Analyse en composantes principales

Gilles Gasso, Stéphane Canu

INSA Rouen - Département ASI
Laboratoire LITIS [1]

17 septembre 2014

1. **Ce cours est librement inspiré du cours DM de Alain Rakotomamonjy**

Plan

Introduction

Objectifs

- $\{x_i \in \mathbb{R}^D\}_{i=1,\cdots,N}$: ensemble de N points décrits par D attributs.
- Objectifs de l'analyse en composantes principales
  1. représentation (graphique) des points dans un sous-espace de dimension d ($d << D$) telle que la déformation du nuage de points soit minimale
  2. réduction de la dimension, ou approximation des points à partir de d variables ($d \leq D$).

Notations

- Observation : $x_i \in \mathbb{R}^D$ avec $x_i = \begin{pmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,D} \end{pmatrix}^\top$
- Variable (attribut) : x^j

Les données : description

Données

Soit X la matrice des données ($x_i \in \mathbb{R}^D$) :

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,D} \\ \vdots & & & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,D} \end{pmatrix} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_N \end{pmatrix}^\top$$

Statistiques sommaires : moyenne et variance

- Moyenne $\bar{x} = \begin{pmatrix} \bar{x}^1 & \bar{x}^2 & \dots & \bar{x}^D \end{pmatrix}^\top$ avec $\bar{x}^j = \frac{1}{N} \sum_{i=1}^N x_{i,j}$,
- Variance des variables $\text{var}(x^j) = \frac{1}{N} \sum_{i=1}^N (x_{i,j} - \bar{x}^j)^2$

Covariance et Matrice de covariance

Covariance entre variables j et k

$$\text{cov}(x^j, x^k) = \frac{1}{N}\sum_{i=1}^{N}(x_{i,j} - \bar{x}^j)(x_{i,k} - \bar{x}^k)$$

- Si covariance grande (en valeur absolue) $\Longrightarrow$ variables j et k dépendantes. Covariance nulle $\Longrightarrow$ variables indépendantes

Matrice de covariance $\Sigma \in \mathbb{R}^{D\times D}$

- Σ est une matrice symétrique de terme général $\Sigma_{j,k} = \text{cov}(x^j, x^k)$:

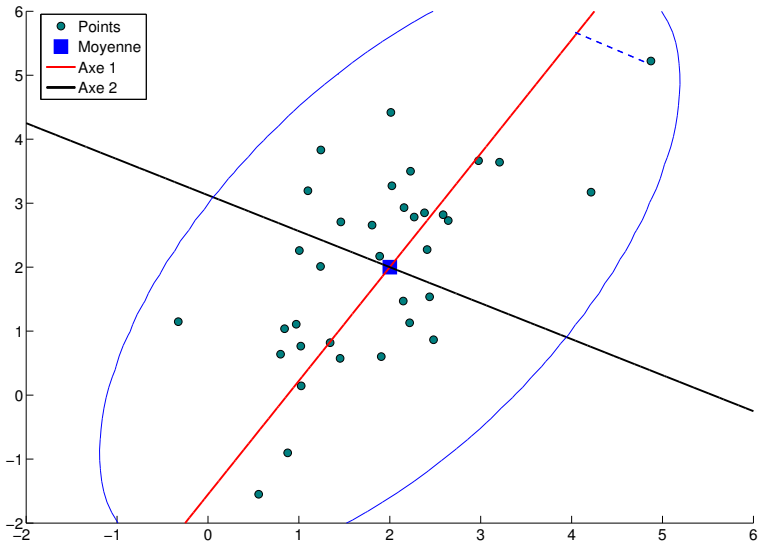
$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\top$$

- données centrées : $\Sigma = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top$, ou encore $\Sigma = \frac{1}{N}X^\top X$

Analyse en Composantes Principales

Principe

- Soit $x_i \in \mathbb{R}^D, i = 1, \cdots, N$ des données centrées.
- Objectif : trouver un sous-espace de dimension $d \leq D$ où projeter les x_i de façon à perdre le moins d'informations possibles



- Trouver une "meilleure base orthonormale" de représentation des données par combinaison linéaire de la base originale.
- p_1, p_2 : vecteurs orthonormés (axes 1 et 2). Projeter les données sur l'espace engendré par p_1 et $p_2 \Longrightarrow$ changement de base
- Quel est le meilleur sous-espace de dimension 1 ?

