Master IA2S (Intelligence Artificielle, Science des données et Systèmes Cyber-Physiques)

Apprentissage Automatique

Ferhat ATTAL
ferhat.attal@u-pec.fr
Université Paris Est Créteil (UPEC)

Novembre 2021

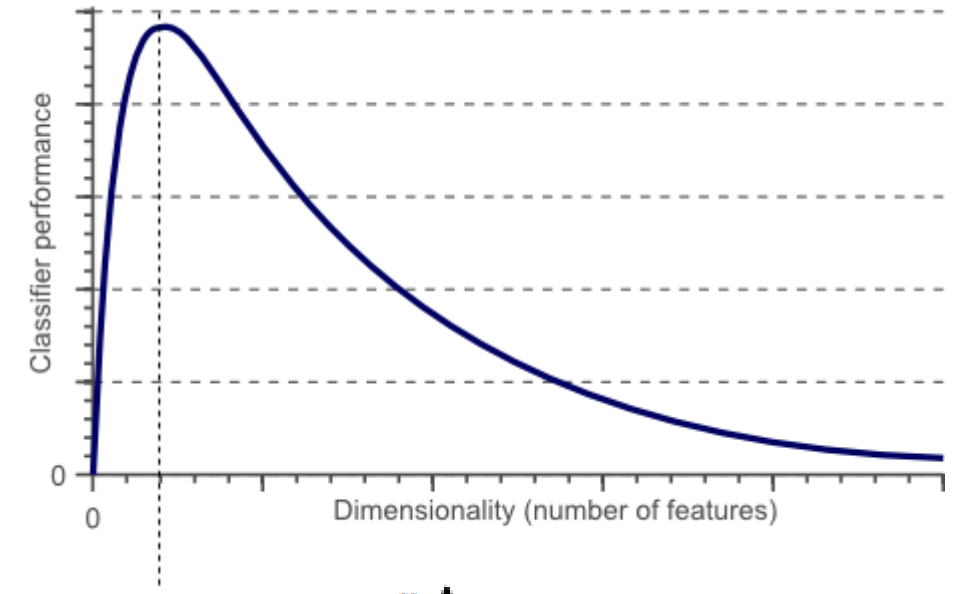
Cours 5 : Réduction de la dimension

Réduction de la dimension

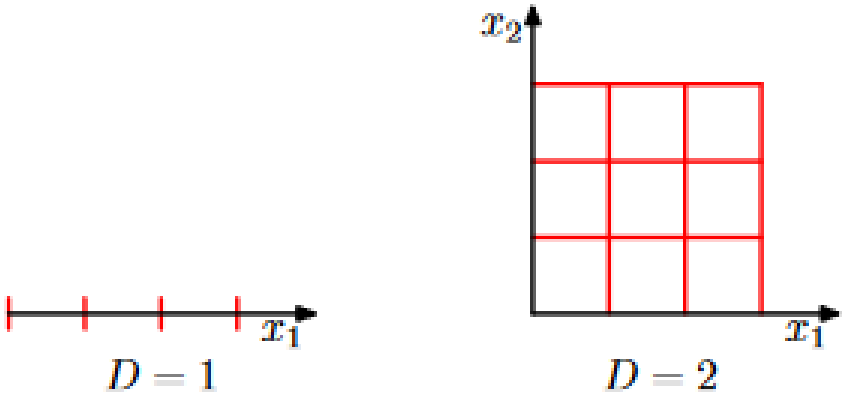
- Réduction de la dimension
- Analyse en composantes principales
- Analyse factorielle

Malédiction de la dimensionnalité

- L'augmentation du nombre de variables (caractéristiques) n'améliorera pas toujours les performances de la classification.

