Auto-Associative Models and Generalized Principal Component Analysis

Stéphane Girard, Serge Iovleff

To cite this version:

Stéphane Girard, Serge Iovleff. Auto-Associative Models and Generalized Principal Component Analysis. Journal of Multivariate Analysis, 2005, 93 (1), pp.21-39. 10.1016/j.jmva.2004.01.006 . hal-00383139

HAL Id: hal-00383139

https://hal.science/hal-00383139

Submitted on 23 Apr 2013

Auto-associative models and generalized Principal component analysis

Girard Stéphane[1]
INRIA Rhône-Alpes, projet IS2, ZIRST,
655, avenue de l'Europe, Montbonnot,
38334 Saint-Ismier Cedex, France.
E-mail: Stephane.Girard@inrialpes.fr
Phone: +(33) 4 76 61 53 25, Fax: +(33) 4 76 61 52 52

and

Iovleff Serge
SABRES, Université de Bretagne-Sud,
Campus Tohannic, rue Yves Mainguy,
56000 Vannes, France.
E-mail: Serge.Iovleff@univ-ubs.fr

Version:

In this paper, we propose auto-associative (AA) models to generalize Principal component analysis (PCA). AA models have been introduced in data analysis from a geometrical point of view. They are based on the approximation of the observations scatter-plot by a differentiable manifold. In this paper, they are interpreted as Projection pursuit models adapted to the auto-associative case. Their theoretical properties are established and are shown to extend the PCA ones. An iterative algorithm of construction is proposed and its principle is illustrated both on simulated and real data from image analysis.

Key Words: Auto-Associative models, Principal Component Analysis, Projection Pursuit, Regression.

1. INTRODUCTION.

Principal component analysis (PCA) [23] is a widely used method for dimension reduction in multivariate data analysis. It benefits from a simple geometrical interpretation. Given a set of points from $\mathbb{R}^p$ and an integer $0 \leq d \leq p$, PCA builds the d-dimensional affine subspace minimizing the Euclidean distance to the scatter-plot [30]. Starting from this point of view, many nonlinear extensions have been proposed. Principal curves or principal surfaces [19, 7] belong to this family of approaches. PCA

[1] corresponding author

can also be interpreted as a Projection pursuit [22, 24] method. It builds the d-dimensional affine subspace maximizing the projected variance [21]. Indeed, other criteria than the variance yields various data exploration methods [13, 29]. In PCAIV-Spline (Principal component analysis of instrumental variables [10]) and curvilinear PCA [1] approaches, nonlinear transformations of the coordinates are combined with a criteria of projected variance on the transformed data. More recently, new algorithms have been proposed to compute low dimensional neighborhood-preserving embeddings of high dimensional data. For instance, Isomap [33], local Isomap [9], LLE (Locally linear embedding) [31] and CDA (Curvilinear distance analysis) [8] do not use a criterion based on variance preservation but attempt to reproduce in the projection space the structure of the local neighborhood in the data space. Such methods are dedicated to visualization purposes. Their drawback is that they cannot produce an analytic form of the transformation function, making difficult to map new points into the dimensionality-reduced space. We refer to [27] for a comparison between Isomap and CDA and to [35] for a comparison of LLE and Isomap classification and visualization power. Finally, it is also possible to associate a Gaussian probabilistic model to PCA [32], the affine subspace is then obtained through a maximization-likelihood estimation. This approach yields new dimension-reduction methods by considering some non Gaussian models such as mixture models.

The extension of PCA to the nonlinear case without losing these interpretations is a difficult problem. Moreover, the definition of a satisfying probabilistic model is often impossible without specifying the observations distribution. As a consequence, such a method would be very specific and thus of little practical interest. Besides, introducing nonlinearities can lead to lose the geometrical interpretation of the model and the related concepts of principal variables, principal directions or residual inertia. Furthermore, existence, unicity or implementation problems often occur.

In this paper, the auto-associative (AA) models are proposed as candidates to the generalization of PCA. AA models have been introduced in [14] from a geometrical point of view. They are based on the approximation of the observations scatter-plot by a manifold. We show here that these models can also be interpreted as Projection pursuit regression models (PPR) [12, 25] adapted to the auto-associative case. Consequently, a simple algorithm, similar to an iterative PCA, is available to implement them. A probabilistic framework permitting to prove many theoretical properties is introduced as well.

Let us first consider PCA from the Projection pursuit point of view. If X is a $\mathbb{R}^p$ random vector with finite second order moment, it can be expanded as a sum of d orthogonal random variables and a residual by applying iteratively the following steps: [A] computation of the Axes, [P] Projection, [R] Regression and [U] Update (see Section 3.1 for a proof):

Algorithm 1.

```
• For $j = 0$, define $R^0 = X - \mathbb{E}[X]$.
• For $j = 1, \dots, d$ :
    [A] Determine $a^j = \arg\max_{x \in \mathbb{R}^p} \mathbb{E}\left[\left\langle x, R^{j-1} \right\rangle^2\right]$
        u.c. $\|x\| = 1$ and $\left\langle x, a^k \right\rangle = 0,\ 1 \leq k < j$.
    [P] Compute $Y^j = \left\langle a^j, R^{j-1} \right\rangle$.
    [R] Determine $b^j = \arg\min_{x \in \mathbb{R}^p} \mathbb{E}\left[\left\|R^{j-1} - Y^j x\right\|^2\right]$ u.c. $\left\langle x, a^j \right\rangle = 1$,
        (we find $b^j = a^j$) and define $s^j(t) = tb^j$, $t \in \mathbb{R}$.
    [U] Compute $R^j = R^{j-1} - s^j(Y^j)$.
```

The vectors a^j are called principal directions, the random variables Y^j principal variables, the functions s^j regression functions and the random vectors R^j residuals. Step [A] consists in computing an axis perpendicular to the previous ones and maximizing a given criteria: Here the projected variance. In our opinion, this is an arbitrary choice when X is not Gaussian. Step [P] consists in projecting the residuals on this axis to determine the principal variables, and step [R] is devoted to the search of the linear function of the principal variables best approximating the residuals. Moreover, the limitation to a set of linear functions can be restrictive as soon as X is not Gaussian. Step [U] simply consists in updating the residuals.

AA models extend the previous algorithm by considering more general steps [A] and [R]. Step [A] is considered as a Projection pursuit step, where many different criteria can be implemented. Step [R] is seen as a regression problem that can be addressed by general tools such as spline or kernel estimates. We show that this extension benefits from PCA main theoretical properties (construction of an exact model, decrease of the residuals, ...) or extends them (approximation of the scatter-plot by a manifold instead of an affine subspace).

This article is organized as follows. In Section 2, auto-associative models are defined and their main properties are given. Two particular AA models are presented in Section 3 and their characteristics are studied. In Section 4, we present a Projection pursuit algorithm adapted to the framework of AA models. Finally, some illustrations are provided in Section 5 both on simulated data and on an image analysis application.

2. AUTO-ASSOCIATIVE MODELS.

In the first part of this section, auto-associative models and some related objects are defined. In the second part, an algorithm is introduced to compute them and its theoretical properties are established.

