3. It seems to be a good compromise between the computational time (the computation of $\mathbf{V}^*$ increase with k) and the stabilization of the index.

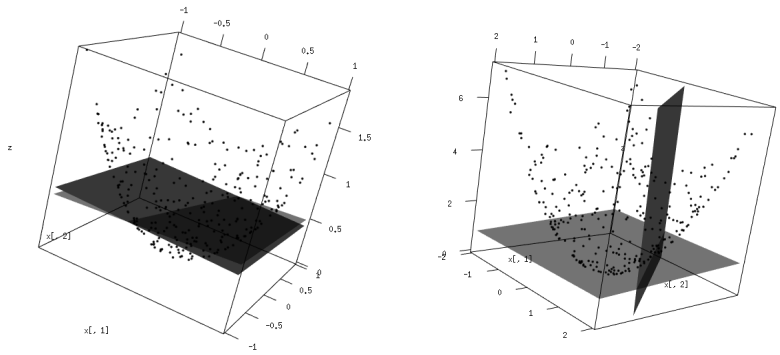


Fig. 1 Planes of projection using the PCA index (dark gray) and the contiguity index (light gray) with 3 neighbors. The planes have been shifted from the origin for display purpose.

4.2 Estimating the Regression function: Linear regression and additive B-Spline regression

4.2.1 Linear Regression

In the linear case, we are looking for a vector μ and a $(d, p-d)$ matrix $\mathbf{R}$ minimizing

$$\sum_{i=1}^{n} \left\| \mathbf{z}_i - \mu - \mathbf{R}'\mathbf{x}_i \right\|^2.$$

It is easily verified that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{z}_i - \mathbf{R}'\mathbf{x}_i\right) = \hat{\mu}_z - \mathbf{R}'\hat{\mu}_x.$$

Setting $\bar{\mathbf{X}} = \mathbf{X} - \mathbf{1}\hat{\mu}_x'$ and $\bar{\mathbf{Z}} = \mathbf{X} - \mathbf{1}\hat{\mu}_z'$, where $\mathbf{1}$ represent a vector of size n with 1 on every coordinates. Assuming that the matrix $\bar{\mathbf{X}}'\bar{\mathbf{X}}$ is invertible, standard calculus show that

$$\hat{\mathbf{R}} = (\bar{\mathbf{X}}'\bar{\mathbf{X}})^{-1}\bar{\mathbf{X}}'\bar{\mathbf{Z}}.$$

Finally, using the decomposition in eigenvalues of the covariance matrix of $\mathbf{Y}$, it is straightforward to verify the following theorem

Theorem 1 *If the d orthonormal vectors* $\mathbf{a}^1, \ldots, \mathbf{a}^d$ *are the eigenvectors associated with the first d eigenvalues of the covariance matrix of* $\mathbf{Y}$ *then the estimated auto-associative model is the on obtain by PCA.*

4.2.2 Additive B-Spline regression

B-Spline regression allows to approximate a curve $\mathbf{s}(x)$ defined from the interval $[a,b]$ into $\mathbb{R}^p$ using a set of n observations $(x_i, \mathbf{y}_i)_{i=1}^n$ ranging over $[a,b] \times \mathbb{R}^p$. Given a set of $m+k+1$ real values t_i, called knots, such that

$$t_0 \leq t_1 \leq \ldots \leq a = t_{k-1} \leq \ldots \leq t_{m+1} = b \leq \ldots \leq t_{m+k}.$$

A B-Spline curve of degree k is defined by

$$\mathbf{s}(x) = \sum_{l=0}^{m} \alpha_l s^l(x), \quad x \in [t_{k-1}, t_{m+1}].$$

Points $(\alpha_l)_{l=0}^m$ are called the control points and the functions $(s^l)_{l=0}^m$ are the $m+1$ B-Spline basis of order k computed using the Cox-De Boor recursion formula (see de Boor (1978) for details). m and k are numbers fixed by the user.

Given d sets of $m+k+1$ knots t_i^j, an additive B-spline curve of degree k is defined by

$$\mathbf{s}(\mathbf{x}) = \sum_{j=1}^{d} \sum_{l=0}^{m} \alpha_l s^{jl}(x_j), \quad x_j \in [t_{k-1}^j, t_{m+1}^j].$$

For simplicity, the number of control point have been set equal for all j.