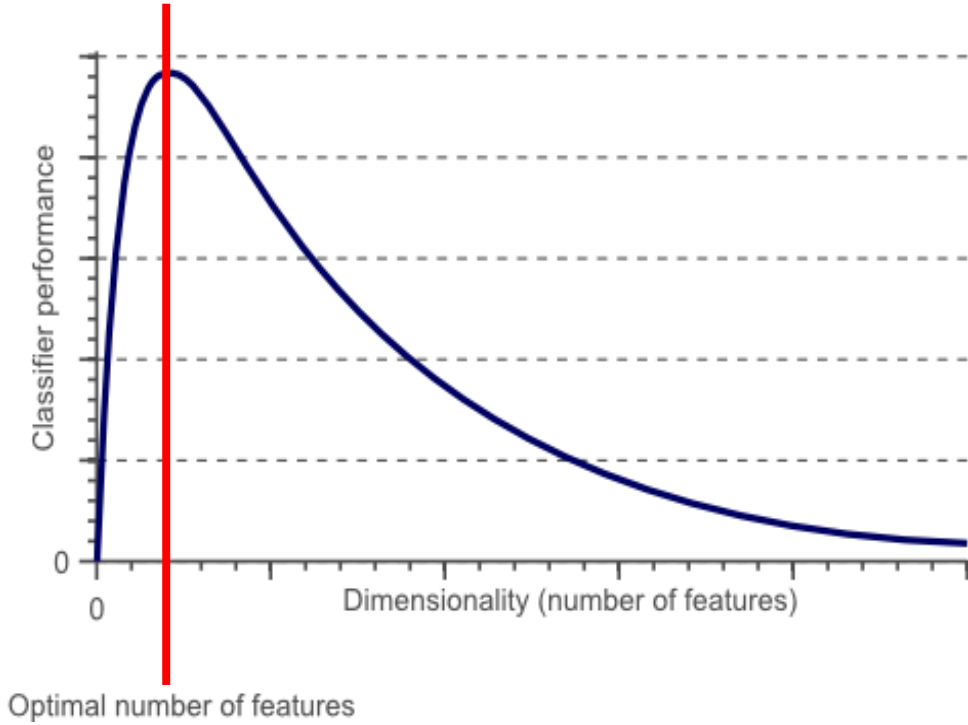


- Le nombre d'exemples (individus) nécessaires pour l'apprentissage augmente de manière exponentielle avec la dimension d.



Réduction de la dimension

- Choisir un ensemble optimal de variables de dimension inférieure pour améliorer la capacité discriminative des algorithmes
- Réduire le temps de calcul des algorithmes
- Réduit la complexité de l'espace des variables
- faciliter l’interprétation des données
- Augmenter la capacité de généralisation des algorithmes
- Réduire la redondance



Réduction de la dimension

Extraction de caractéristiques (**Feature extraction)**: permet de trouver un ensemble de **nouvelles** caractéristiques (c'est-à-dire par le biais d'une fonction de projection **f()**) à partir des caractéristiques existantes.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_j \\ \vdots \\ x_d \end{bmatrix} \xrightarrow{\mathbf{f()}} z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ \vdots \\ z_K \end{bmatrix}$$

K<<N

Sélection de caractéristiques (Feature selection) : choisir un sous-ensemble de variables à partir de l'ensemble de variables d'originales.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_j \\ \vdots \\ x_d \end{bmatrix} \rightarrow z = \begin{bmatrix} x_1 \\ x_5 \\ \vdots \\ \vdots \\ x_K \end{bmatrix}$$

K<<N

Réduction de la dimension

D'un point de vue mathématique, trouver une projection optimale $z = f(x)$ équivaut à optimiser une fonction objective.

Exemple de fonctions objectives:

- Minimiser la perte d'information : le but est de représenter les données aussi précisément que possible (c'est-à-dire sans perte d'information) dans l'espace de dimension inférieure.

- Augmenter la capacité discriminatoire : l'objectif est de trouver une projection permettant d'améliorer la capacité discriminatoires dans l'espace de dimension inférieure.

Analyse en composantes principales (ACP) transformation de Karhunen–Loève (KLT)