2.1. Definitions

DEFINITION 1. A function $F\colon \mathbb{R}^p \to \mathbb{R}^p$ is a d-dimensional auto-associative function if there exist d unit orthogonal vectors a^j and d functions $s^j\colon \mathbb{R} \to \mathbb{R}^p$ such that

$$F = \left(\mathrm{Id}_{\mathbb{R}^p} - s^d \circ P_{a^d}\right) \circ \ldots \circ \left(\mathrm{Id}_{\mathbb{R}^p} - s^1 \circ P_{a^1}\right) = \coprod_{k=d}^{1} \left(\mathrm{Id}_{\mathbb{R}^p} - s^k P_{a^k}\right),$$

$P_{a^j} s^j = \mathrm{Id}_{\mathbb{R}^p}$ and $P_{a^k} s^j = 0$, $1 \le k < j \le d$, with $P_{a^j}(x) = \langle a^j, x \rangle$. The vectors a^j are called principal directions, the functions s^j are called regression functions and we note $F \in \mathcal{A}^d_{a,s}$.

In the sequel, in order to keep the text concise, the product represents the composition. The proof of the following lemma can be found in [14].

LEMMA 1. *Consider $F \in \mathcal{A}^d_{a,s}$, and suppose that the s^j, $j = 1, \ldots, d$ are $C^1(\mathbb{R}, \mathbb{R}^p)$. Then, the equation $F(x) = 0$ defines a differentiable d-dimensional manifold.*

Let $X \in \mathbb{R}^p$ be a square integrable random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $\mathbb{P}_X$ the distribution of X on $\mathbb{R}^p$, and by $\mathrm{L}^2_X(\mathbb{R}, \mathbb{R}^p)$ the set of functions s from $\mathbb{R}$ to $\mathbb{R}^p$ such that $s \circ P_a$ is $\mathbb{P}_X$ square integrable for all $a \in \mathbb{R}^p$.

DEFINITION 2. X satisfies a d-dimensional auto-associative model with principal directions $(a^1, \ldots, a^d)$, regression functions $(s^1, \ldots, s^d)$ and residual ε, if there exist $F \in \mathcal{A}^d_{a,s}$, $\mu \in \mathbb{R}^p$ and a centered random vector ε such that $F(X - \mu) = \varepsilon$.

Besides, X is said to satisfy a linear AA model when the regression functions are linear. Let us give two simple examples of auto-associative models:

- Every X satisfies a 0-dimensional AA model (choose $F = \mathrm{Id}$, $\mu = \mathbb{E}[X]$ and $\varepsilon = X - \mathbb{E}[X]$). We then have $\mathrm{Var}\left[\|\varepsilon\|^2\right] = \mathrm{Var}\left[\|X\|^2\right]$.
- Similarly, X always satisfies a p-dimensional AA model. In this case $F = 0$, $\mu = 0$ and $\varepsilon = 0$ yield $\mathrm{Var}\left[\|\varepsilon\|^2\right] = 0$.

In practice, it is important to find a balance between these two extreme cases by constructing a d-dimensional model with $d \ll p$ and $\mathrm{Var}\left[\|\varepsilon\|^2\right] \ll$

Var $\left[\|X\|^2\right]$. For instance, in the case where X is centered with a covariance matrix Σ of rank d, it satisfies a d-dimensional linear AA model with a null residual. Let us denote by a^j, $j = 1, \dots, d$ the eigenvectors of Σ associated to the positive eigenvalues. We show in Corollary 2 that

$$F(x) = \coprod_{k=d}^{1} \left(\mathrm{Id}_{\mathbb{R}^p} - P_{a^k} a^k\right)(x) = x - \sum_{k=1}^{d} \left\langle a^k, x \right\rangle a^k$$

and $\varepsilon = 0$ $\mathbb{P}$-.a.s. define a linear auto-associative model for X. This is the expansion of X obtained by PCA.
We now propose an algorithm to build some auto-associative models which are not necessarily linear, with small dimension and small residual variance. In this aim, we introduce a definition:

Definition 3. A closed subset $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ of $\mathrm{L}^2_X(\mathbb{R}, \mathbb{R}^p)$ is admissible if
$$(\mathcal{R}) : \begin{cases} \forall b \in \mathbb{R}^p \quad s \in \mathcal{S}(\mathbb{R}, \mathbb{R}^p) \Rightarrow s + b \in \mathcal{S}(\mathbb{R}, \mathbb{R}^p) \\ \qquad\qquad \mathrm{Id}_{\mathbb{R}} b \in \mathcal{S}(\mathbb{R}, \mathbb{R}^p). \end{cases}$$

$(\mathcal{R})$ can be interpreted as an invariance condition with respect to translation. A possible choice of $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ is the set of affine functions from $\mathbb{R}$ to $\mathbb{R}^p$. This example is examined in Section 3.1.
Let us recall that, given an unit vector $a \in \mathbb{R}^p$, an index I: $\mathbb{R} \to \mathbb{R}$ is a functional measuring the interest of the projection of the random vector X on a (i.e. $\langle a, X \rangle$) with a non negative real number. For instance, a possible choice of I is the projected variance $I(\langle a, X \rangle) = \mathrm{Var}\left[\langle a, X \rangle\right]$. Some other examples are presented in Section 4.2.

2.2. Construction of auto-associative models

Let $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ be a set of admissible functions, $d \in \{0, \dots, p\}$, and consider the following algorithm:

Algorithm 2.

```
• For $j = 0$, define $\mu = \mathbb{E}[X]$ and $R^0 = X - \mu$.
• For $j = 1, \dots, d$ :
    [A] Determine $a^j = \arg\max_{x \in \mathbb{R}^p} I(\langle x, R^{j-1} \rangle)$
        u.c. $\|x\| = 1$, $\langle x, a^k \rangle = 0$, $1 \le k < j$.
    [P] Compute $Y^j = \langle a^j, R^{j-1} \rangle$.
    [R] Choose $s^j \in \arg\min_{s \in \mathcal{S}(\mathbb{R}, \mathbb{R}^p)} \mathbb{E}\left[\left\|R^{j-1} - s(Y^j)\right\|^2\right]$ u.c. $P_{a^j} s^j = \mathrm{Id}$.
    [U] Compute $R^j = R^{j-1} - s^j(Y^j)$.
```

Theorem 1 below states that this algorithm builds a d-dimensional auto-associative model and an exact representation of X in p iterations.
It is clear that step [R] strongly depends on the choice of $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$. The existence of a solution to the minimization problem is established thanks to the conditions imposed on $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$. In particular, condition $(\mathcal{R})$ ensures that there exist some functions in $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ meeting the constraint of the minimization problem. The unicity of the solution is not established without an additional convexity condition. In this paper we focus on two extreme cases. The choice $\mathcal{S}(\mathbb{R}, \mathbb{R}^p) = \mathcal{A}(\mathbb{R}, \mathbb{R}^p)$, the set of the affine functions from $\mathbb{R}$ to $\mathbb{R}^p$ is examined in Section 3.1, and the choice $\mathcal{S}(\mathbb{R}, \mathbb{R}^p) = \mathrm{L}^2_X(\mathbb{R}, \mathbb{R}^p)$ is considered in Section 3.2. The choice of the index I is discussed in Section 4.2. The constraint $P_{a^j} s^j = \mathrm{Id}$ which is imposed in step [R] plays an important role in the algorithm. It ensures that the residuals R^j are orthogonal to the axis a^j since

$$\left\langle a^j, R^j \right\rangle = \left\langle a^j, R^{j-1} \right\rangle - \left\langle a^j, s^j(Y^j) \right\rangle = Y^j - Y^j = 0.$$

Thus, it is natural to iterate the model construction in the subspace orthogonal to a^j. The next theorem is mainly a consequence of this property.