In order to estimate the regression function $\tilde{\mathbf{r}}$, we express it as an additive linear combination of $m+1$ B-Spline basis functions r^{jl}. We have thus to estimate the set of coefficients (α_{jl}), $j = 1, \ldots, d$, $l = 0, \ldots m$ by minimizing

$$\sum_{i=1}^{n} \left\| \mathbf{z}_i - \alpha_0 - \sum_{j=1}^{d} \sum_{l=0}^{m} \alpha_{jl} r^{jl}(x_{ij}) \right\|^2 .$$

Standard regression techniques give then the estimates $\hat{\alpha} = ((\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{Z})$ where $\mathbf{S}$ is the design matrix which depends on the knots position, the degree of the B-Spline and the number of control points chosen by the user Prautzsch et al (2002). The regression function $\tilde{\mathbf{r}}$ is then estimated by the formula

$$\hat{\tilde{\mathbf{r}}}(\mathbf{x}) = \hat{\alpha}_0 + \sum_{j=1}^{d} \sum_{l=0}^{m} \hat{\alpha}_{jl} r^{jl}(x_j) = \hat{\alpha}_0 + \sum_{j=1}^{d} \hat{\tilde{\mathbf{r}}}^j(x_j).$$

Remark: In the examples presented below we use cubic B-Spline basis but the aam library allow the user to chose its own degree. The knots $\mathbf{t}^j$ are regularly spaced in the range of $(x_{ij})_{i=1}^n$.

4.3 Model Selection

Since a Semi-linear PCA model depends highly of the projection matrix $\mathbf{P}$ and the regression function, model selection allows to select among various candidates the best projection and the number of control points. Several criteria for model selection have been proposed in the literature and the widely used are penalized likelihood criteria. Classical tools for model selection includes Akaike Akaike (1974) and Bayesian Schwarz (1978) information criteria. The Akaike Information Criterion (AIC) is a measure of the relative goodness of fit of a statistical model given by $-2\ln(L) + 2\gamma(\mathscr{M})$ and the Bayesian Information Criterion (BIC) consists in selecting the model which penalizes the log-likelihood by $\frac{\gamma(\mathscr{M})}{2}\log(n)$ where $\gamma(\mathscr{M})$ is the number of parameters of the model $\mathscr{M}$ and n is the number of observations.
Remark: In this paper we use BIC because of its popularity but the aam library allows to use AIC.

4.4 Estimation in practice

The drawback of the previous maximum likelihood equations (7) and (10) is that we have to perform a rotation of the original data set in order to estimate the regression function and next to perform and inverse rotation of the estimated model. In practice, having fixed the dimension d of the model and the number of control point m, we avoid such computations by estimating the model using the following steps:

- Center and (optionally) standardize the data set $\mathbf{Y}$: obtain $\bar{\mathbf{Y}}$,
- Compute the projected data set $\mathbf{X} = \bar{\mathbf{Y}}\mathbf{P}'$ ($\mathbf{X}$ is centered),
- Compute the regression $\bar{\mathbf{Y}} \sim \mathbf{X}$ (without intercept),
- Compute the residual variance $\hat{\sigma}^2$, the log-likelihood and the BIC.

As we can see the main difference is in the regression step: we estimate directly a function from $\mathbb{R}^d$ to $\mathbb{R}^p$. In practice, as the non-linear part of the model is in $V^\perp$, the regression function we obtain numerically give the identity function in the V space.

These estimations are repeated for d (the dimension of the auto-associative model) varying from 1 to a fixed value $d_{\max}$, and for m (the number of control point) varying from 4 to a fixed value $m_{\max}$ (for a cubic B-Spline interpolation we cannot have less than 4 control points). The model with smallest BIC is selected.

5 Examples

We first present two illustrations of the estimation principle of semi-linear PCA on low dimensional data (Section 5.1 and 5.2). Second, semi-linear PCA is applied to an astronomical analysis problem in section 5.3. In all the examples, we use an additive B-Spline regression model for the estimation of the regression function $\tilde{R}$ (section 4.2.2). The B-Spline are of degree 3 and we select the number of control points using BIC. All the results are compared with PCA results.

5.1 First example on simulated data

The data are simulated using a one-dimensional regression function in $\mathbb{R}^3$. The equation of the AA model is given by

$$u \to (u, \sin u, \cos u), \tag{15}$$

and thus $P(u,v,w) = u$. The first coordinate of the random vector is sampled 1000 times from a centered Gaussian distribution with standard deviation $\sigma_x = 3$. An independent noise with standard deviation $\sigma = 1$ is added to the v and w coordinates.