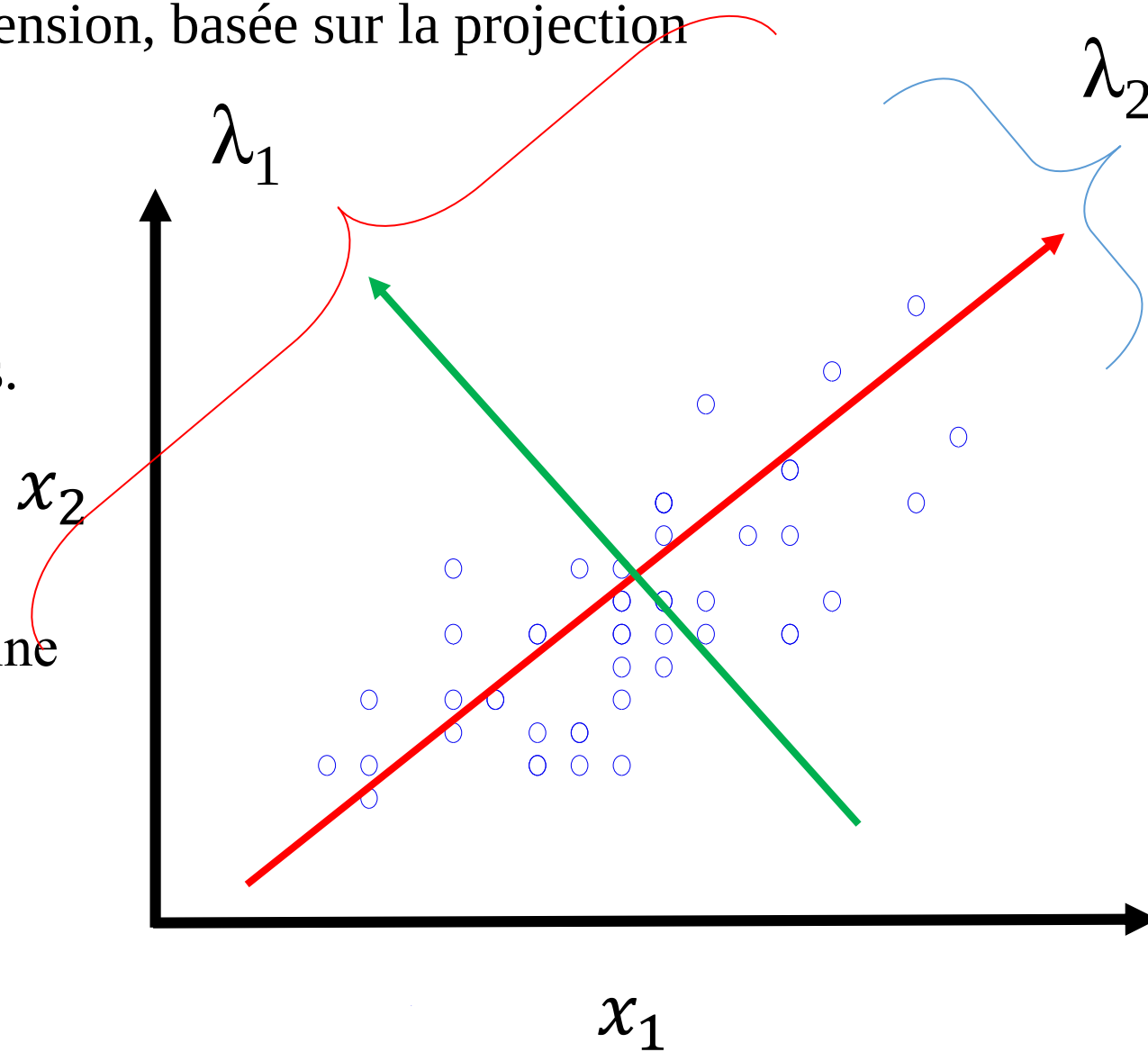
L'ACP est une technique de réduction de la dimension, basée sur la projection linéaire des données.

Soit en

- Maximisant la variance des données projetées.

Ou bien en

- Minimisant l’erreur entres les données d’origine et les données projetées



Analyse en composantes principales

Soit un échantillon $S = \{x_i\}_{i \leq n}$ où x_i appartient à $\mathbb{R}^d$.

	1	***…***	***j***	***…***	***d***
1	x_{11}	…	x_{1j}	…	x_{1d}
⋮	⋮		⋮		⋮
i	x_{i1}	…	x_{ij}	…	x_{id}
⋮	⋮		⋮		⋮
n	x_{n1}	…	x_{nj}	…	x_{nd}

L’objectif est de projeter les données sur un espace ayant une dimension K < d tout en maximisant la variance des données projetées

Analyse en composantes principales

Soit le vecteur moyen $\bar{\boldsymbol{x}}$ et la matrice de covariance Σ

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

Avec

$$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{id})^T$$

$$\bar{\boldsymbol{x}} = (\overline{\boldsymbol{x}_1}, \overline{\boldsymbol{x}_2}, \dots, \overline{\boldsymbol{x}_j}, \dots, \overline{\boldsymbol{x}_d})^T$$