Theorem 1. *Algorithm 2 builds a d-dimensional AA model with principal directions $(a^1, \ldots, a^d)$, regression functions $(s^1, \ldots, s^d)$ and residuals $\varepsilon = R^d$. Moreover, when $d = p$ then $\varepsilon = R^p = 0$ and the exact expansion holds:*

$$X = \mathbb{E}[X] + \sum_{k=1}^{p} s^k(Y^k), \ \mathbb{P} - a.s.$$

Note that these properties are quite general, since they do not depend neither on the index I, nor on the subset of admissible functions $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$. Some additional properties are provided in Section 3 for particular choices of I and $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$. We first prove the following proposition:

Proposition 1. *The residuals and the regression functions obtained with Algorithm 2 share the following properties :*

(i) For all $1 \le j \le d$, $\mathbb{E}\left[R^j\right] = \mathbb{E}\left[Y^j\right] = \mathbb{E}\left[s^j(Y^j)\right] = 0$.

(ii) For all $1 \le k \le j \le d$, $\left\langle a^k, R^j \right\rangle = 0$, $\mathbb{P}$-a.s.

(iii) For all $1 \le k < j \le d$, $\left\langle a^k, s^j(Y^j) \right\rangle = 0$, $\mathbb{P}$-a·s.

(iv) The sequence of the residual norms is $\mathbb{P}$-a.s. non increasing.

Proof.

(i) The proof is by induction on j. Let us note H_j the hypothesis $\mathbb{E}\left[R^j\right] = 0$. H_0 is clearly true. Supposing H_{j-1} is true, we thus have,

$$\mathbb{E}\left[R^j\right] = \mathbb{E}\left[R^{j-1}\right] - \mathbb{E}\left[s^j(Y^j)\right] = -\mathbb{E}\left[s^j(Y^j)\right].$$

Now, s^j is a solution of step [R] and then $\mathbb{E}\left[s^j(Y^j)\right] = 0$. This last equality can be proved by contradiction. If $\mathbb{E}\left[s^j(Y^j)\right] \neq 0$, then introduce $\mu^j = \mathbb{E}\left[s^j(Y^j)\right]$ and $s'^j = s^j - \mu^j$. Since $\left\langle a^j, s^j \right\rangle = \mathrm{Id}$ and $\mathbb{E}\left[Y^j\right] = \mathbb{E}\left[\left\langle a^j, R^{j-1}\right\rangle\right] = 0$ by H_{j-1}, we have $\left\langle a^j, \mu \right\rangle = 0$ and thus $\left\langle a^j, s'^j \right\rangle = \mathrm{Id}$. Moreover, from condition $(\mathcal{R})$, we have $s'^j \in \mathcal{S}(\mathbb{R}, \mathbb{R}^p)$, and therefore

$$\mathbb{E}\left[\left\|R^{j-1} - s'^j(Y^j)\right\|^2\right] < \mathbb{E}\left[\left\|R^{j-1} - s^j(Y^j)\right\|^2\right],$$

since R^{j-1} is centered. This contradicts the minimality property of s^j. As a conclusion, $\mathbb{E}\left[R^j\right] = -\mathbb{E}\left[s^j(Y^j)\right] = 0$.

(ii) and (iii) The proof is also by induction on j. Note H_j the hypothesis $\forall k \leq j$, $\left\langle a^k, R^j \right\rangle = 0$. H_1 is true since

$$\left\langle a^1, R^1 \right\rangle = \left\langle a^1, R^0 \right\rangle - \left\langle a^1, s^1(Y^1) \right\rangle = Y^1 - Y^1 = 0.$$

Supposing H_{j-1} is true, we now prove H_j. For $k = j$, we have

$$\left\langle a^j, R^j \right\rangle = \left\langle a^j, R^{j-1} \right\rangle - \left\langle a^j, s^j(Y^j) \right\rangle = Y^j - Y^j = 0.$$

For $k < j$, H_{j-1} yields:

$$\left\langle a^k, R^j \right\rangle = \left\langle a^k, R^{j-1} \right\rangle - \left\langle a^k, s^j(Y^j) \right\rangle = \left\langle a^k, s^j(Y^j) \right\rangle.$$

Now, s^j is a solution of step [R] and thus minimizes

$$\left\|R^{j-1} - s^j(Y^j)\right\|^2 = \left\langle a^k, R^{j-1} - s^j(Y^j) \right\rangle^2 + \sum_{i \neq k} \left\langle a^i, R^{j-1} - s^j(Y^j) \right\rangle^2.$$

From H_{j-1} and $(\mathcal{R})$, the minimum is reached for a function s^j such that $\left\langle a^k, s^j \right\rangle = 0$ (the proof is done by contradiction as in (i)). To conclude, $\left\langle a^k, R^j \right\rangle = 0$ and $\left\langle a^k, s^j \right\rangle = 0$, which both prove H_j and (iii).

(iv) Consider $j \geq 1$ and $s'^j \in \mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ given by $s'^j = \left\langle s^j, a^j \right\rangle a^j$. We have

$$\begin{aligned}
\left\|R^j\right\|^2 &= \left\|R^{j-1} - s^j(Y^j)\right\|^2 \\
&\leq \left\|R^{j-1} - s'^j(Y^j)\right\|^2 \\
&= \sum_{k=1}^{j-1} \left\langle a^k, R^{j-1} - s'^j(Y^j) \right\rangle^2 + \left\langle a^j, R^{j-1} - s'^j(Y^j) \right\rangle^2 \\
&+ \sum_{k=j+1}^{p} \left\langle a^k, R^{j-1} - s'^j(Y^j) \right\rangle^2.
\end{aligned}$$

The first term is null in view of (ii). Condition $\langle a^j, s^j \rangle = \text{Id}$ entails that the second term is null too. Finally, in view of s'^j definition:

$$\left\|R^j\right\|^2 \leq \sum_{k=j+1}^{p} \left\langle a^k, R^{j-1} - s'^j(Y^j)\right\rangle^2 = \sum_{k=j+1}^{p} \left\langle a^k, R^{j-1}\right\rangle^2 \leq \left\|R^{j-1}\right\|^2.$$

▮

The proof of Theorem 1 is now straightforward. It only remains to show that $R^p = 0$ $\mathbb{P}$-a.s. In view of (ii) and Proposition 1, R^p is orthogonal to a $\mathbb{R}^p$ basis, and therefore it is $\mathbb{P}$-a.s null. The following corollary will reveal useful to select the model dimension similarly to the PCA case.

Corollary 1. *Let Q_d be the information ratio represented by the d-dimensional AA model:*

$$Q_d = 1 - \mathbb{E}\left[\left\|R^d\right\|^2\right] \Big/ \text{Var}\left[\left\|X\right\|^2\right].$$

Then, $Q_0 = 0$, $Q_p = 1$ and the sequence (Q_d) is non decreasing.