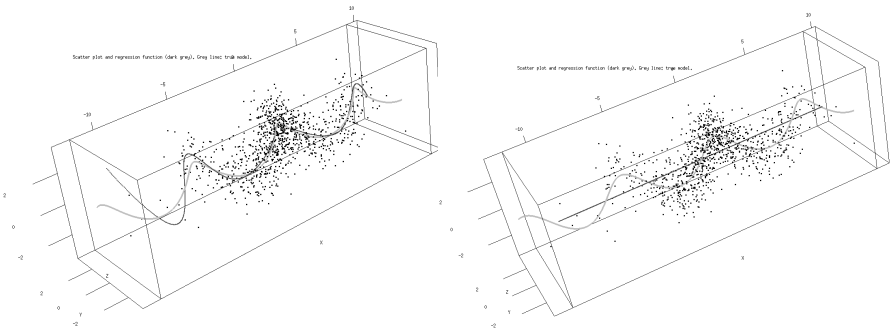
The axis of projection have been computed by PCA and by contiguity analysis (section 4.1) using the 3 nearest neighbors for the proximity graph. The correlations between the projected data set x and the original data set are

	y_1	y_2	y_3
x (contiguity)	0.999968	-0.000579458	0.0089
x (PCA)	0.999997	0.008813556	0.0118

which show that the first axis given either by contiguity analysis or PCA is very close from the x-axis as it was expected. The projected variance on the first axis find by contiguity analysis is 9.20496 which is also very close to 9. We use BIC in order to select the dimension of the model and the number of control points using an additive B-Spline regression. A summary of the tested model is given in the table 1 and the estimated AAM using B-Spline regression or PCA is drawn in the figures 2 and 3.

		Contiguity Analysis		PCA	
	dim	BIC	Residual Variance	BIC	Residual Variance
linear	1	9987.13(5)	1.439	11495.8	1.43864
	2	10609.7(10)	1.40815		
9	1	11247.6(29)	1.06557	11058.4	1.06407
	2	11621(58)	1.1086		
10	1	11060.7(32)	0.986316	10926.3	0.985777
	2	11469(64)	0.913605		
11	1	10853.7(35)	0.941064	10855	0.92711
	2	11503(70)	0.90682		
12	1	10844.6(38)	0.92717	10845(38)	0.941661
	2	11525(76)	0.892669		
13	1	10871.4(41)	0.929965	10871.6	0.927976
	2	11568.4(82)	0.891119		
14	1	108887.5(44)	0.927834	10888.3	0.927976
	2	11604(88)	0.885939		

Table 1 Values of BIC for $d = 1$ and $d = 2$ and for various number of control points (given in the first column). The number of free parameters of each model is given in parenthesis. BIC selects the model of dimension 1 with 12 control points. The axis of projection can be either obtained by contiguity analysis or obtained using PCA.



(a) (b)

Fig. 2 (a) The simulated scatter-plot, the estimated AAM (dark grey) and the true AAM (thick grey) using a contiguity analysis and a B-Spline regression. (b) Results using an usual PCA (the straight line). This graphic is obtained with R using the *draw3d* command of the aam library.

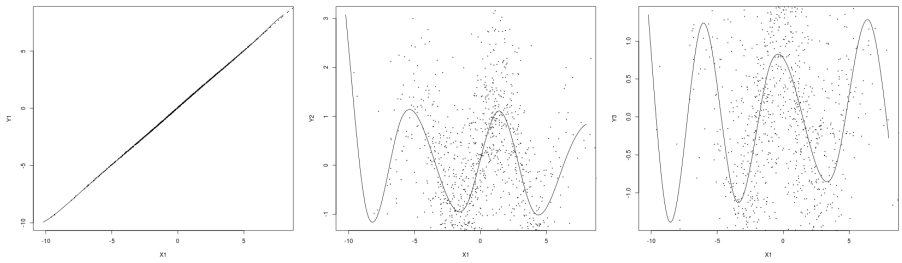


Fig. 3 Estimated regression functions $x \to R_k(x)$ and the scatter plots $(x_i, y_{ik})_{i=1}^n$ for $k = 1,2,3$. The true model is given in the equation 15. This graphic is obtained with R using the *draw.regression* command of the aam library.

