AUTO-ASSOCIATIVE MODELS
AND GENERALIZED PRINCIPAL COMPONENT ANALYSIS

Stéphane Girard
* INRIA, Université Grenoble 1

Joint work with Serge Iovleff, Université Lille 1
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
  
  ($n$ observations in a $p$– dimensional space)

- **Goal**: Dimension reduction.
  - Data visualization (dimension less than 3),
  - To find which variables are important,
  - 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):

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

Explicit solution

- $a^1, \ldots, a^d$ are the eigenvectors associated to the $d$ largest eigenvalues of $\mathbb{E} [X^t X]$, the covariance matrix of $X$.
- The $a^k$'s are called principal axes, the $Y^k = \langle X, a^k \rangle$ the principal variables.
- The associated residual is defined by

$$R^d = X - \sum_{k=1}^{d} \langle X, a^k \rangle a^k,$$

and it can be shown that $\|R^d\| \leq \|R^{d-1}\|$.
**PCA: Projection Pursuit interpretation**

**Equivalent problem**

- Estimate the $d-$ dimensional linear subspace $d \in \{0, \ldots, p\}$ maximizing the projected variance.
- Maximize iteratively with respect to $a^1, \ldots, a^d$ (orthonormal):

$$\text{Var} \left[ \langle X, a^1 \rangle \right], \ldots, \text{Var} \left[ \langle X, a^d \rangle \right].$$
Algorithm

- For \( j = 0 \), let \( R^0 = X \).
- For \( j = 1, \ldots, d \):
  
  Determine \( a^j = \arg \max_{x \in \mathbb{R}^p} E \left[ \langle x, R^{j-1} \rangle^2 \right] \) such that \( \|a^j\| = 1 \) and \( \langle a^j, a^k \rangle = 0 \), \( 1 \leq k < j \).

  [P] Projection.
  Compute the principal variable \( Y^j = \langle a^j, R^{j-1} \rangle \).

  [R] Linear regression.
  Determine \( b^j = \arg \min_{x \in \mathbb{R}^p} E \left[ \|R^{j-1} - Y^j x \|^2 \right] \) such that \( \langle b^j, a^j \rangle = 1 \) and \( \langle b^j, a^k \rangle = 0 \), \( 1 \leq k < j \). The solution is \( b^j = a^j \), and let the regression function be \( s^j(t) = ta^j \).

  [U] Residual update.
  Compute \( R^j = R^{j-1} - s^j(Y^j) \).
Algorithm output. After $d$ iterations, we have the following expansion:

$$X = \sum_{k=1}^{d} s^k(Y^k) + R^d,$$

(1)

with $s^k(t) = ta^k$ and $Y^k = \langle a^k, X \rangle$, or equivalently

$$X = \sum_{k=1}^{d} \langle a^k, X \rangle a^k + R^d.$$

This equation can be rewritten as

$$F(X) = R^d$$

(2)

where we have defined

$$F(x) = x - \sum_{k=1}^{d} \langle a^k, x \rangle a^k.$$

The equation $F(x) = 0$ 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 an expansion similar to (2):
   \[ F(X) = R^d, \]
   but with a non necessarily linear function \( F \), such that the equation \( F(x) = 0 \) could model more general subspaces.

2. To keep an expansion “principal variables + residual” similar to (1):
   \[ X = \sum_{k=1}^{d} s^k(Y^k) + R^d, \]
   but with non necessarily linear functions \( s^k \).

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:

   Introduction of an index $I$ which measures the quality of the projection axis. For instance:
   - Dispersion,
   - Deviation from normality,
   - Clusters detection,
   - Outliers detection,

[R] Regression.
   Estimation of the regression function from $\mathbb{R}$ to $\mathbb{R}^p$ in a given set:
   - Linear functions,
   - Splines, kernels,
New algorithm.

- For \( j = 0 \), let \( R^0 = X \).
- For \( j = 1, \ldots, d \):
  
  Determine \( a^j = \arg \max_{x \in \mathbb{R}^p} I(\langle x, R^{j-1} \rangle) \) such that \( \| a^j \| = 1 \) and \( \langle a^j, a^k \rangle = 0, 1 \leq k < j \).
  
  [P] Projection.
  Compute the principal variable \( Y^j = \langle a^j, R^{j-1} \rangle \).
  