Analyse en composantes principales

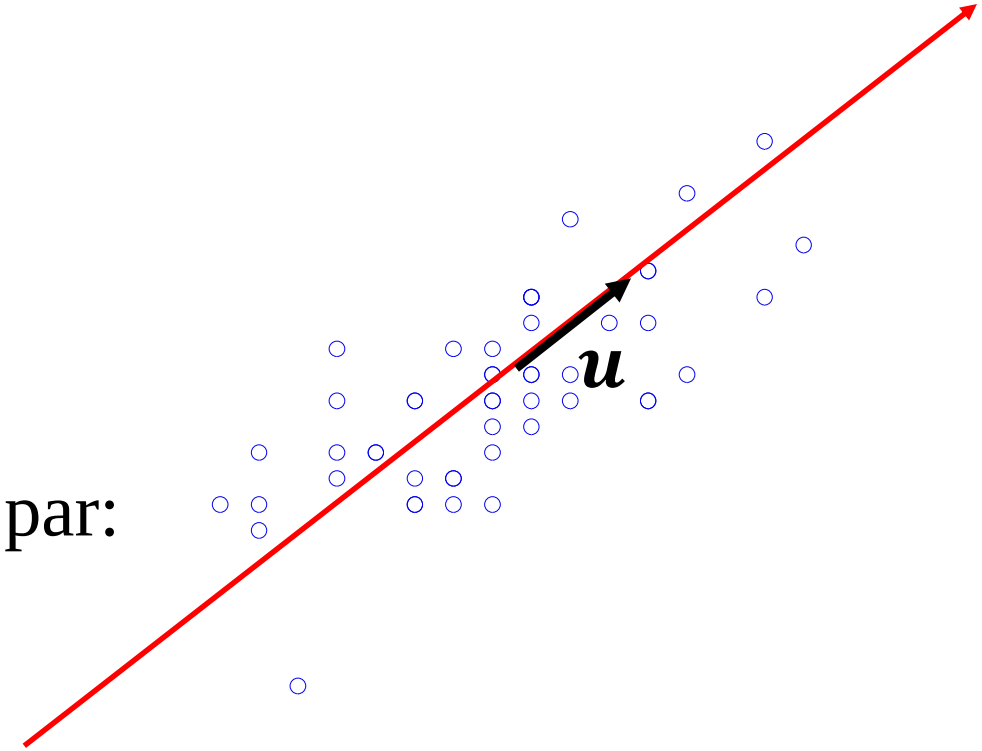
Considérons $\boldsymbol{z_i}$ la projection de $\boldsymbol{x_i}$ sur un espace unidimensionnel (K=1) muni d'un vecteur unitaire $\boldsymbol{u} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d)^{\mathrm{T}}$

$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=\mathbf{1}$

$\boldsymbol{z_i} = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{x_i}$

Ainsi

La variance des points projetés est données par:



$\mathrm{Var}(\mathbf{z}) = \mathrm{var}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x})$

$= \boldsymbol{u}^{\mathrm{T}}\Sigma\,\boldsymbol{u}$

Analyse en composantes principales

Maximisation de la variance des données projetées

$$\begin{cases} \text{M}a\text{ximiser } \boldsymbol{u}^{\mathrm{T}} \Sigma\, \boldsymbol{u} \\ \boldsymbol{s}.\, c \\ \boldsymbol{u}^{\mathrm{T}} \boldsymbol{u} = \mathbf{1} \end{cases}$$

Il s'agit d'un problème de maximisation sous contraintes ⇒ Utilisation du multiplicateur de Lagrange

Analyse en composantes principales

Maximisation de la variance des données projetées

$$\mathcal{L}(\boldsymbol{u}, \lambda) = \boldsymbol{u}^{\mathrm{T}} \Sigma \, \boldsymbol{u} - \lambda \, (\boldsymbol{u}^{\mathrm{T}} \boldsymbol{u} - \mathbf{1})$$

$$\frac{\delta \mathcal{L}(\boldsymbol{u}, \lambda)}{\delta \boldsymbol{u}} = 0 \;\Rightarrow 2\Sigma \, \boldsymbol{u} - 2\lambda \, \boldsymbol{u} = \mathbf{0}$$

$$\Rightarrow \Sigma \, \boldsymbol{u} = \lambda \, \boldsymbol{u}$$

$\lambda \geq 0$ multiplicateur de Lagrange

Analyse en composantes principales

Maximisation de la variance des données projetées

$$\Sigma \boldsymbol{u} = \lambda \boldsymbol{u}$$

Avec

λ : valeurs propre de la matrice de covariance Σ

$\boldsymbol{u}$: vecteurs propre de la matrice de covariance Σ

On a

$$\begin{aligned} \text{Var}(\mathbf{z}) &= \boldsymbol{u}^{\mathrm{T}} \Sigma \boldsymbol{u} \\ &= \boldsymbol{u}^{\mathrm{T}} \lambda \boldsymbol{u} \\ &= \lambda \boldsymbol{u}^{\mathrm{T}} \boldsymbol{u} \\ &= \lambda \end{aligned}$$