3. TWO PARTICULAR AUTO-ASSOCIATIVE MODELS.

We consider two important cases in practice where step [R] has an explicit solution: the linear auto-associative models (LAA) and the auto-associative regression models (AAR). Clearly, these models inherit from the properties established in the previous section. In both cases, we precise these general properties by giving some further characteristics.

3.1. Linear auto-associative models

We focus on the case where $\mathcal{S}(\mathbb{R}, \mathbb{R}^p) = \mathcal{A}(\mathbb{R}, \mathbb{R}^p)$. From Proposition 1(i), it is straightforward that we can restrict ourselves to linear regression functions s i.e. such that $s(t) = tb$, $t \in \mathbb{R}$, $b \in \mathbb{R}^p$. Thus, step [R] can be rewritten as:

[R] Find $b^j = \arg\min_{x \in \mathbb{R}^p} \mathbb{E}\left[\|R^{j-1} - Y^j x\|^2\right]$, u.c. $\langle a^j, x\rangle = 1$,

and a result similar to Theorem 1 holds:

Theorem 2. *Algorithm 2 builds a d-dimensional LAA model with regression functions $s^j(t) = tb^j$. Moreover, for $d = p$, the following expansion holds:*

$$X = \mathbb{E}[X] + \sum_{k=1}^{p} Y^k b^k, \quad \mathbb{P} - a.s.$$

and the principal variables Y^k, $k = 1, \ldots, p$ are orthogonal.

We first prove the following properties.

PROPOSITION 2. *Let Σ^j be the covariance matrix of R^j. The regression functions and the principal variables obtained with Algorithm 2 share the following properties :*

(i) For all $1 \le j \le d$, $b^j = \Sigma^{j-1} a^j / ({}^t a^j \Sigma^{j-1} a^j)$.

(ii) For all $1 \le i < j \le p$, $\mathbb{E}\left[Y^i Y^j\right] = 0$.

Proof.

(i) Let $\mathcal{L}(x,\lambda)$ be the Lagrangian associated to the minimization problem of step [R]:

$$\mathcal{L}(x,\lambda) = \mathbb{E}\left[\left\|R^{j-1} - Y^j x\right\|^2\right] + \lambda\left(\langle a^j, x\rangle - 1\right).$$

Requiring the gradient to vanish at point x, we obtain the equation

$$2\mathbb{E}\left[R^{j-1}Y^j\right] - 2x\mathbb{E}\left[{Y^j}^2\right] + \lambda a^j = 0,$$

and projecting on the axis a^j, it yields $\lambda = 0$ leading to

$$b^j = \mathbb{E}\left[R^{j-1}Y^j\right] / \mathbb{E}\left[{Y^j}^2\right] = \Sigma^{j-1} a^j / ({}^t a^j \Sigma^{j-1} a^j).$$

(ii) The result is proved by induction by noting H_k : $\mathbb{E}\left[Y^i Y^j\right] = 0$, $1 \le i < j \le k$. H_1 is straightforwardly true. Let us suppose that H_k is true and prove H_{k+1}. The random vector X can be expanded as :

$$X = \mathbb{E}[X] + \sum_{i=1}^{k} Y^i b^i + R^k. \tag{1}$$

Hence, by projection,

$$\left\langle X - \mathbb{E}[X], a^{k+1}\right\rangle = \sum_{i=1}^{k} Y^i \left\langle b^i, a^{k+1}\right\rangle + Y^{k+1},$$

and for $1 \le j < k+1$ we thus obtain:

$$\begin{aligned}
\mathbb{E}\left[Y^j Y^{k+1}\right] &= \mathbb{E}\left[Y^j \left\langle X - \mathbb{E}[X], a^{k+1}\right\rangle\right] - \sum_{i=1}^{k} \mathbb{E}\left[Y^i Y^j\right] \left\langle b^i, a^{k+1}\right\rangle \\
&= \mathbb{E}\left[Y^j \left\langle X - \mathbb{E}[X], a^{k+1}\right\rangle\right] - \mathbb{E}\left[{Y^j}^2\right] \left\langle b^j, a^{k+1}\right\rangle,
\end{aligned}$$

by H_k. Taking into account of (i), we have $b^j = \mathbb{E}\left[R^{j-1}Y^j\right] / \mathbb{E}\left[{Y^j}^2\right]$, and consequently,

$$\mathbb{E}\left[Y^j Y^{k+1}\right] = \mathbb{E}\left[Y^j \left\langle a^{k+1}, X - \mathbb{E}[X] - R^{j-1}\right\rangle\right].$$

An expansion similar to (1) yields

$$X - \mathbb{E}\left[X\right] - R^{j-1} = \sum_{i=1}^{j-1} Y^i b^i,$$

and then

$$\mathbb{E}\left[Y^j Y^{k+1}\right] = \sum_{i=1}^{j-1} \mathbb{E}\left[Y^i Y^j\right] \left\langle a^{k+1}, b^i \right\rangle = 0.$$

by H_k since $j-1 < k$.

▮

Theorem 2 is then a consequence of Theorem 1 and Proposition 2. Let us note that, from part (i) of the proof, the constraint of step [R] is always satisfied and thus inactive.
It appears from Theorem 2 that the limitation to a family of linear functions allows to recover an important property of PCA models: the orthogonality of the principal variables. It is now shown that Algorithm 2 can also compute a PCA model for a well suited choice of index.

Corollary 2. *If, moreover, the index I of step [A] is the projected variance, i.e.* $I(\left\langle x, R^{j-1} \right\rangle) = \mathrm{Var}\left[\left\langle x, R^{j-1} \right\rangle\right]$*, then Algorithm 2 computes the PCA model of* X*.*

Proof. It is well-known that the solution a^j of step [A] is the eigenvector associated to the maximum eigenvalue λ_j of Σ^{j-1}. From Proposition 2(i), we then obtain $b^j = a^j$. Introducing $A^j = a^j\,{}^t a^j$, we consider the induction hypothesis

$$H_k:\ \Sigma^k = \Sigma^0 - \sum_{j=1}^{k} \lambda_j A^j,\ R^k = R^0 - \sum_{j=1}^{k} A^j R^0.$$

H_0 is straightforwardly true. Supposing H_k holds, we now prove that H_{k+1} is also true. We have on one hand :

$$R^{k+1} = R^k - \left\langle a^{k+1}, R^k \right\rangle a^{k+1} = R^k - \left\langle a^{k+1}, X \right\rangle a^{k+1},$$

and on the other hand :

$$\Sigma^{k+1} = \Sigma^k +^t a^{k+1} \Sigma^k a^{k+1} A^{k+1} - A^{k+1}\Sigma^k - \Sigma^k A^{k+1} = \Sigma^k - \lambda_{k+1} A^{k+1},$$

and thus H_{k+1} is true. It yields

$$\lambda_{k+1} a^{k+1} = \Sigma^k a^{k+1} = \Sigma^0 a^{k+1} - \sum_{j=1}^{k} \lambda_j \left\langle a^j, a^{k+1} \right\rangle a^j = \Sigma^0 a^{k+1},$$

which proves that a^{k+1} is also an eigenvector of Σ^0 associated to the eigenvalue λ_{k+1}. Introducing Jordan's expansion

$$\Sigma^0 = \sum_{k=1}^{d} \lambda_k A^k,$$

we deduce from H_d that $\Sigma^d = 0$ and thus that R^d is almost surely constant. Since the residuals are centered, it follows that $R^d = 0$, $\mathbb{P}$-a.s. and

$$X = \mathbb{E}\left[X\right] + \sum_{k=1}^{d} \left\langle a^k, X - \mathbb{E}\left[X\right]\right\rangle a^k, \;\; \mathbb{P} - a.s. \tag{2}$$

which is the expansion produced by a PCA. ∎

Let us note that the auto-associative function F associated to a PCA by (2) is linear. It is possible to show that, conversely, PCA is the only AA model associated to a linear function F [15].