  [R] Regression.
  Determine \( s^j = \arg \min_{s \in S(\mathbb{R}, \mathbb{R}^p)} \mathbb{E} \left[ \| R^{j-1} - s(Y^j) \|^2 \right] \) such that \( P_{a^j} \circ s^j = \text{Id}_\mathbb{R} \) and \( P_{a^k} \circ s^j = 0, 1 \leq k < j \).
  
  [U] Residual update
  Compute \( R^j = R^{j-1} - s^j(Y^j) \).
Remark: At the end of iteration $j$, the residual is given by

\[
R^j = R^{j-1} - s^j (Y^j) \\
= R^{j-1} - s^j (\langle a^j, R^{j-1} \rangle) \\
= R^{j-1} - s^j \circ P_{a^j} (R^{j-1}) \\
= (\text{Id}_{\mathbb{R}^p} - s^j \circ P_{a^j}) (R^{j-1}) \\
= (\text{Id}_{\mathbb{R}^p} - s^j \circ P_{a^j}) \circ (\text{Id}_{\mathbb{R}^p} - s^{j-1} \circ P_{a^{j-1}}) (R^{j-2}) \\
= \ldots \\
= (\text{Id}_{\mathbb{R}^p} - s^j \circ P_{a^j}) \circ \ldots \circ (\text{Id}_{\mathbb{R}^p} - s^1 \circ P_{a^1}) (R^0) \\
= (\text{Id}_{\mathbb{R}^p} - s^j \circ P_{a^j}) \circ \ldots \circ (\text{Id}_{\mathbb{R}^p} - s^1 \circ P_{a^1}) (X).
\]

Auto-associative composite model.
**Remark:** The constraint $P_{aj} \circ s^j = \text{Id}_R$.

- Natural constraint.


\begin{align*}
\text{Important consequence: At the end of iteration } j, \text{ the residual is given by } \\
R^j &= (\text{Id}_R - s^j \circ P_{aj}) (R^{j-1}), \text{ and thus is projection on } a^j \text{ is} \\
P_{aj} R^j &= (P_{aj} - P_{aj} \circ s^j \circ P_{aj}) (R^{j-1}) \\
&= (P_{aj} - P_{aj}) (R^{j-1}) \\
&= 0.
\end{align*}

Thus, iteration $(j + 1)$ can be performed on the linear subspace orthogonal to $(a^1, \ldots, a^j)$, which is of dimension $(p - j)$. 
Goal 1. After $d$ iterations:

- One always has an auto-associative model
  \[ F(X) = R^d, \]
  with
  \[ F = (\text{Id}_{\mathbb{R}^p} - s^d \circ P_{a^d}) \circ \ldots \circ (\text{Id}_{\mathbb{R}^p} - s^1 \circ P_{a^1}) = \prod_{k=d}^{1} (\text{Id}_{\mathbb{R}^p} - s^k \circ P_{a^k}), \]
  and $P_{a^j}(x) = \langle a^j, x \rangle$.
- The equation $F(x) = 0$ defines a $d-$ dimensional manifold.
Goal 2. After \( d \) iterations:

- One always as the expansion “principal variables + residual” similar to (1):

\[
X = \sum_{k=1}^{d} s^k(Y^k) + R^d,
\]

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: \( \mathbb{E} [R^k] = 0, \quad k = 0, \ldots, d. \)
Goal 3. After $d$ iterations, we have:

- Some orthogonality properties
  \[
  \langle a_k, a^j \rangle = 0, \ 1 \leq k < j \leq d, \\
  \langle a_k, R^j \rangle = 0, \ 1 \leq k \leq j \leq d, \\
  \langle a_k, s^j(Y^j) \rangle = 0, \ 1 \leq k < j \leq d.
  \]

- Since the norm of the residuals is decreasing, we can define, similarly to the PCA case, the information ratio represented by the $d$– dimensional model as
  \[
  Q_d = 1 - \mathbb{E} \left[ \| R^d \|^2 \right] / \text{Var} \left[ \| X \|^2 \right].
  \]
  One can show that $Q_0 = 0$, $Q_p = 1$ and $(Q_d)$ is increasing.

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

\[
\mathbb{E} \left[ Y^k Y^j \right] \neq 0, \ 1 \leq k < j \leq d.
\]
Goal 4.

- We still have an iterative algorithm. It converges at most in $p$ steps.
- Its complexity depends on the two steps [A] et [R].