5.2 Second example on simulated data

In our second example the AAM is given by

$$(u,v) \to \left(u, v, u^2 + v^2\right) \tag{16}$$

and thus $P(u,v,w) = (u,v)$. The first two coordinates of the random vector are sampled from a centered Gaussian distribution with covariance matrix

$$\Sigma_x = \begin{pmatrix} 0.5^2 & 0 \\ 0 & 0.55^2 \end{pmatrix}$$

and $n = 1000$ points are simulated. An independent noise with standard deviation $\sigma = 0.2$ has then been added to the z coordinate.

The result of the contiguity analysis with $k = 3$ neighbors is displayed in the figure 4 (but only the 2-neighbors graph is drawn). The correlation circle shows that the first component obtains by the contiguity analysis is essentially correlated with the second variable (y-axis) and that the second component is very correlated with the first variable (x-axis) and a bit with the third variable (z-axis). The last graphic shows that in contrast the first component obtain by PCA is essentially correlated with the z-axis.

The model is estimated using an additive B-Spline regression. BIC selects 6 control points (40 parameters) and the residual standard deviation is 0.309132 which overestimate a bit the true level of noise. The true model and the estimated model obtained with an additive B-Spline regression are given in the figure 5 and compared with the usual PCA. The figure 6 draw the estimation of the regression functions (R_k^j) for $j = 1,\ldots,2$ and $k = 1,2,3$.

5.3 Example in spectrometry analysis

Finally we illustrate the performance of the semi-linear PCA on a real data set. The data consists of 19-dimensional spectral information of 487 stars Stock and Stock (1999); Garcia et al (2004); Stock et al (2002); Garcia et al (2008) and they have

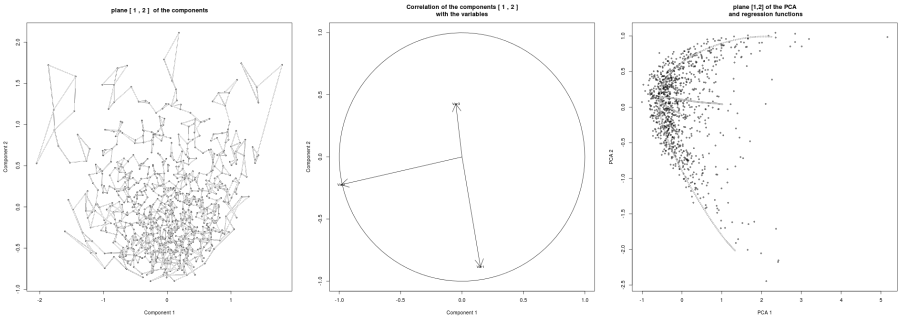
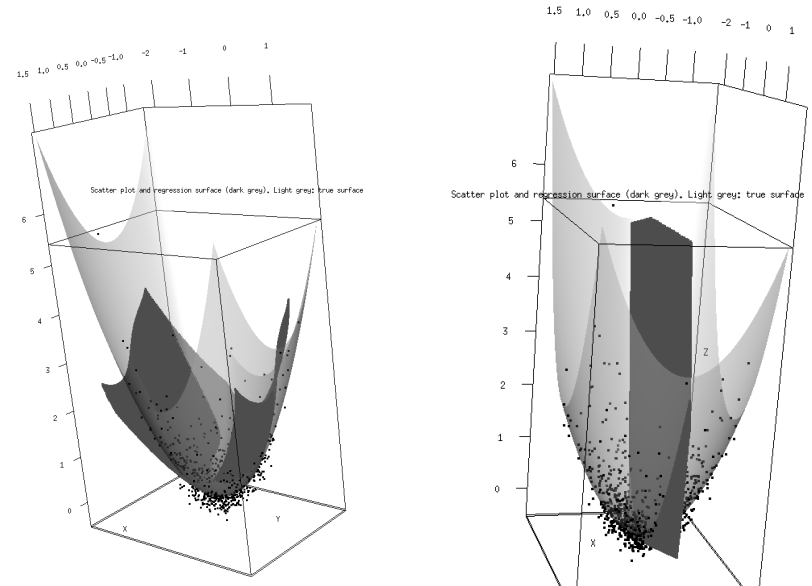


Fig. 4 The AAM components and the 2-neighbors graph, the correlation circle of the AAM components with the variables and representation of the scatter-plot in the PCA plan. These pictures have been obtained using the *plot* command of the aam library.