Analyse en composantes principales

La matrice de covariance admet $\boldsymbol{d}$ valeurs propres $(\lambda_1, \lambda_2, \ldots, \lambda_d)$ et $\boldsymbol{d}$ vecteurs propres $(\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_d})$

Avec
$\lambda_1 > \lambda_2 > \ \ldots > \lambda_d$

- Maximiser la variance des données projetées $\boldsymbol{u}^{\mathrm{T}} \Sigma \boldsymbol{u}$ revient à trouver la valeur propre maximale λ c.à.d λ_1

- Le vecteur propre $\boldsymbol{u_1}$ correspondant à la valeur propre maximale λ_1 represente le vecteur directeur du premier axe principal (première composante principale).

Analyse en composantes principales

Nous pouvons définir des composantes principales supplémentaires de manière progressive en choisissant chaque nouvelle direction comme étant celle qui maximise la variance projetée parmi toutes les directions possibles orthogonales à celles déjà considérées.

Nous pouvons démontrer facilement que les vecteurs directeurs des composantes principales correspondent aux vecteurs propres $\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_d}$.

Remarque :

Σ est symétrique $\Rightarrow \boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_d}$ forment une base orthogonale

Analyse en composantes principales

L'espace de projection (l'espace des axes principaux) est alors obtenu en prenant les K premiers vecteurs propres $\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_K}$

Comment choisir K

Choisir les K composantes permettant d'avoir une proportion de la variance expliquée supérieur à un certain seuil (0.9, 0.95 , etc.)

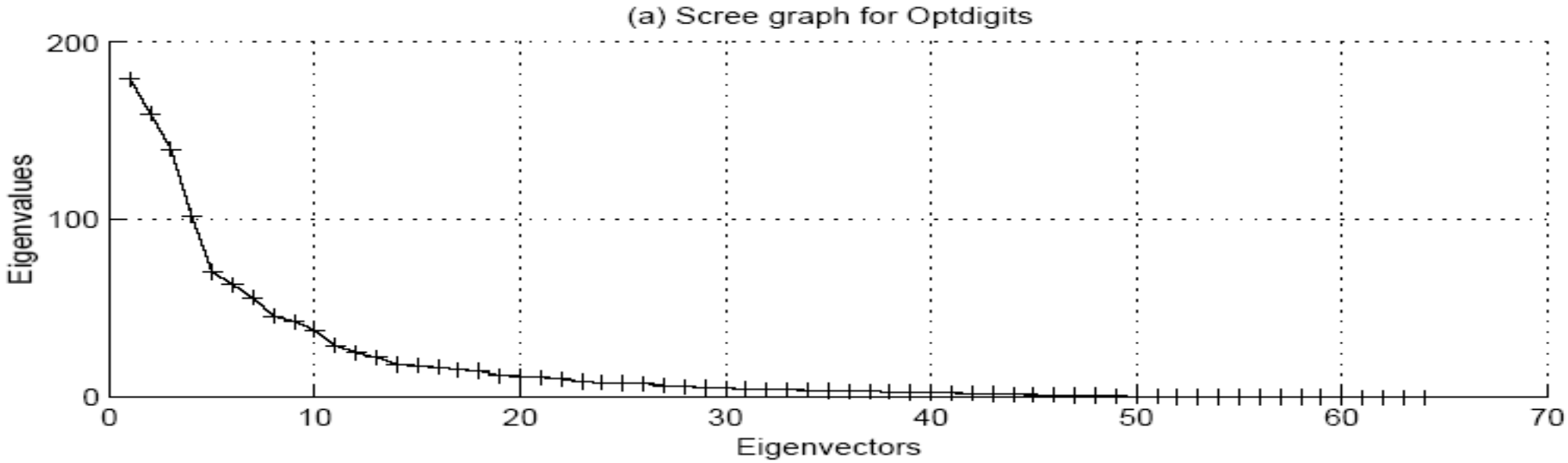
$$\text{Proportion de la variance expliquée} = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{d} \lambda_j}$$

