Joint work with Serge Iovleff, Université Lille I

Stéphane Girard

Auto-Associative Models

And Generalized Principal Component Analysis
1. Principal Component Analysis, 2 points of view,
2. Generalized PCA, theoretical aspects,
3. Implementation aspects,
4. Illustration on simulated datasets,
5. Illustration on real datasets.
1. Principal Component Analysis

**Background**: Multidimensional data analysis

**Goal**: Dimension reduction.

- To find which variables are important,
- Data visualization (dimension less than 3),
- Compression.

**Method**: Projection on low $d$-dimensional linear subspaces.
Problem

• Let $X$ be a centered random vector in $\mathbb{R}^p$.
• Estimate the $d$-dimensional linear subspace $d \in \{0, \ldots, p\}$ minimizing the mean distance to $X$.

- Minimize with respect to $\{a_1, \ldots, a_d\}$ (orthonormal):
- Estimate the $p - d$-dimensional linear subspace $\{d' \in \mathbb{R}^p \mid d' \perp X\}$ minimizing the mean distance to $X$.

Explicit solution

- $a_1, \ldots, a_d$ are the eigenvectors associated to the $d$ largest eigenvalues of $\mathbb{E}[X^tX]$, the covariance matrix of $X$.
- The $Q$'s are called principal axes, the $Y_k = \langle X, a_k \rangle$ the principal variables.
- The associated residual is defined by

$$\|X - d \sum_{k=1}^d \langle X, a_k \rangle a_k\| \leq \|X\| - 1.$$
Equivalent problem

- Estimate the \( d \)-dimensional linear subspace \( \{d, \ldots, \} \) \( p \) \( d \) \( \text{maximizing} \) the projected variance.
- Maximize iteratively with respect to \( a_1, \ldots, a_d \) (orthonormal):

\[
\text{Var}[\langle p_i, X \rangle], \ldots, \text{Var}[\langle p_i, X \rangle].
\]

**PCA: Projection Pursuit Interpretation**
Algorithm

For \( j = 0 \) let \( R_0 = X \).

For \( j = 1, \ldots, d \):


Determine \( a_j = \arg \max_{x \in \mathbb{R}^p} \mathbb{E} [ \langle x, R_{j-1} \rangle^2 ] \) such that

\[
\|a_j\|_2 = 1 \quad \text{and} \quad \langle a_j, a_k \rangle = 0, \quad 1 \leq k < j.
\]

[B] Linear regression.

Determine \( b_j = \arg \min_{x \in \mathbb{R}^p} \mathbb{E} \left[ \| x - R_{j-1} - s_j(Y_j) \|_2^2 \right] \) such that

\[
\langle b_j, a_j \rangle = 1 \quad \text{and} \quad \langle b_j, a_k \rangle = 0, \quad 1 \leq k < j.
\]

The solution is \( b_j = a_j \), and let the regression function be

\[
s_j(t) = ta_j.
\]

[C] Residual update.

Compute \( R_j = R_{j-1} - s_j(Y_j) \).

\( \cdots \)  \( \forall \ j \geq 1 \quad 0 = \langle \gamma, \gamma \rangle \) and let the regression function be

\( \ell_j = \langle \gamma, \ell \rangle \). The solution is \( \ell > \gamma \geq 1 \)  \( \ell \). The solution is \( \ell_j = \ell \).

For \( j = 1, \ldots, \ell \) let \( R_0 = 0 \).
Algorithm output.
After $p$ iterations, we have the following expansion:

$$X = \sum_{k=1}^{d} s_k(Y_k) + R_d, \quad (1)$$

with

$$s_k(t) = t a_k$$

and

$$Y_k = \langle a_k, X \rangle$$

This equation can be rewritten as

$$pH = (X)_d$$

$$pH + \sum_{p} \langle X', y_p \rangle = X$$

with

$$\langle X', y_p \rangle = \lambda \langle X', y_p \rangle$$

and

$$qH = (q)_d$$

$$pH + (q)_d \sum_{p} = X$$

Equation (2) defines a $d$-dimensional linear auto-associative model for $X$.