Analyse en Composantes Principales

Objectifs et hypothèses

- $X \in \mathbb{R}^{N \times D}$: matrice de données centrées.
- Objectif ACP : trouver un sous-espace de dimension $d \leq D$ qui permet d'avoir une représentation réduite de X.

Comment ?

- Projection linéaire de $x_i \in \mathbb{R}^D$ sur $t_i \in \mathbb{R}^d$

$$t_i = P^\top x_i \qquad \text{avec} \qquad P = \begin{pmatrix} p_1 & \cdots & p_d \end{pmatrix}, \quad p_i \in \mathbb{R}^D$$

$P \in \mathbb{R}^{D \times d}$: matrice de transformation linéaire

- Contrainte : $P^\top P = I$
  Les vecteurs de la nouvelle base sont orthogonaux 2 à 2 c'est-à-dire $p_j^\top p_i = 0 \quad \forall\, i \neq j$

Analyse en Composantes Principales

- Reconstruction de x_i à partir de t_i
  - Si $d = D$, la matrice P est orthogonale

$$t_i = P^\top x_i \quad \Longrightarrow \quad Pt_i = PP^\top x_i \quad \Longrightarrow \quad x_i = Pt_i$$

Dans ce cas, pas de réduction de dimension, juste un changement de base et donc pas de perte d'information

  - $d < D$ (réduction de dimension)
Reconstruction de x_i par l'approximation

$$\hat{x}_i = Pt_i \qquad ou \qquad \hat{x}_i = PP^\top x_i$$

Problématique

Construire P de sorte que l'erreur $\|x_i - \hat{x}_i\|^2$ entre le vrai x_i et sa reconstruction $\hat{x}_i$ soit minimale et ceci pour tous les points x_i, $i = 1, \cdots, N$

Minimisation d'erreur/maximisation variance

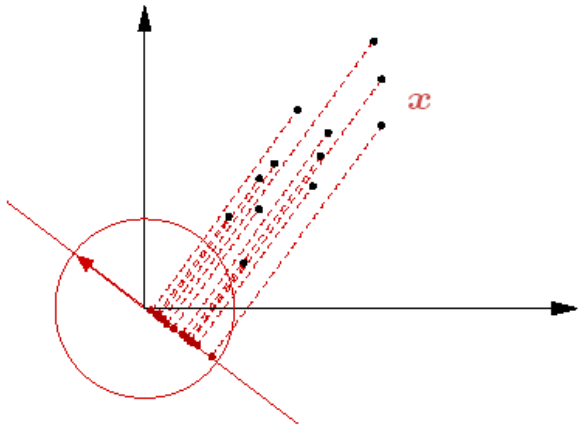
- Soit $J_e(P)$ l'erreur quadratique d'estimation. On a :