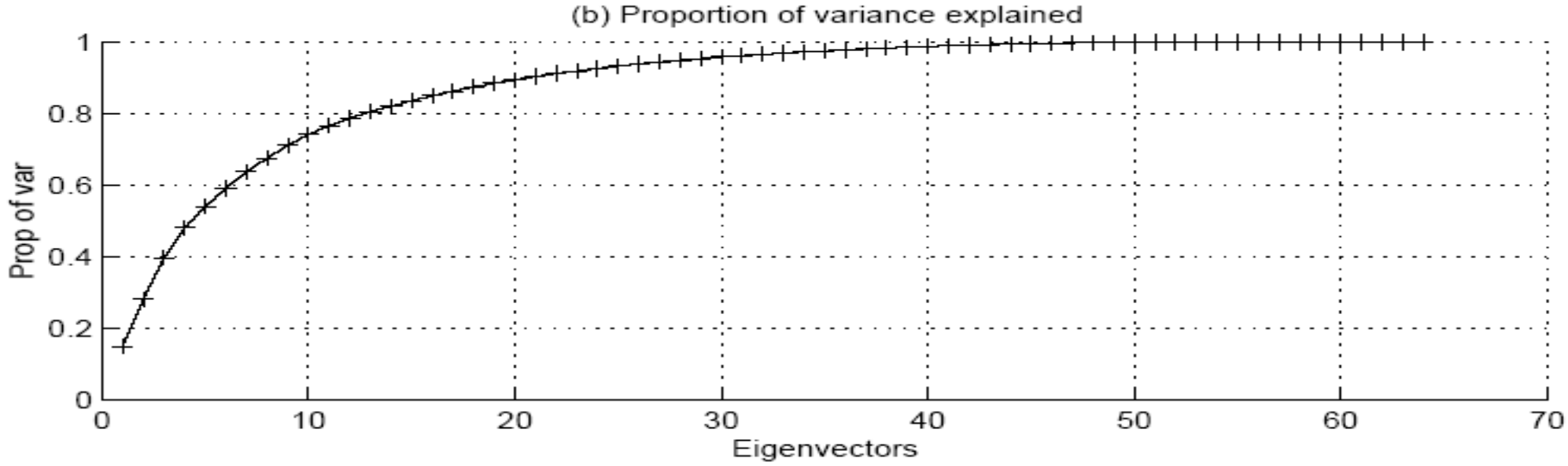
Avec

Variance totale (inertie totale) des données projetés $\sum_{j=1}^{d} \lambda_j$

Variance des données projetées sur K axes principaux $\sum_{j=1}^{K} \lambda_j$

Analyse en composantes principales





La mise en œuvre de l'ACP peut être divisée en 6 étapes principales :

1. Calculer les centres des données

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

2. Calculer la matrice de covariance

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$

3. Extraire les valeurs et vecteurs propres de Σ

4. Classer les vecteurs propres dans l'ordre décroissant des valeurs propres associées

$$(\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_d}) \text{ Avec } \lambda_1 > \lambda_2 > \dots > \lambda_d$$

5. Choisir K composantes principales vérifiant $\frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{d} \lambda_j} \geq$ seuil

6. Projeter les données sur le nouveaux espace $Z = \boldsymbol{U}^T X$

Avec $X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_i}, \dots, \boldsymbol{x_n})$ et **U=** $(\boldsymbol{u_1}^T, \boldsymbol{u_2}^T, \dots, \boldsymbol{u_K}^T)$

Analyse factorielle

L'**analyse factorielle** est un outils très pertinent dans la mise en évidence des relations entre variables et, plus généralement, dans la compréhension des données à modéliser. Cet méthode vise à réduire la **dimension** des données (le nombre de variables) en conservant au mieux l'information utile.