Equation (1) defines a $p$-dimensional linear model for $X$.

Equation (2) defines a $d$-dimensional linear subspace, spanned by $a_1, \ldots, a_d$.

Equation (2) defines a $d$-dimensional linear auto-associative model for $X$.
Goals of a generalized PCA

1. To keep a simple iterative algorithm.
2. To keep an expansion “principal variables + residual” similar to (1):
   \[
   \mathbf{pH} + \sum_{j=1}^{q} s_j \mathbf{Y}_j = \mathbf{X}
   \]
   but with non necessarily linear functions \(s_j\).
3. To benefit from the “nice” theoretical properties of PCA.
4. To keep a simple iterative algorithm.
2. Generalized PCA, theoretical aspects

We adopt the Projection Pursuit point of view. The steps [A] and [R] are generalized:

- Estimation of a projection axis.
- Outliers detection.
- Clusters detection.
- Deviation from normality.
- Dispersion.

Introduction of an index \( I \) which measures the quality of the projection axis, for instance:

- Dispersion, Kernel, etc.
- Linear functions.

Estimation of the regression function from \( \mathbb{R} \) to \( \mathbb{R}^p \) in a given set:

- Linear functions,
- Splines, kernels, etc.
Auto-Associative models and generalized Principal Component Analysis August 2006

New algorithm.

\[ 0 = \rho s \circ \rho^p \text{ and } \rho = \rho s \circ \rho^p \text{ such that } \rho^p(\langle \rho X, R^{-1} \rangle - |) = |\text{ and } \langle \rho^p, \rho^p \rangle = 1 \text{ and } \langle \rho^p, \rho^k \rangle = 0, 1 \leq k < j. \]

\[ \rho^p, \rho^p \text{ is a projection axis.} \]

\[ \rho^p, \rho^p \text{ is a regression.} \]

\[ \rho^p = \rho X \text{ is the principal variable.} \]

\[ \rho^p = \rho X \text{ is a residual update.} \]

For \( j \geq 1 \), \( j \geq 1 \), and \( X = 0 \), let \( \rho^p \).

For \( j = 1 \), let \( \rho^p \).

\[ \rho^p = \rho X \text{ is a new algorithm.} \]
Auto-associative composite model:

\[ (X) \left( \right. \right. (^{i}D \circ \iota^{S} - \iota^{P} I) \circ \cdots \circ (^{i}D \circ \iota^{S} - \iota^{P} I) = \\
\left. \left. (0) \right( (^{i}D \circ \iota^{S} - \iota^{P} I) \circ \cdots \circ (^{i}D \circ \iota^{S} - \iota^{P} I) = \\
\cdots = \\
(\varepsilon_{i} R) (1-^{o}D \circ \iota^{S} - \iota^{P} I) \circ (^{o}D \circ \iota^{S} - \iota^{P} I) = \\
(1-^{o}R) (^{o}D \circ \iota^{S} - \iota^{P} I) = \\
\left( \left. \left. \left( (1-^{o}R) (^{o}D \circ \iota^{S} - 1-^{o}R = \\
\left( \left. \left. \left( (1-^{o}R) (^{o}D \circ \iota^{S} - 1-^{o}R = \\
\left( \left. \left. \left( (1-^{o}R) (^{o}D \circ \iota^{S} - 1-^{o}R = \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) \right) (\varepsilon_{i} X) = \right)

Remark: At the end of iteration \( j \), the residual is given by
Remark:
The constraint \( P_a \circ s \circ P_a = I \)

\[ R_j = (P_a - P_a \circ s \circ P_a)(R_{j-1}) = 0. \]

Thus, iteration \( (j+1) \) can be performed on the linear subspace orthogonal to \( \langle a_1, \ldots, a_j \rangle \) which is of dimension \((l - d)\) and thus is projection on \( \mathcal{IP} \) is given by