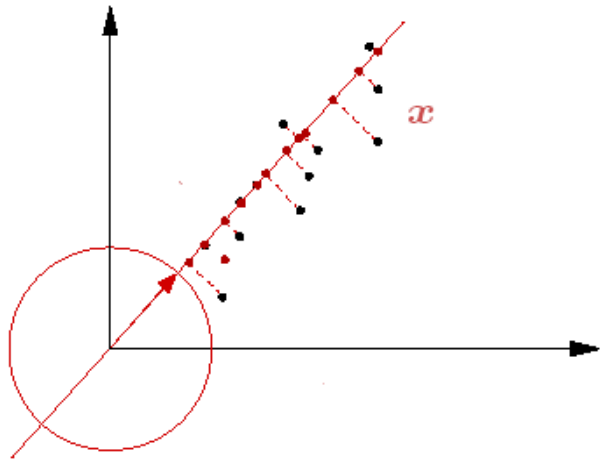
$$
\begin{aligned}
J_e(P) &= \frac{1}{N}\sum_{i=1}^{N}\|x_i-\hat{x}_i\|^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-PP^\top x_i)^\top(x_i-PP^\top x_i)\\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i^\top x_i-2x_i^\top PP^\top x_i+x_i^\top PP^\top PP^\top x_i)\\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^\top x_i-\frac{1}{N}\sum_{i=1}^{N}x_i^\top PP^\top x_i=\frac{1}{N}\sum_{i=1}^{N}x_i^\top x_i-\frac{1}{N}\sum_{i=1}^{N}t_i^\top t_i\\
&= trace\left(\frac{1}{N}\sum_{i=1}^{N}x_i^\top x_i-\frac{1}{N}\sum_{i=1}^{N}t_i t_i^\top\right)=trace\left(\frac{1}{N}\sum_{i=1}^{N}x_i x_i^\top-\frac{1}{N}\sum_{i=1}^{N}P^\top x_i x_i^\top P\right)\\
J_e(P) &= trace\left(\Sigma\right)-trace\left(P^\top\Sigma P\right) \quad \text{pour des données } x_i \text{ centrées}
\end{aligned}
$$

- $min J_e(P)$ revient à maximiser par rapport à P la variance $P^\top\Sigma P$ des points projetés.

Axes factoriels et composantes principales

- Soit X la matrice des données et $p \in \mathbb{R}^D$ tq $\|p\| = 1$. Soit le vecteur de $\mathbb{R}^N$, $c_1 = Xp_1 = \left(x_1^\top p_1 \ \dots \ x_N^\top p_1\right)^\top$.
- On appelle premier axe factoriel de X le vecteur p_1 tel que la variance de Xp_1 soit maximale. Le vecteur c_1 est appelé première composante principale.





Le $k$ième axe factoriel est le vecteur p_k unitaire ($\|p_k\| = 1$) tel que la variance de $c_k = Xp_k$ soit maximale et que p_k soit orthogonal aux $k-1$ premiers axes factoriels.

Minimisation de l'erreur quadratique d'estimation

Premier axe factoriel

On cherche le sous espace engendré par p_1 tq $p_1^\top p_1 = 1$.

- Problème d'optimisation sous contrainte égalité :

$$\min_{p_1} J_e(p_1) = \frac{1}{N}\sum_{i=1}^{N} x_i^\top x_i - \frac{1}{N}\sum_{i=1}^{N} x_i^\top p_1 p_1^\top x_i \quad \text{avec} \quad p_1^\top p_1 = 1$$

  - Simplification de $J_e(p_1)$

$$J_e(p_1) = -p_1^\top \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top\right) p_1 = -p_1^\top \Sigma p_1$$

- Le lagrangien s'écrit

$$\mathcal{L}(p_1, \lambda_1) = -p_1^\top \Sigma p_1 + \lambda_1(p_1^\top p_1 - 1)$$

Minimisation de l'EQE

Optimisation

- Conditions d'optimalité

$$\nabla_{p_1}\mathcal{L} = 0 = -2\Sigma p_1 + 2\lambda_1 p_1 \quad \text{et} \quad \nabla_{\lambda_1}\mathcal{L} = 0 = p_1^\top p_1 - 1$$

$$\Longrightarrow \Sigma p_1 = \lambda_1 p_1 \quad \text{et} \quad p_1^\top \Sigma p_1 = \lambda_1$$

- Interprétation
  1. $(\lambda_1,\ p_1)$ représente la paire (valeur propre, vecteur propre) de la matrice de covariance Σ
  2. $J_e(p_1) = -\lambda_1$ est la fonctionnelle que l'on cherche à minimiser

Solution

Le premier axe factoriel p_1 est le vecteur propre associé à la plus grande valeur propre de Σ.

k-ième axe factoriel

Lemme

Le sous-espace de dimension k minimisant l'erreur quadratique d'estimation des données contient nécessairement le sous-espace de dimension $k-1$.