Objectif général : recherche de « facteurs » (variables dérivées) permettant de résumer les (caractéristiques) des données

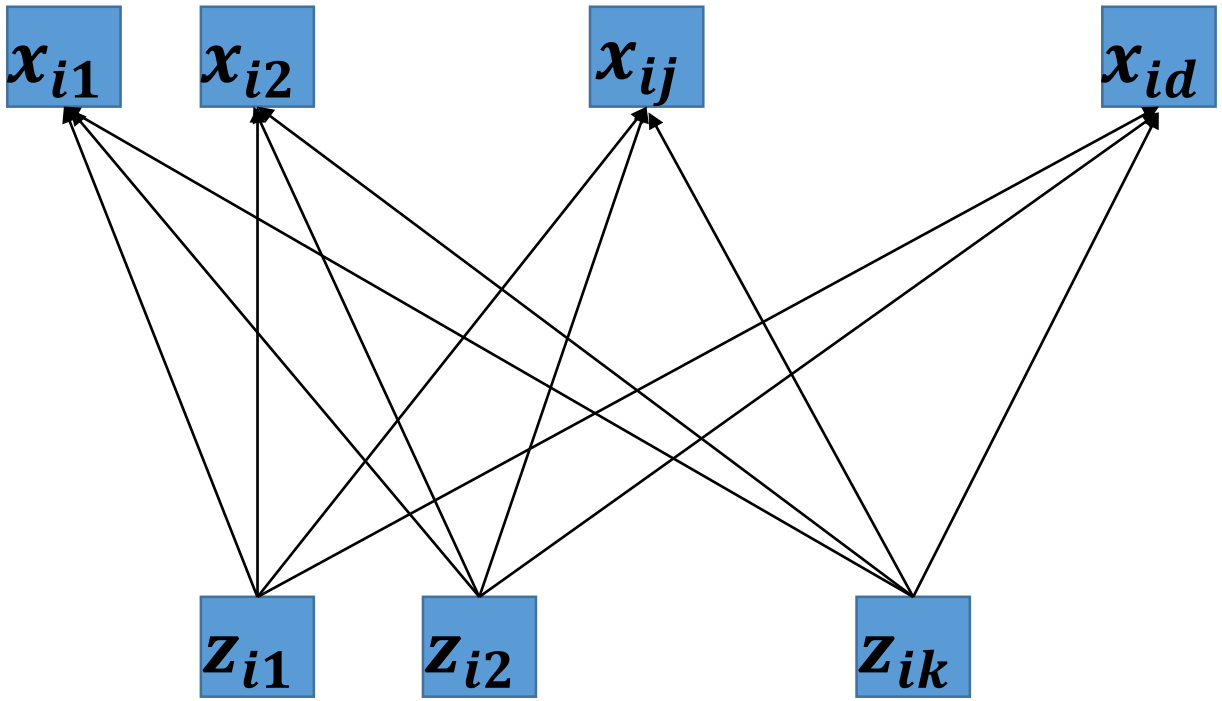
- Améliorer la « lisibilité » des données
- Réduire le nombre de variables en conservant au mieux l'information utile

Analyse factorielle

- Soit $X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_i}, \ldots, \boldsymbol{x_n})$ avec $\boldsymbol{x_i} \in \mathbb{R}^d$
- On fait l'hypothèse que ces variables dépendent (linéairement) en partie de k variables non observables, ou variables latentes ou facteurs $\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_k}$.

On cherche donc à décomposer les variables observées $\boldsymbol{x_i}$ (supposées centrées) de la façon suivante

$$\boldsymbol{x_i} - \mu = \boldsymbol{W z_i} + \boldsymbol{\varepsilon_i}$$



$\boldsymbol{x_i} \in \mathbb{R}^d$

$\boldsymbol{W} \in \mathbb{R}^{dxk}$

$\boldsymbol{z_i} \in \mathbb{R}^k$

Analyse factorielle

Avec

$$\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{id})^T$$
$$\boldsymbol{\mu} = (\mu_1, \dots, \mu_j, \dots, \mu_d)^T$$
$$\boldsymbol{z_i} = (z_{i1}, z_{i2}, \dots, z_{ik})^T$$
$$\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{ij}, \dots, \varepsilon_{id})^T$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w}_{11} & \boldsymbol{w}_{12} & \dots & \boldsymbol{w}_{1k} \\ \boldsymbol{w}_{21} & \boldsymbol{w}_{22} & \dots & \boldsymbol{w}_{2k} \\ & & \vdots & \\ \boldsymbol{w}_{j1} & \boldsymbol{w}_{j2} & \dots & \boldsymbol{w}_{jk} \\ & & \vdots & \\ \boldsymbol{w}_{d1} & \boldsymbol{w}_{d2} & \dots & \boldsymbol{w}_{dk} \end{bmatrix}$$