Important consequence: At the end of iteration \( j \), the residual is given by

\( (\hat{\lambda})s = \lambda \)
After $d$ iterations:

• One always has an auto-associative model.

\[ F(X) = R_d, \]

and

\[ \langle x, \ell p \rangle = (x)^{\ell p} \]

with

\[ (\ell p \circ s - \text{id}_p) \prod_{i=1}^{p=q} = (\ell p \circ s - \text{id}_p) \circ \cdots \circ (\ell p \circ s - \text{id}_p) = F \]

\[ \ell p F = (X)F \]
After $d$ iterations:

- One always has the expansion \( \text{principal variables + residual} \) similar to (1):

\[
P = \begin{bmatrix} Y \end{bmatrix} = \sum_{k=1}^{d} s_k \left( X_k \right) + R_d,
\]

and the functions \( s_k \) are non necessarily linear.

- For \( d = p \), the expansion is exact:

\[
R_p = 0.
\]

- We can still define principal axes \( a_k \) and principal variables \( Y_k \).

- The residuals are centered:

\[
E[R_k] = 0, \quad k = 0, \ldots, d.
\]
Goal 3. After $p$ iterations, we have:

- Some orthogonality properties

Remark. Except in particular cases, the non-correlation of the principal variables is lost:

\[ \langle a_k, a_j \rangle = 0, \quad 1 \leq k < j \leq d, \]
\[ \langle a_k, R_j \rangle = 0, \quad 1 \leq k \leq j \leq d, \]
\[ \langle a_k, s_j(Y_j) \rangle = 0, \quad 1 \leq k < j \leq d. \]

Since the norm of the residuals is decreasing, we can define, similarly to the PCA case, the information ratio represented by the $p -$ dimensional model as:

\[ Q_p = 1 - \frac{\|X\|_A^2}{\|H\|_F^2} - 1 = p \mathcal{O} \]

Since the norm of the residuals is decreasing, we can define, similarly to the PCA case, the

Remark. Except in particular cases, the non-correlation of the principal variables is lost:

\[ \mathbb{E}[Y_k Y_j] \neq 0, \quad 1 \leq k < j \leq d. \]
We still have an iterative algorithm. It converges at most in $p$ steps. Its complexity depends on the two steps $\mathbf{A}$ and $\mathbf{R}$.

**Goal 4.**

Note that the above theoretical properties do not depend on these steps.

\[ 0 = \mathbf{s} \circ \phi^p \mathbf{I} \text{ and } \mathbf{s} = \mathbf{s} \circ \phi^p \mathbf{I} \text{ such that } \| \mathbf{s} \circ (\mathbf{A} - \mathbf{I}) \mathbf{s} - 1 \| \text{ is minimal such that } \mathbf{s} \circ \mathbf{a}_j = \mathbf{I} \text{ and } \langle \mathbf{a}_j, \mathbf{a}_k \rangle = 0, 1 \leq k < j. \]

Regression.

\[ \mathbf{s}_j = \text{arg min}_{\mathbf{s} \in \mathbf{S}} (\mathbf{R}, \mathbf{R}_p \mathbf{I}) \text{ such that } \mathbf{P}_{\mathbf{a}_j \circ \mathbf{s}_j} = \mathbf{I} \text{ and } \mathbf{P}_{\mathbf{a}_k \circ \mathbf{s}_j} = 0, 1 \leq k < j. \]

Estimation of a projection axis. Its complexity depends on the two steps $\mathbf{A}$ and $\mathbf{R}$.