(a) (b)

Fig. 5 (a) Result using a contiguity analysis and an additive B-Spline regression. (b) Result using PCA. These images have been obtained using the *draw3d* function of the aam library.

been classified in 6 groups. They have been modeled by Scholz et al (2007) using an auto-associative neural networks based on a 19-30-10-30-19 network. Using the terminology of this article the model proposed by M. Scholz and its co-authors is an auto-associative model of dimension 10. We select the model using BIC. The main results are the following:

1. The axis of projection given by the PCA outperform largely the results we obtain with the contiguity analysis, for any choice of the number of control points.

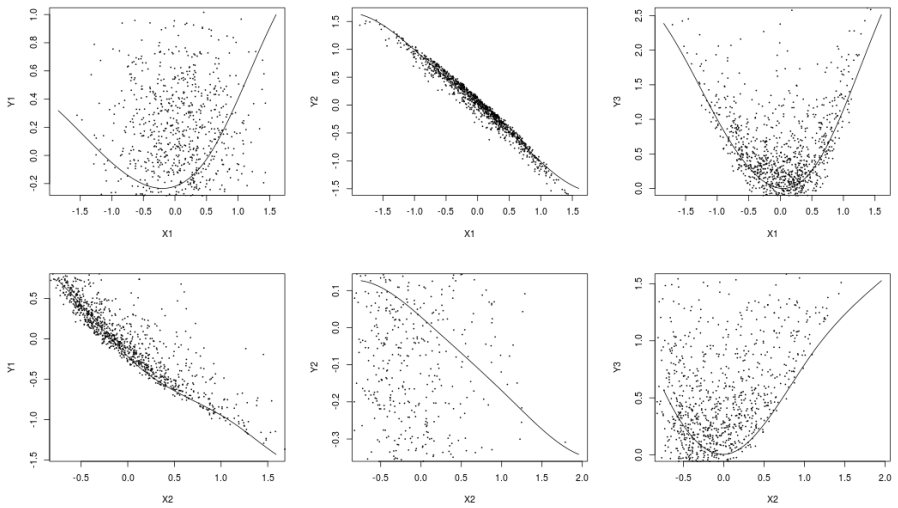


Fig. 6 Estimated regression functions $x_j \to R_k^j(x_j)$ and the scatter plots $(x_{ij}, y_{ik})_{i=1}^n$ for $k = 1,2,3$ and $j = 1,2$. The true model is given in the equation 16. This graphic is obtained with R using the *draw.regression* command of the aam library.

2. BIC retains a model of dimension 5 with 9 Control Points (871 parameters) when we use a non-linear regression step. The residual variance is $\sigma^2 = 0.0080763$ while the total variance (inertia) of the data was 26.59832.
3. BIC retains a model of dimension 12 (307 parameters) when we use a linear regression step. Observe that in this case, we are performing an usual PCA (theorem 1).

The scatter plot in the main PCA space with the correlation circle and the individuals regression functions in the plane (6,7) of the PCA. are given in the figure 7.

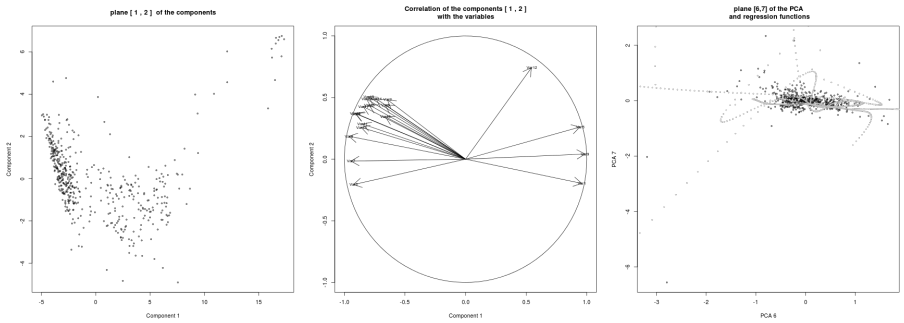


Fig. 7 Scatter plot in the PCA space, correlation circle of the first two components with the variables and plane (6,7) of the PCA with the estimated regression functions.