3.2. Auto-associative regression models

Herein, we consider the case where $\mathcal{S}(\mathbb{R}, \mathbb{R}^p) = \mathrm{L}^2_X(\mathbb{R}, \mathbb{R}^p)$ leading to an explicit solution for step [R]:

[R] $s^j(Y^j) = \mathbb{E}\left[R^{j-1} | Y^j\right]$,

since the conditional expectation is an orthogonal projector in L^2_X meeting the constraint. We thus have the following result:

Theorem 3. *Algorithm 2 builds a d-dimensional auto-associative model. Moreover, when $d = p$, the following expansion holds:*

$$X = \mathbb{E}\left[X\right] + \sum_{j=1}^{p} s^j(Y^j), \;\; \mathbb{P} - a.s.$$

where the principal variables Y^j et Y^{j+1} are orthogonal, $j = 1, \ldots, p-1$.

We first prove the following proposition:

Proposition 3. *The residuals and the principal variables obtained with Algorithm 2 share the following properties :*

(i) For all $1 \le j \le d$, $\mathbb{E}\left[R^j \middle| Y^j\right] = 0$, $\mathbb{P}$-a.s.

(ii) For all $1 \le j < d$, $\mathbb{E}\left[Y^j Y^{j+1}\right] = 0$.

Proof.

(i) Since $R^j = R^{j-1} - s^j(Y^j)$, we have $\mathbb{E}\left[R^j \,\middle|Y^j\right] = \mathbb{E}\left[R^{j-1} \,\middle|Y^j\right] - \mathbb{E}\left[s^j(Y^j) \,\middle|Y^j\right]$ and consequently $\mathbb{E}\left[R^j | Y^j\right] = 0$, $\mathbb{P}$-a.s.

(ii) We have $\mathbb{E}\left[Y^j Y^{j+1}\right] = \mathbb{E}\left[Y^j \left\langle a^{j+1}, R^j \right\rangle\right] = \mathbb{E}\left[Y^j \left\langle a^{j+1}, \mathbb{E}\left[R^j | Y^j\right] \right\rangle\right] = 0$ from (i).

▮

Theorem 3 is a direct consequence of Theorem 1 and Proposition 3(ii). The choice $\mathcal{S}(\mathbb{R}, \mathbb{R}^p) = \mathrm{L}^2_X(\mathbb{R}, \mathbb{R}^p)$ provides then a convenient framework to propose a nonlinear PCA with interesting theoretical properties (Theorem 3) and a simple computation scheme (Algorithm 2). The implementation aspects are discussed in the next section.

4. IMPLEMENTATION.

Consider a sample $(X_1, \dots, X_n)$ iid from the unknown distribution $\mathbb{P}_X$. The parameter μ is estimated by the empirical mean $\bar{X} = 1/n \sum X_i$. The two crucial steps in Algorithm 2 are [A] and [R]: the determination of the principal directions and the estimation of the regression functions. The index I and the set of functions $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ both determine the nature of the obtained model and the complexity of the computation associated to the optimization problems [A] and [R].

4.1. Estimation of the regression function

Remark that, when $\mathcal{S}(\mathbb{R}, \mathbb{R}^p)$ is the set of affine functions from $\mathbb{R}$ to $\mathbb{R}^p$ then, from Proposition 2, $b^j = \Sigma^{j-1} a^j / ({}^t a^j \Sigma^{j-1} a^j)$ where Σ^{j-1} is the covariance matrix of the residual R^{j-1}. Then, b^j is estimated by replacing in the above formula Σ^{j-1} by its empirical estimate and a^j by the estimation obtained at step [A].

In the case of AAR models, the problem reduces to estimating the conditional expectation of R^{j-1} given Y^j. This standard problem [18] can be tackled by kernel [2] or spline [16] regression estimates.

Here, a kernel estimate has been chosen to deal with the simulated and real data. For an example of the use of spline regression in a similar context, we refer to [5]. Compared to a classical regression problem, an additional constraint on the function to estimate at the j-th iteration: $P_{a^j} s^j = \mathrm{Id}$ has to be taken into account. Fortunately, in the orthogonal basis B^j of $\mathbb{R}^p$ obtained by completing $\{a^1, \dots, a^j\}$, step [R] reduces to $(p-j)$ independent regressions. Hence, each coordinate $k \in \{j+1, \dots, p\}$ of the estimate can be written in the basis B^j as:

$$\tilde{s}^j_k(u) = \sum_{i=1}^n \tilde{R}^{j-1}_{i,k} K_h(u - Y^j_i) \Bigg/ \sum_{i=1}^n K_h(u - Y^j_i) \,, \tag{3}$$

where $\tilde{R}^{j-1}_{i,k}$ represents the k-th coordinate of the residual of the observation i at the $(j-1)$-th iteration in the basis B^j, Y^j_i is the value of the j-th principal variable for the observation i and the kernel K_h is for example a centered Gaussian density with standard deviation h, called window in this context. More generally, any Parzen-Rosenblatt kernel is convenient. For an automatic choice of h, we refer to [20], Chapter 6.

4.2. Computation of principal directions

The choice of the index I is the key point of any Projection pursuit problem where it is needed to find "interesting" directions. We refer to [22] and [24] for a review on this topic. The meaning of the word "interesting" depends on the considered data analysis problem. For instance, Friedman *et al* [11, 13], and more recently Hall [17], proposed an index to find clusters or use deviation from the normality measures to reveal more complex structures of the scatter-plot. An alternative approach can be found in [4] where a particular metric is introduced in PCA so as to detect clusters. We can also mention indices dedicated to outliers detection [29].

In the framework of AAR models, we are interested in finding parameterization directions for the manifold to be estimated. In this aim, Demartines and Herault [8] introduce an index to detect the directions in which the projection approximatively preserves distances. From a similar principle, Girard [5] proposes an index revealing the directions in which the neighborhood structure is invariant with respect to projection. Both criteria require complex optimization algorithms. For instance, in [5], the optimization step [A] is performed with a simulated annealing technique, leading to a data analysis procedure heavy to use in practice.