Calcul du 2e axe factoriel p_2 sachant que p_1 est connu

$$\min_{p_2} \quad J_e(p_2) = -p_2^\top \Sigma p_2$$

$$\text{tel que} \quad p_2^\top p_2 = 1, \quad p_2^\top p_1 = 0$$

- Interprétation : on cherche un vecteur unitaire p_2 qui maximise la variance $p_2^\top \Sigma p_2$ et qui soit orthogonal au vecteur p_1

Exercice

Montrer que p_2 est le vecteur propre associé à λ_2, la seconde plus grande valeur propre de Σ

Algorithme

1. Centrer les données : $\{x_i \in \mathbb{R}^D\}_{i=1}^N \longrightarrow \{x_i = x_i - \bar{x} \in \mathbb{R}^D\}_{i=1}^N$
2. Calculer la matrice de covariance $\Sigma = \frac{1}{N} X^\top X$ avec $X^\top = \begin{pmatrix} x_1 & \cdots & x_N \end{pmatrix}$
3. Calculer la décomposition en valeurs propres $\{p_j \in \mathbb{R}^D, \lambda_j \in \mathbb{R}\}_{j=1}^D$ de Σ
4. Ordonner les valeurs propres λ_j par ordre décroissant
5. Nouvelle base de représentation des données :

$$P = (p_1, \cdots, p_d) \in \mathbb{R}^{D \times d}$$

   $\{p_1, \cdots, p_d\}$ sont les d vecteurs propres associés aux d plus grandes λ_j.

6. Projection de tous les points via P s'obtient matriciellement :

$$C = XP = \begin{pmatrix} c_1 & \cdots & c_d \end{pmatrix}$$

Note : la projection d'un point quelconque x est donnée par $t = P^\top (x - \bar{x})$