Determine $a^j = \arg \max_{x \in \mathbb{R}^p} I(\langle x, R^{j-1} \rangle)$ such that $\|a^j\| = 1$ and $\langle a^j, a^k \rangle = 0$, $1 \leq k < j$.

[R] Regression.
Determine $s^j = \arg \min_{s \in S(\mathbb{R}, \mathbb{R}^p)} \mathbb{E} \left[ \| R^{j-1} - s(Y^j) \|^2 \right]$ such that $P_{a^j} \circ s^j = \text{Id}_\mathbb{R}$ and $P_{a^k} \circ s^j = 0$, $1 \leq k < j$.

- Note that the above theoretical properties do not depend on these steps.
3. Implementation aspects, step [A]

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

\[
I(\langle x, R_j^{j-1} \rangle) = \frac{\sum_{i=1}^{n} \langle x, R_i^{j-1} \rangle^2}{\sum_{k=1}^{n} \sum_{\ell=1}^{n} m_{k\ell} \langle x, R_k^{j-1} - R_{\ell}^{j-1} \rangle^2},
\]

where \( M = (m_{k\ell}) \) is the contiguity matrix defined by \( m_{k\ell} = 1 \) if \( R_{\ell}^{j-1} \) is the closest neighbor of \( R_k^{j-1} \), \( m_{k\ell} = 0 \) otherwise.

• **Optimization.** Explicit solution.

[A] \( a^j \) is the eigenvector associated to the largest eigenvalue of \( V_j^* V_j^{-1} \), where

\[
V_j = \sum_{k=1}^{n} t^{t} R_k^{j-1} R_k^{j-1}, \quad V_j^* = \sum_{k=1}^{n} \sum_{\ell=1}^{n} m_{k\ell} (R_k^{j-1} - R_{\ell}^{j-1}) (R_k^{j-1} - R_{\ell}^{j-1})
\]

are proportional to the covariance and local covariance matrices of \( R^{j-1} \).
Implementation aspects, step [R]

- **Set of $L^2$ functions.** The regression step reduces to estimating the conditional expectation:

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

- **Estimation of the conditional expectation.**
  
  - Classical problem since the constraints $P_{a_j} \circ s^j = \text{Id}$ and $P_{a_k} \circ s^j = \text{Id}$, $1 \leq k < j$ are easily taken into account in the $a^k$’s basis. Step [R] reduces to $(p - j)$ independent regressions from $\mathbb{R}$ to $\mathbb{R}$.
  
  - Numerous estimates are available: splines, local polynomials, kernel estimates, ...
  
  - For instance, for the coordinate $k \in \{j + 1, \ldots, p\}$, the kernel estimate of $s^j(u)$ can be written as

    \[
    \hat{s}^j_k(u) = \sum_{i=1}^{n} \hat{R}^{j-1}_{i,k} K_h(u - Y^j_i) \bigg/ \sum_{i=1}^{n} K_h(u - Y^j_i),
    \]

    where $h$ is a smoothing parameter (the bandwidth).
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 $h = 0.3 \rightarrow Q_1 = 99.97\%$.

Theoretical curve

Estimated 1–dimensional manifold
Second illustration on a simulated dataset

- $n = 1000$ points in $\mathbb{R}^3$ randomly chosen on the surface
  $(x, y) \rightarrow (x, y, \cos(\pi \sqrt{x^2 + y^2})(1 - \exp\{-64(x^2 + y^2)\}))$.
- Two iterations: $Q_1 = 84.1\%$ et $Q_2 = 97.6\%$.

Theoretical surface  Simulated points  Estimated 2– dimensional manifold
Auto-Associative models and generalized Principal Component Analysis

$s^1$ (blue) and $s^2$ (red)  
Residuals $R_i^1$  
Residuals $R_i^2$
5. First illustration on a real dataset

- Set of $n = 45$ images of size $256 \times 256$.

- Interpretation: $n = 45$ points in dimension $p = 256^2$.

- Rotation: $n = 45$ points in dimension $p = 44$. 
\begin{itemize}
  \item Information ratio $Q_d$ as a function of $d$ (blue: classical PCA, green: generalized PCA).
\end{itemize}
- Projection on the 3 first PCA axes of the estimated manifolds (dimension 1 & dimension 2).
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 $Q_d$ as a function of $d$ (blue: classical PCA, green: generalized PCA).
Estimated 1–dimensional manifold projected on the principal plane.
Estimated 1-dimensional manifolds projected on the principal plane, for each type of cancer.