Our approach is similar to Lebart one's [26]. It consists in defining a contiguity coefficient whose minimization allows to unfold nonlinear structures. At each iteration j, the following ratio of quadratic functions is maximized with respect to x:

$$I(\langle x, R^{j-1}\rangle) = \sum_{i=1}^{n} \left\langle x, R_i^{j-1} \right\rangle^2 \Bigg/ \sum_{k=1}^{n}\sum_{\ell=1}^{n} m_{k\ell} \left\langle x, R_k^{j-1} - R_\ell^{j-1} \right\rangle^2 . \quad (4)$$

The matrix $M = (m_{k\ell})$ is a first order contiguity matrix, whose value is 1 when R_ℓ^{j-1} is the nearest neighbor of R_k^{j-1}, 0 otherwise. The upper part of (4) is the usual projected variance. The lower part is the distance between the projection of points which are nearest neighbor in $\mathbb{R}^p$. Then, the maximization of (4) should reveal directions in which the projection best preserves the first order neighborhood structure. In this sense, the index (4) can be seen as a first order approximation of the index proposed in [5]. Thanks to this approximation, the maximization step benefits from an explicit solution: The resulting principal direction a^j is the eigenvector associated to the maximum eigenvalue of $V_j^{\star -1} V_j$ where

$$V_j^\star = \sum_{k=1}^{n}\sum_{\ell=1}^{n} m_{k\ell}{}^t(R_k^{j-1} - R_\ell^{j-1})(R_k^{j-1} - R_\ell^{j-1})$$

is proportional to the local covariance matrix. The matrix

$$V_j = \sum_{k=1}^{n} {}^t R_k^{j-1} R_k^{j-1}$$

is proportional to the empirical covariance matrix of R^{j-1}. $V_j^{\star -1}$ should be read as the generalized inverse of the singular matrix $V_j^{\star}$, R^j being orthogonal to $\{a^1, \dots, a^j\}$ from Proposition 1(ii). Note that this approach is equivalent to Lebart's one when the contiguity matrix M is symmetric.

5. EXAMPLES.

We first present two illustrations of the AAR models construction principle on low dimensional data (Section 5.1 and 5.2). Second, AAR models are applied to an image analysis problem in Section 5.3. In all cases, the principal directions are computed thanks to the contiguity index (4). A Gaussian kernel method (3) is applied in the regression step [R].

5.1. First example on simulated data

The data are simulated from a distribution whose support is a one-dimensional manifold in $\mathbb{R}^3$. The equation of the manifold is given by

$$x \to (x, \sin x, \cos x). \tag{5}$$

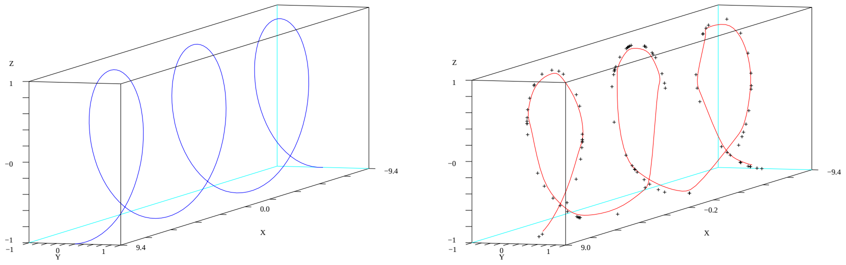
The first coordinate of the random vector is uniformly distributed on $[-3\pi, 3\pi]$ and $n = 100$ points are simulated. One iteration of Algorithm 2 is used. The squared cosine between the natural axis of parameterization (the x-axis) and the axis estimated at step [A] is as high as 0.998. The window of the kernel estimate is chosen equal to $h = 0.3$. At the end of the first iteration, the information ratio is $Q_1 = 99.97\%$. The theoretical manifold, the simulated scatter-plot and the estimated manifold are presented on Figure 1 for comparison.

5.2. Second example on simulated data

The data are simulated from a distribution whose support is a two-dimensional manifold in $\mathbb{R}^3$. The equation of the manifold is given by

$$(x, y) \to \left(x, y, \cos(\pi\sqrt{x^2+y^2})(1 - \exp\{-64(x^2+y^2)\})\right). \tag{6}$$

The first two coordinates of the random vector are uniformly distributed on $[-1/2, 1/2] \times [-1, 1]$ and $n = 1000$ points are simulated.
We limit ourselves to two iterations. The squared cosine between the first natural axis of parameterization (the y-axis) and the first estimated axis a^1 is as high as 0.998 and the squared cosine between the second natural axis



(a) (b)

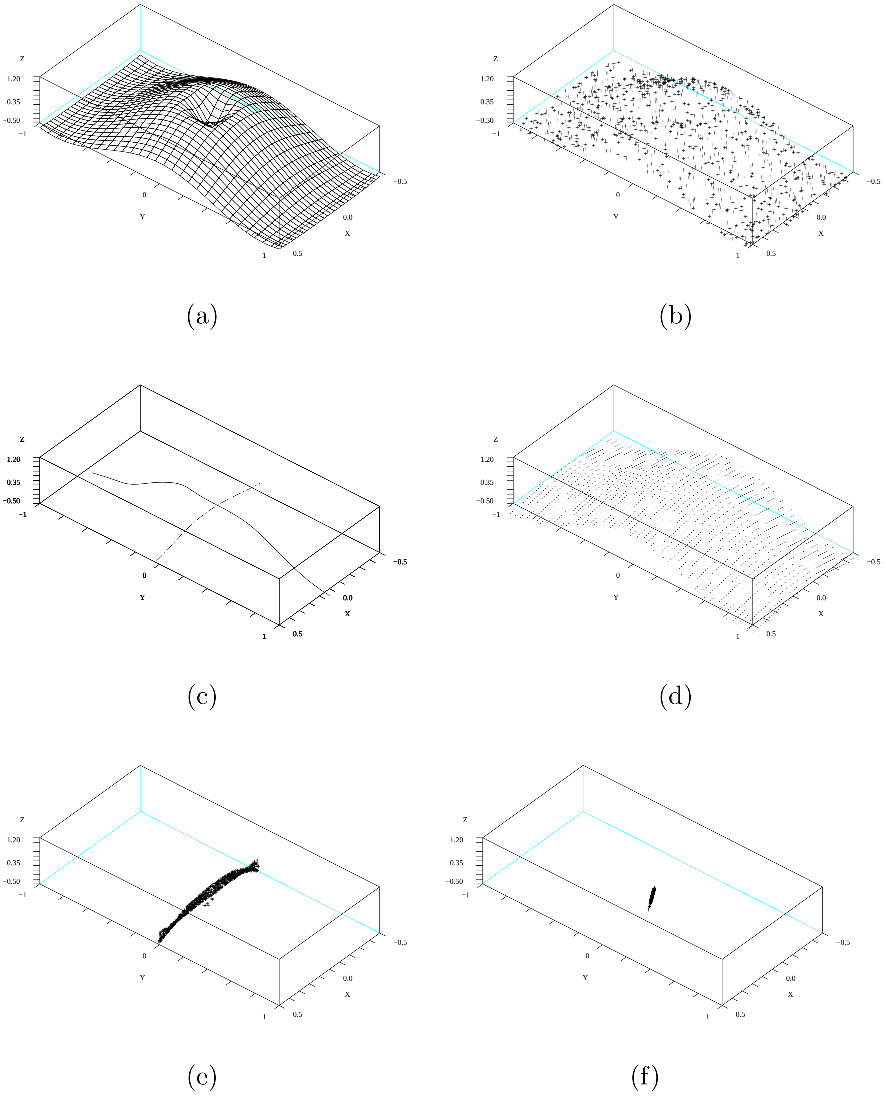
FIG. 1 Manifold (a), simulated scatter-plot and estimated manifold (b).

of parameterization (the x-axis) and the second estimated axis a^2 is 0.999. The window of the kernel estimate is chosen equal to $h = 0.12$. After the first and second iterations, the information ratio is respectively equal to $Q_1 = 84.1\%$ and $Q_2 = 97.6\%$.