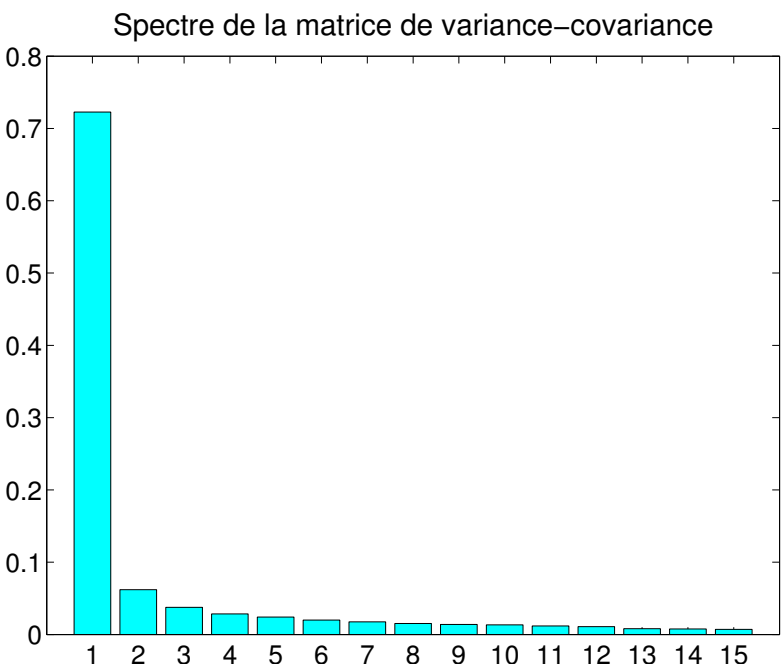
Propriétés des axes factoriels

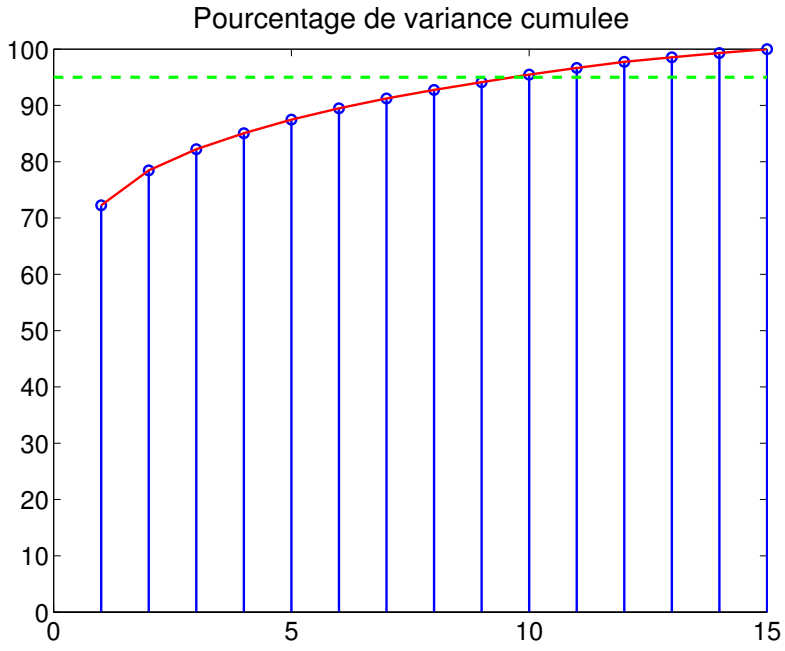
- Les valeurs propres de Σ sont positives car Σ est une matrice semi-definie positive
- Le nombre d'axes factoriels est égal au nombre de valeurs propres non-nulles de Σ.
- La variance expliquée par l'axe factoriel p_k (homogène à une inertie) s'écrit $I_k = p_k^\top \Sigma p_k = p_k^\top \lambda_k p_k = \lambda_k$.
- La variance totale des axes factoriels est $I = \sum_{k=1}^{d} \lambda_k$
- Pourcentage de variance expliquée par les d premiers axes

$$\frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{D} \lambda_k} \cdot 100$$

Propriétés des axes factoriels

- Choix de la dimension d du sous-espace
  - Validation croisée
  - Détection "d'un coude" sur le graphique des valeurs propres
  - On choisit d de sorte qu'un pourcentage fixé (par exemple 95%) de la variance soit expliqué





Propriétés de l'ACP

- Les composantes principales $\{c_i\}_{i=1,\cdots,D}$ sont centrées et non-corrélées ie

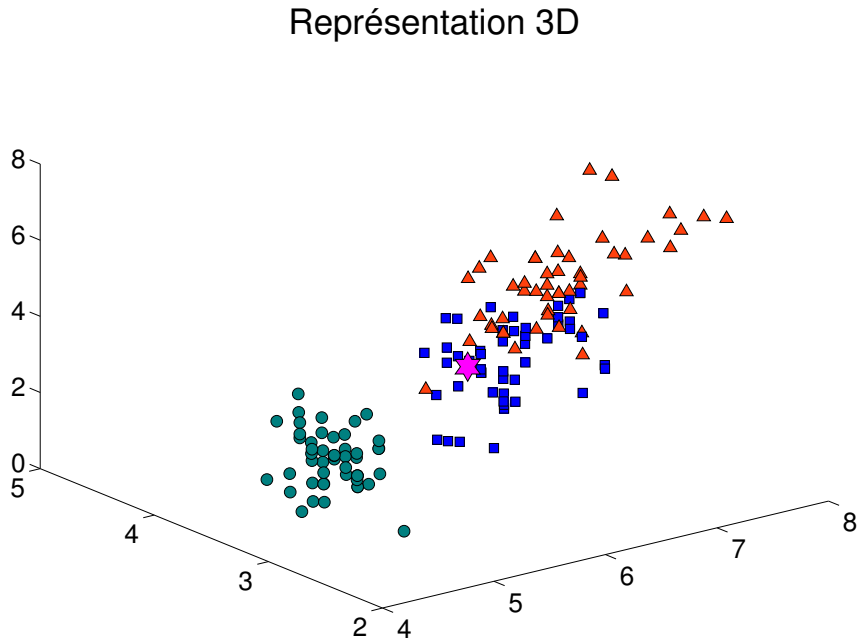
$$\text{cov}(c_i, c_k) = 0 \quad \text{si} \quad i \neq k$$