A summary of some tested model are given in the table 2. We present a visualization of some of the estimated regression functions in figure 8. There is $5 \times 19 = 95$ regression functions to plot but we show only the first five variables $(y_1, \ldots, y_5)$ in term of the projected variables $\mathbf{x} = (x_1, \ldots, x_5)$. The first projected variable x_1 fits very well the variables y_j.

PCA			
Control Points	BIC	dim	Residual variance
Linear	1829,95(307)	12	0,0049702
7	1187,26(820)	6	0,00727521
8	1147,62(776)	5	0,00975073
9	453,387(871)	5	0,0080763
10	701,769(966)	5	0,00768342
11	1333,45(1061)	5	0,00773327

Table 2 Values of BIC for various number of control points (given in the first column). BIC selects the model of dimension 5 with 9 control points using as projection matrix the 5 axis given by PCA. The total variance (inertia) of the data set was 26.59832.

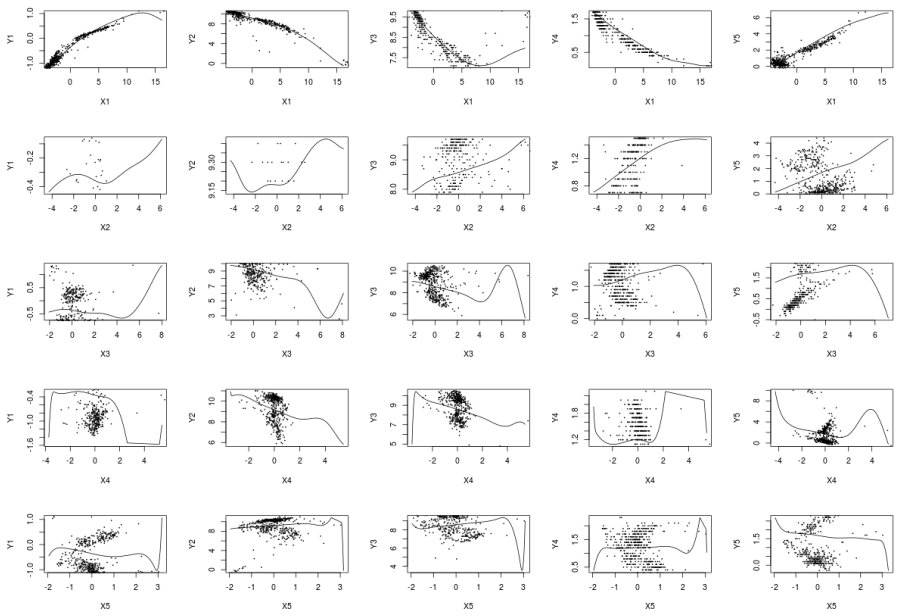


Fig. 8 The estimated regression functions from $\mathbb{R}^5$ to $\mathbb{R}^{19}$. Only the first five dimensions are given.

6 Conclusion

We have presented a class of auto-associative models for data modeling and visualization called semi-linear auto-associative models. We provided theoretical groundings for these models by proving that the principal component analysis and the probabilistic principal component analysis are special cases. This model allows to model data set with a simple non-linear component and is truly generative with an underlying probabilistic interpretation. However it does not allow to model data with a strong non-linear component and it depends highly on the choice of the projection matrix.

The Semi-Linear PCA have been implemented in C++ using the *stk++* library Iovleff (2012) and is available at: `https://sourcesup.renater.fr/projects/aam/`. The program is accompanied with a set of *R* scripts which allows to simulate data sets and display the results of the *aam* program.

Acknowledgements The author thanks two anonymous referees for comments that improved the readability of this article.

References

Akaike H (1974) A new look at the statistical mode identification. IEEE Transaction on Automatic Control 19:716–723

Baldi P, Hornik K (1989) Neural networks and principal component analysis: Learning from examples without local minima. Neural Networks 2(1):53–58

Besse P, Ferraty F (1995) A fixed effect curvilinear model. Computational Statistics 10(4):339–351

Bishop CM (2006) Pattern recognition and machine learning, information science and statistics. Springer, Berlin

de Boor C (1978) A practical guide to splines. Springer-Verlag

Caussinus H, Ruiz-Gazen A (1995) Metrics for finding typical structures by means of Principal Component Analysis. Data science and its Applications, Harcourt Brace Japan pp 177–192

Delicado P (2001) Another look at Principal curves and surfaces. Journal of Multivariate Analysis 77:84–116