The manifold (6) and the simulated scatter-plot are depicted in Figure 2(a)–(b). The first regression function s^1 is plotted on Figure 2(c) with a solid line. It approximatively represents the shape of the scatter-plot in the y-direction. It can be noted that it does not take into account of the hole induced by the exponential function. The corresponding residuals (at the end of the first iteration) are represented on Figure 2(e). Remark that, accordingly to Proposition 1(ii), they are orthogonal to the first principal direction a^1. The second regression function is drawn with a dashed line on Figure 2(c). Figure 2(d) shows the estimated manifold after two iterations. The associated residuals are represented on Figure 2(f). They are orthogonal to the two principal directions a^1 and a^2. In fact, they are a consequence of the poor reconstruction of the hole due to the non additive structure of the manifold equation (6).

5.3. Example in image analysis

Image analysis is a natural application field for multivariate analysis [6], since an image with $M \times M$ pixels can be represented by a vector of $\mathbb{R}^p$ with $p = M^2$. Even images of moderate size yield data in spaces of extremely large dimension. PCA usually is an accurate tool to reduce the dimension of such data [28, 34]. However, even some very simple deformation in the image space can lead to important nonlinearities in $\mathbb{R}^p$. In such situations, PCA efficiency is significantly decreased. This remark is the starting point of Capelli *et al* [3] work who propose a "piecewise" PCA. The idea is to split the nonlinear structure of $\mathbb{R}^p$ into approximatively linear sub-structures. We study here a database of 45 images of size 256 × 256 taken from the



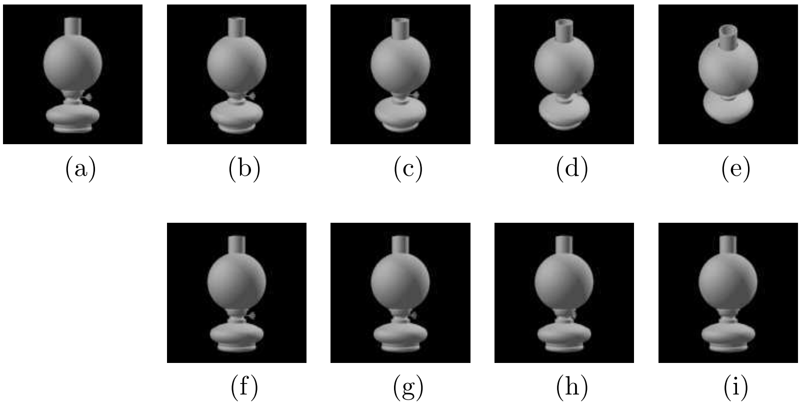
(a) (b)

(c) (d)

(e) (f)

FIG. 2 Manifold (a), simulated scatter-plot (b), regression functions (c), estimated two-dimensional manifold (d), residuals after the first iteration (e) and the second one (f).

archive of Centre For Intelligent Systems, Faculty of Human Sciences and Faculty of Technology, University of Plymouth. It is made up with images of a synthesis object viewed under different elevation and azimuth angles. A sample from the database is presented on Figure 3.


(a) (b) (c) (d) (e)

(f) (g) (h) (i)

FIG. 3 A sample from the image database. (a) reference image, (b-e) rotation using the elevation angle, (f-i) rotation using the azimuth angle.

Each image is represented by a vector of dimension $M^2 = 256^2$ leading to a scatter-plot of $n = 45$ points in dimension 65536. However, a simple rotation of axes allows to represent this set of points in dimension $p = 44$. In the following, our aim is to compare the modeling results obtained by a classical PCA and by AAR models. The smoothing parameter is fixed to $h = 200$. Figure 4 shows the compared information percentage $100Q_d$ represented by AAR and by PCA models of increasing dimension $d = 0, \ldots, 10$ (see Corollary 1).

The one-dimensional AAR model allows to represent more than 96% of the information. As a comparison, a linear model built by PCA should be of dimension 4 to reach this percentage. Moreover, the elbow in the curve associated to AAR models seems to indicate that $d = 1$ is a convenient choice. The projection of the corresponding manifold in the linear subspace spanned by the first three PCA axes is represented on Figure 5(a) where it is superimposed to the scatter-plot projection. Modeling this scatter-plot by a two-dimensional manifold could also be justified since the image database is generated by rotating the object in two orthogonal directions. The projection of the two-dimensional manifold estimated and sampled is presented on Figure 5(b).

It is worth remarking that the principal variable Y^1 associated to the one-dimensional AAR model has a simple interpretation. It corresponds to the rotation with respect to the elevation angle. As an illustration, we simulate uniform realizations of this variable and represent the corresponding images obtained with the one-dimensional AAR model (Figure 6). The variable

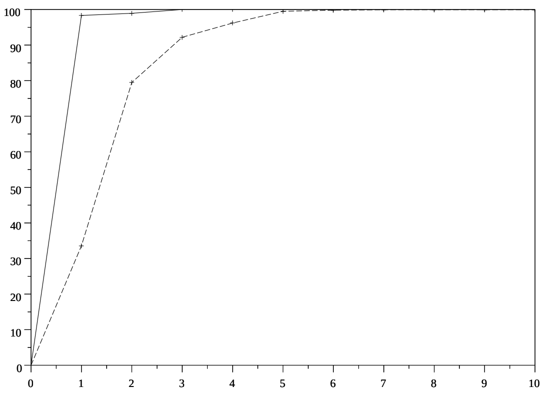
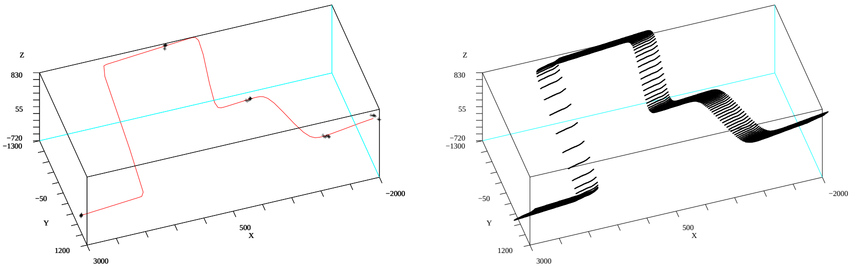


FIG. 4 Percentage of represented information as a function of the model dimension (dashed line: PCA model, solid line: AAR model).

Y^2 is not so easily interpretable. For this reason, the one-dimensional AAR model should be preferred.