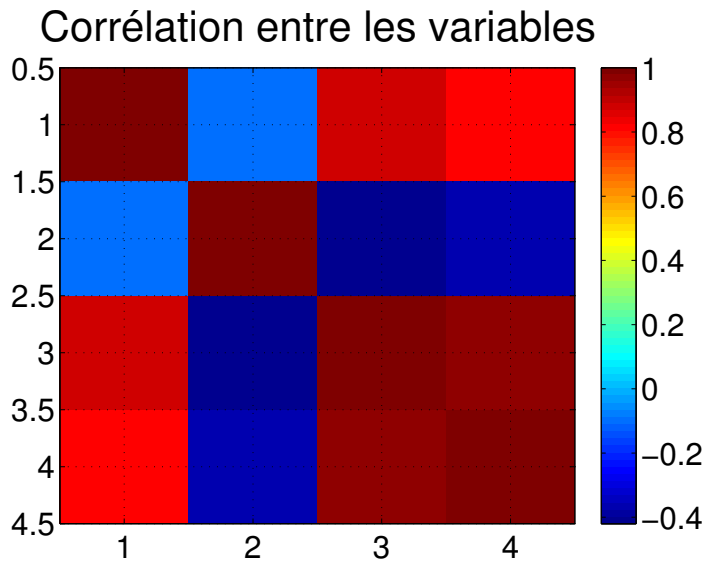
$$\text{cov}(c_i, c_k) = \frac{1}{N} c_i^\top c_k = \frac{1}{N} p_i^\top X^\top X p_k = p_i^\top \Sigma p_k = p_i^\top (p_k \lambda_k) = 0$$

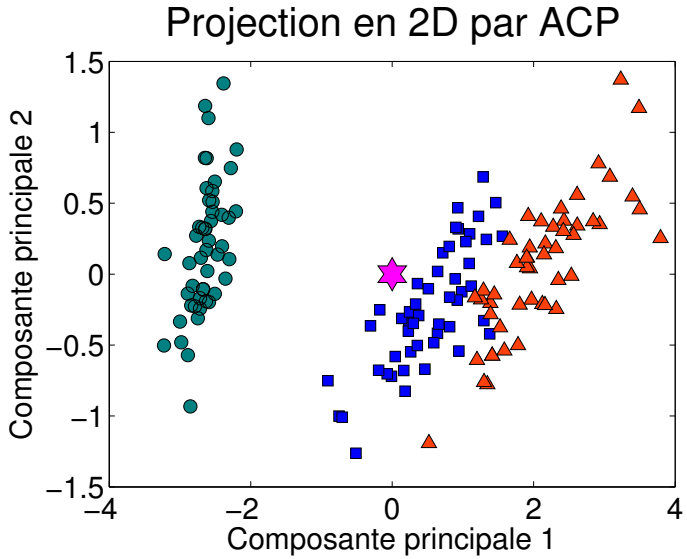
- Soit $c_k = X p_k$, le vecteur représentant la projection de X sur le k-ième axe p_k. La variance de la composante principale c_k est

$$\frac{1}{N} c_k^\top c_k = \frac{1}{N} p_k^\top X^\top X p_k = p_k^\top \Sigma p_k = p_k^\top \lambda_k p_k = \lambda_k$$