Durand JF (1993) Generalized principal component analysis with respect to instrumental variables via univariate spline transformations. Computational Statistics and Data Analysis 16:423–440

Friedman J (1987) Exploratory projection pursuit. Journal of the American statistical association 82(397):249–266

Friedman J, Stuetzle W (1981) Projection pursuit regression. Journal of the American statistical Association 76(376):817–823

Friedman JH, Tukey JW (1974) A Projection Pursuit algorithm for exploratory data analysis. IEEE Trans on computers 23(9):881–890

Garcia J, Stock J, Stock MJ, Sanchez N (2004) Quantitative Stellar Spectral Classification. III. Spectral Resolution. RevMexAstronAstrofis 41 (2005)

31-40 URL http://arxiv.org/abs/astro-ph/0410532v1;http://arxiv.org/pdf/astro-ph/0410532v1

Garcia J, Sanchez N, Velasquez R (2008) Quantitative Stellar Spectral Classification. IV. Application to the Open Cluster IC 2391. RevMexAstronAstrofis 45 (2009) 13-24 URL http://arxiv.org/abs/0809.4188v1;http://arxiv.org/pdf/0809.4188v1

Girard S, Iovleff S (2005) Auto-Associative models and generalized principal component analysis. Journal of Multivariate Analysis 93:21–39

Gower JC, Hand DJ (1995) Biplots, vol 54. Chapman & Hall/CRC

Hall P (1990) On polynomial-based projection indices for exploratory projection pursuit. The Annals of Statistics 17(2):589–605

Hastie T, Stuetzle W (1989) Principal curves. Journal of the American Statistical Association 84(406):502–516

Hastie T, Tibshinari R, Friedman J (2001) The elements of statistical learning, second edition edn. Springer Series in Statistics, Springer

Hastie TJ, Tibshirani RJ (1990) Generalized Additive Models. Monographs on Statisics and Applied Probability 43

Hinton GE, Dayan P, Revow M (1997) Modeling the manifolds of images of handwritten digits. IEEE transactions on Neural networks 8(1):65–74

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:417–441

Huber PJ (1985) Projection Pursuit. The Annals of Statistics 13(2):435–475

Iovleff S (2012) The Statitiscal ToolKit. http://www.stkpp.org/

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

Jones MC, Sibson R (1987) What is projection pursuit? Journal of the Royal Statistical Society, Ser A 150:1–36

Klinke S, Grassmann J (2000) Projection pursuit regression. Wiley Series in Probability and Statistics pp 471–496

Lebart L (2000) Contiguity analysis and classification. In: Gaul W, Opitz O, Schader M (eds) Data Analysis, Springer-Verlag, pp 233–244

Lu BW, Pandolfo1 L (2010) Quasi-objective nonlinear principal component analysis. Neural Networks 24(2):159–170

Pan JX, Fung WK, Fang KT (2000) Multiple outlier detection in multivariate data using projection pursuit techniques. Journal of Statistical Planning and Inference 83(1):153–167

Pearson K (1901) On lines and planes of closest fit to systems of points in space. The London, Edinburgh and Dublin philosophical magazine and journal of science Sixth Series(2):559–572

Prautzsch R, Boehm W, Paluszny M (2002) Bézier and B-Spline Techniques. Mathematics and visualization, Springer

Roweis ST, Saul LK (2000) Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500):2323–2326

Schölkopf B, Smola A, Müller KR (1999) Kernel Principal Component Analysis. In: Advances in Kernel Methods—Support Vector Learning, pp 327–352

Scholz M, Fraunholz M, Selbig J (2007) Nonlinear Principal Component Analysis: Neural Network Models and Applications. In: Principal Manifolds for Data Visu-

alization and Dimension Reduction, volume 28,, Springer-Verlag, pp 205–222
Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6:461–464
Stock J, Stock M (1999) Quantitative stellar spectral classification. Revista Mexican de Astronomia y Astrofisica 34:143–156
Stock MJ, Stock J, Garcia J, Sanchez N (2002) Quantitative Stellar Spectral Classification. II. Early Type Stars. RevMexAstronAstrofis 38 (2002) 127-140 URL http://arxiv.org/abs/astro-ph/0205315v1;http://arxiv.org/pdf/astro-ph/0205315v1
Tenenbaum JB (2000) A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500):2319–2323
Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. Journal of the Royal Statistical Society, Ser B 61(3):611–622