6. CONCLUSION AND FURTHER WORK.

As a conclusion, AA models offer a nice theoretical framework to the generalization of PCA. They extend the main PCA properties while offering more flexibility: projection indices and regression functions available in the literature allow to build numerous data analysis methods. Moreover, all these methods would benefit from a simple implementation thanks to the proposed iterative algorithm. In this paper, this principle is illustrated by building AAR models combined with a contiguity index on simulated and real data from image analysis. The resulting method is computationally efficient since it does not require any optimization procedure, neither for the Projection pursuit step, nor for the regression step. Possible extensions of this work involve practical aspects and theoretical research. On the practical point of view, it would be interesting to compare, at least visually, the mapping produced by LLE [31] and Isomap [33] methods to the scatter-plot of principal variables associated to AA models. From a theoretical point of view, we plan to establish the asymptotic properties of the estimates (3) and (4) in order to build tests on the model dimension. The generalization properties of AA models should be investigated. In this aim, it would be necessary to introduce a criterion measuring the distance between simulations from the AA model (see for instance Figure 6) and



(a) (b)

FIG. 5 Projections in the subspace spanned by the first three PCA axes: (a) the one-dimensional manifold estimated and superimposed to the scatter-plot, (b) the two-dimensional manifold estimated and sampled.

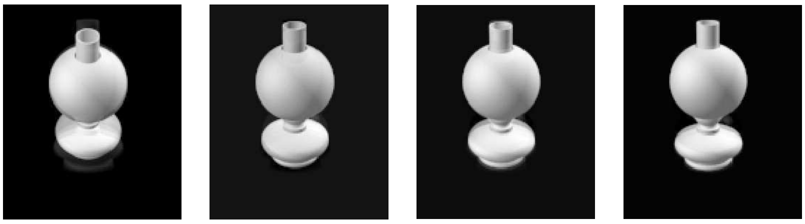


FIG. 6 Simulation of 4 images with the one-dimensional AAR model. The variable Y^1 is simulated uniformly on the interval $[\min_i Y_i^1, \max_i Y_i^1]$.

the original sample (Figure 3). This distance would provide a good tool for the adaptive choice of the smoothing parameters.

ACKNOWLEDGMENTS

The authors thank an anonymous referee for making them aware of LLE and Isomap references.

REFERENCES

[1] P. Besse & F. Ferraty, (1995). ”A fixed effect curvilinear model”, *Computational Statistics*, 10(4), p. 339–351.

[2] D. Bosq & J.P. Lecoutre, (1987). *Théorie de l’estimation fonctionnelle*, Economica, Paris.

[3] R. Capelli, D. Maio & D. Maltoni, (2001). "Multispace KL for pattern representation and classification", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(9), p. 977–996.

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

[5] B. Chalmond & S. Girard, (1999). "Nonlinear modeling of scattered multivariate data and its application to shape change", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(5), p. 422–432.

[6] B. Chalmond, (2002). *Modeling and inverse problems in image analysis*, Applied Mathematics Science series, 155, Springer, New-York.

[7] P. Delicado, (2001). "Another look at Principal curves and surfaces", *Journal of Multivariate Analysis*, 77, p. 84–116.

[8] P. Demartines & J. Hérault, (1997). "Curvilinear Component Analysis: A self-organizing neural network for nonlinear mapping of data sets", *IEEE Trans. on Neural Networks*, 8(1), p. 148–154.

[9] D.L. Donoho & C. Grimes, (2003). "Local ISOMAP perfectly recovers the underlying parametrization for families of occluded/lacunary images", *IEEE Computer Vision and Pattern Recognition*, Madison, Wisconsin, June 16–22.

[10] J.F. Durand, (1993). "Generalized principal component analysis with respect to instrumental variables via univariate spline transformations", *Computational Statistics and Data Analysis*, 16, p. 423–440.

[11] J.H. Friedman & J.W. Tukey, (1974). "A Projection Pursuit algorithm for exploratory data analysis", *IEEE Trans. on Computers*, C23 (9), p. 881–890.

[12] J.H. Friedman & W. Stuetzle, (1981). "Projection Pursuit Regression", *Journal of the American Statistical Association*, 76 (376), p. 817–823.

[13] J.H. Friedman, (1987). "Exploratory Projection Pursuit", *Journal of the American Statistical Association*, 82 (397), p. 249–266.

[14] S. Girard, (2000). "A nonlinear PCA based on manifold approximation", *Computational Statistics*, 15(2), p. 145–167.

[15] S. Girard, B. Chalmond & J-M. Dinten, (1998). "Position of Principal component analysis among auto-associative composite models", *Comptes Rendus de l'Académie des Sciences de Paris*, t.326, Série 1, p. 763–768.

[16] P.J. Green & B.W. Silverman, (1994). *Non-parametric regression and generalized linear models*, Chapman and Hall, London.

[17] P. Hall, (1990). "On polynomial-based projection indices for exploratory projection pursuit", *The Annals of Statistics*, 17(2) p. 589–605.

[18] W. Härdle, (1990). *Applied nonparametric regression*, Cambridge University Press, Cambridge.

[19] T. Hastie & W. Stuetzle, (1989). "Principal curves", *Journal of the American Statistical Association*, 84 (406), p. 502–516.

[20] T. Hastie, R. Tibshinari & J. Friedman, (2001). *The elements of statistical learning*, Springer Series in Statistics, Springer.

[21] H. Hotelling, (1933). "Analysis of a complex of statistical variables into principal components", *Journal of Educational Psychology*, 24, p. 417–441.

[22] P.J. Huber, (1985). "Projection Pursuit". *The Annals of Statistics*, 13(2), p. 435–475.

[23] I. Jolliffe, (1986). *Principal Component Analysis*, Springer-Verlag, New York.

[24] M.C. Jones & R. Sibson, (1987). "What is projection pursuit?", *Journal of the Royal Statistical Society, Ser. A*, 150, p. 1–36.

[25] S. Klinke & J. Grassmann, (2000). "Projection pursuit regression", *Wiley Series in Probability and Statistics*, p. 471–496.

[26] L. Lebart, (2000). "*Contiguity analysis and classification*", Data Analysis, Gaul W., Opitz O., Schader M., (eds), Springer, Berlin, p. 233–244.

[27] J.A. Lee, A. Lendasse & M. Verleysen, (2002). "Curvilinear distance analysis versus Isomap", *European Symposium on Artifical Neural Networks*, Bruges, Belgium, April 24–26, p. 185–192.

[28] B. Moghaddam & A. Pentland, (1997). "Probabilistic visual learning for object representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19 (7), p. 696–710.

[29] J-X. Pan, W-K. Fung & K-T. Fang, (2000). "Multiple outlier detection in multivariate data using projection pursuit techniques." *Journal of Statistical Planning and Inference*, 83(1), p. 153–167.

[30] K. Pearson, (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, p. 559–572.

[31] S.T. Roweis & L.K. Saul, (2000). "Nonlinear dimensionality reduction by locally linear embedding", *Science*, 290, p. 2323–2326.

[32] M.E. Tipping & C.M. Bishop, (1999). "Probabilistic principal component analysis", *Journal of the Royal Statistical Society, Ser. B*, 61(3), p. 611–622.

[33] J.B. Tenenbaum, V. de Silva, & J.C. Langford, (2000). "A global geometric framework for nonlinear dimensionality reduction", *Science*, 290, p. 2319–2323.

[34] M. Uenohara & T. Kanade, (1997). "Use of Fourier and Karhunen-Loeve decomposition for fast pattern matching with a large set of templates", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19 (8), p. 891–898.

[35] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios & N. Koudas, (2002). "Non-linear dimensionality reduction techniques for classification and visualization", *Proc. of 8th SIGKDD*, Edmonton, Canada, July 23–26.