Exemple des données iris : $x_i \in \mathbb{R}^4$







Réduction de la dimensionalité

- ACP $\equiv$ représenter les données dans un espace de dim. réduite.
- La nouvelle base de représentation est donnée par la matrice P.
  Chaque vecteur de cette base est combinaison linéaire des vecteurs de la base originale. P vérifie $P^\top P = I$.
- $C = XP$: matrice des composantes principales qui est en fait la matrice de projections de tous les x_i sur les axes factoriels.
- Reconstruction des x_i à partir des composantes principales
  x_i est reconstruit par $\hat{x}_i = Pt_i + \bar{x}$ avec $t_i = P^\top x_i$.
  On déduit que la matrice des données reconstruites est

$$\hat{X} = CP^\top + \mathbf{1}_N \otimes \bar{x}^\top \quad \text{ou} \quad \hat{x}_i = \sum_{k=1}^{d} C_{i,k} p_k + \bar{x}$$

Note : un point quelconque projeté t est reconstruit par $\hat{x} = Pt + \bar{x}$

Réduction de la dimensionalité

- Si $q = d$, c'est à dire que le nouveau sous-espace de représentation est égale à l'espace original alors

$$\hat{X} = X$$

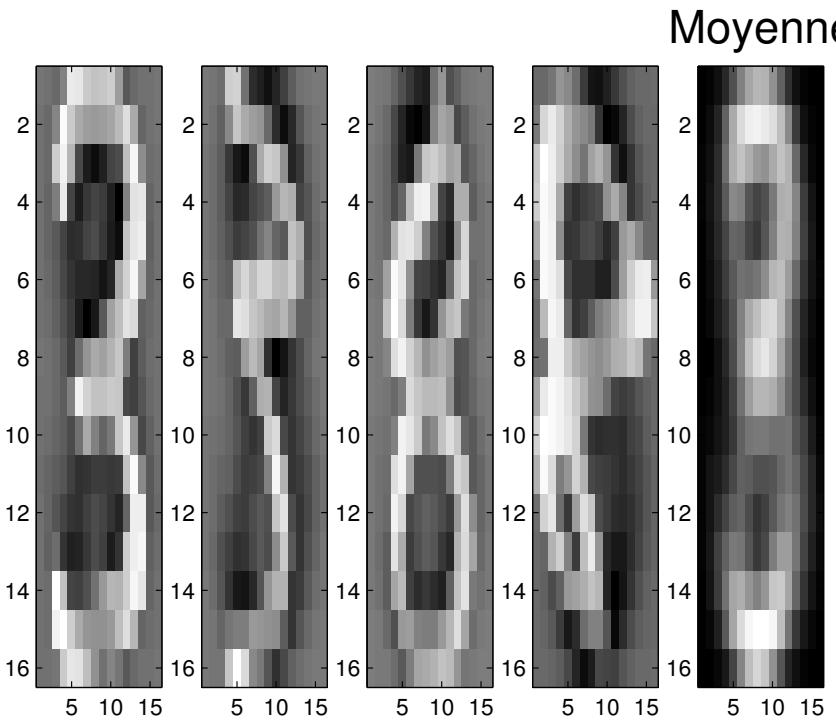
- Erreur d'approximation sur un sous-espace vectoriel de dimension d