We still have an iterative algorithm. It converges at most in $d$ steps.
Contiguity index

Measure of the neighborhood preservation. Points which are neighbor in \( \mathbb{R}^{p-1} \) should stay neighbor on the axis.

\[
\langle \mathbf{J}_{\mathbf{R}} - \mathbf{J}_{\mathbf{R}} \rangle \sum_{u} \sum_{u} = \langle \mathbf{J}_{\mathbf{R}} \rangle \sum_{u} \sum_{u} = \lambda \mathbf{J}
\]

where \( \mathbf{J} \) is the contiguity matrix defined by

\[
\mathbf{J}_{i,j} = \begin{cases} 1 & \text{if } \mathbf{R}_{j-1} \ell \text{ is the closest neighbor of } \mathbf{R}_{j-1} k \\ 0 & \text{otherwise} \end{cases}
\]

Optimization. Explicit solution.

\[
\mathbf{a}_j = \text{eigenvector associated to the largest eigenvalue of } \mathbf{V}_{j} \mathbf{V}_{j}^{-1}
\]

where \( \mathbf{V}_{j} = \sum_{k=1}^{n} \mathbf{R}_{j-1} k \mathbf{R}_{j-1} k \), \( \mathbf{V}_{j} \) is the covariance and local covariance matrices of \( \mathbf{R}_{j-1} \).
where $\eta$ is a smoothing parameter (the bandwidth):

$$\left( f^T \Lambda - n \right)^n \sum_{i=1}^{n} f^T \Lambda \sum_{i=1}^{n} \left( f^T \Lambda - n \right)^n = \left( n \right)^n \eta$$

written as

\[ \sum_{i=1}^{n} \left( f^T \Lambda - n \right)^n = \left( n \right)^n \eta \]

For instance, for the coordinates $\eta \in \{d, \ldots, 1\}$, the kernel estimates of $s_j$ can be

- numerous estimates are available: splines, local polynomials, kernel estimators, ...

- for instance, for the coordinate $k \in \{j+1, \ldots, p\}$, the kernel estimate of $s_j(u)$ can be written as

$$\tilde{s}_j(k)(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{f}(j-1,i,k)}{\sum_{i=1}^{n} K_h(u - Y_j)}$$

where $h$ is a smoothing parameter (the bandwidth).

Implementation aspects, step [R]

- Estimation of the conditional expectation.
  - $[\{f \Lambda \}_{1-f \mathcal{F}}] \mathbb{E} = (f \Lambda) \mathcal{F} [R]$

- Set of functions. The regression step reduces to estimating the conditional expectation:

$$[R]$$
4. First illustration on a simulated dataset

- \( n = 100 \) points in \( \mathbb{R}^3 \) randomly chosen on the curve \( x \rightarrow (x, \sin x, \cos x) \).

- One iteration \( h = 0.3 \rightarrow Q_1 = 99.97\% \).

Estimated 1-dimensional manifold

Theoretical curve

1. First illustration on a simulated dataset

\( \Omega \leftarrow 0.3 \) randomly drawn on the curve \( x \rightarrow (x, \sin x, \cos x) \).

\( \theta = 100 \) points in \( \mathbb{R}^3 \).
Estimated 2-dimensional manifold

Theoretical surface

\[
\begin{align*}
\text{Estimated points on the surface} & \quad \text{Second illustration on a simulated dataset}
\end{align*}
\]
First illustration on a real dataset

Set of \( n = 45 \) images of size \( 256 \times 256 \).

Interpretation: \( n = 45 \) points in dimension \( p = 256 \).

Rotation: \( n = 45 \) points in dimension \( p = 44 \).
Information ratio $Q$ as a function of $d$ (blue: classical PCA, green: generalized PCA).
Auto-Associative models and Generalized Principal Component Analysis

Projection on the 3 first PCA axes of the estimated manifolds.
Auto-Associative models and generalized Principal Component Analysis

August 2006

Second illustration on a real dataset

Dataset I, five types of breast cancer.

Set of $n = 286$ samples in dimension $p = 17816$.

Rotation: $n = 286$ points in dimension $p = 285$.

Forgetting the labels, information ratio $Q$ as a function of $p$ (blue: classical PCA, green: generalized PCA).

- $d = u = 286$ points in dimension $285$.
- $d = 17816$.
- Five types of breast cancer.

Stéphane Girard 24
Estimated $k$-dimensional manifold projected on the principal plane.
Estimated 1-dimensional manifolds projected on the principal plane for each type of cancer.