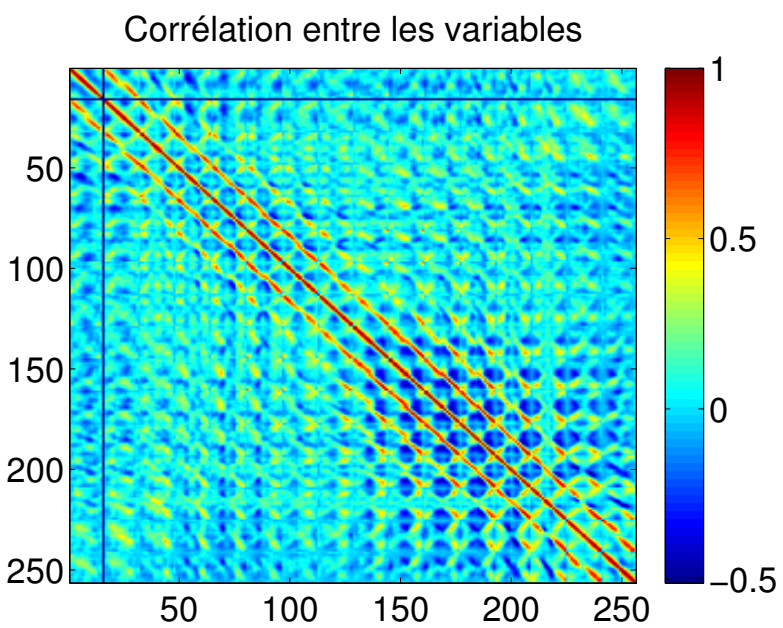
$$E_q = \frac{1}{N} \sum_{i=1}^{N} ||x_i - \hat{x}_i^{(d)}||^2 = \sum_{i=d+1}^{D} \lambda_i$$

- L'analyse en composantes principale est un outil de visualisation des données ...
- ... et permet de faire de la reduction de la dimensionalité.

Exemple : données USPS

- Caractères manuscripts sous forme d'images 16×16
- Chaque image est transformée en un vecteur de dimension 256
- On a pris ici des "3" et des "8" (quelques exemples ci-dessous)





Exemple : données USPS

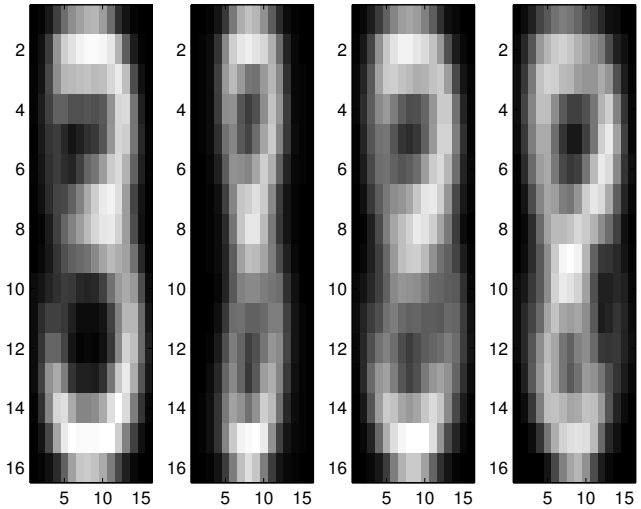


Figure: Reconstruction avec $d = 2$ composantes

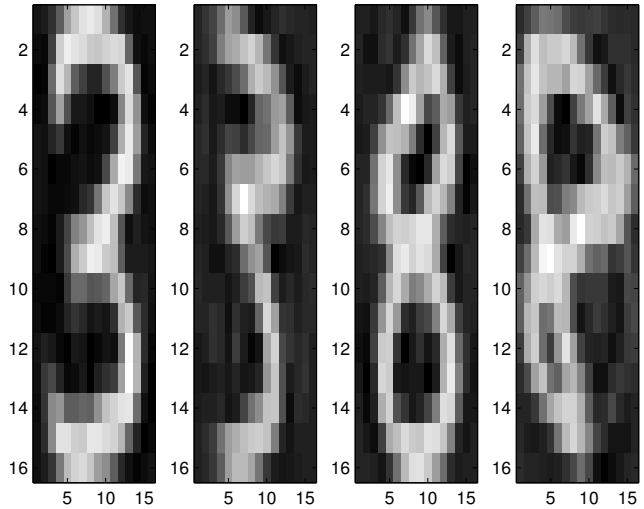


Figure: Reconstruction avec $d = 50$ composantes