Analyse factorielle

$$\boldsymbol{x}_i = \boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu} + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{z_i} \sim \mathcal{N}(0, \boldsymbol{I})$$

$$\boldsymbol{\varepsilon_i} \sim \mathcal{N}(0, \boldsymbol{\psi})$$

$$\boldsymbol{x_i}|\boldsymbol{z_i} \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\psi})$$

$$\boldsymbol{x_i} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\psi})$$

$\boldsymbol{\psi}$ est une matrice diagonale

Analyse factorielle

Estimation des paramètre du modèle (W, ψ)

- Méthode basée sur l'ACP
- Méthode basée sur le maximum de vraisemblance
- Méthode basée sur les facteurs principaux

Estimation des paramètre du modèle ($\boldsymbol{W}, \boldsymbol{\psi}$)

- Méthode basée sur l'ACP

On sait que

$$\boldsymbol{Var}\,(\boldsymbol{x}_i) = \boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\psi}$$

$$\boldsymbol{\Sigma} = \sum_{j=1}^{d} \lambda_j \, \boldsymbol{u_j}\boldsymbol{u_j}^T$$

$$= \underbrace{[\sqrt{\lambda_1}\boldsymbol{u_1}, \sqrt{\lambda_2}\boldsymbol{u_2}, \ldots, \sqrt{\lambda_k}\boldsymbol{u_k} \ldots\ldots \sqrt{\lambda_d}\boldsymbol{u_d}]}_{\boldsymbol{W}} \left.\begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{u_1}^T \\ \sqrt{\lambda_2}\boldsymbol{u_2}^T \\ \vdots \\ \sqrt{\lambda_k}\boldsymbol{u_k}^T \\ \vdots \\ \sqrt{\lambda_d}\boldsymbol{u_d}^T \end{bmatrix}\right\} \boldsymbol{W}^T$$

$\lambda_1, \lambda_2, \ldots, \lambda_d$: valeurs propres de la matrice de covariance Σ

$\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_d}$: vecteurs propres de la matrice de covariance Σ

Estimation des paramètre du modèle ($\boldsymbol{W}$, $\boldsymbol{\psi}$)

- Méthode basée sur l'ACP

$$\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{W}^T + [\sqrt{\lambda_{k+1}}\boldsymbol{u}_{k+1}, \ldots\ldots \sqrt{\lambda_d}\boldsymbol{u}_d] \begin{bmatrix} \sqrt{\lambda_{k+1}}\boldsymbol{u}_{k+1} \\ \vdots \\ \sqrt{\lambda_d}\boldsymbol{u}_d{}^T \end{bmatrix}$$

ainsi

$$\boldsymbol{\psi} = \text{diag}\{[\sqrt{\lambda_{k+1}}\boldsymbol{u}_{k+1}, \ldots\ldots \sqrt{\lambda_d}\boldsymbol{u}_d] \begin{bmatrix} \sqrt{\lambda_{k+1}}\boldsymbol{u}_{k+1} \\ \vdots \\ \sqrt{\lambda_d}\boldsymbol{u}_d{}^T \end{bmatrix}\}$$

Ou bien

$$\boldsymbol{\psi} = \text{diag}\{\boldsymbol{\Sigma} - \boldsymbol{W}\boldsymbol{W}^T\}$$

Estimation des paramètre du modèle ($\boldsymbol{W}$, $\boldsymbol{\psi}$)

- Méthode basée sur le maximum de vraisemblance

$$\boldsymbol{x_i} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{WW}^T + \boldsymbol{\psi})$$

Log- vraisemblance de $\boldsymbol{x}$

$$\mathcal{L}(\theta; \boldsymbol{x}) = log \prod_{i=1}^{n} p(x_i|\theta)$$

$$\mathcal{L}(\theta;\boldsymbol{x}) = log \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{WW}^T + \boldsymbol{\psi})$$

$$\mathcal{L}(\theta;\boldsymbol{x}) = \frac{\mathrm{nd}}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{WW}^T + \boldsymbol{\psi}| \; \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T(\boldsymbol{WW}^T + \boldsymbol{\psi})^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

Estimation des paramètre du modèle ($\boldsymbol{W}, \boldsymbol{\psi}$)

- Méthode basée sur le maximum de vraisemblance

Log- vraisemblance des données complétées

$$\mathcal{L}_c(\theta;\boldsymbol{x},\boldsymbol{z}) = log \prod_{i=1}^{n} p(x_i|z_i;\theta)p(z_i;\theta)$$

$$= log \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu},\boldsymbol{\psi})\ \mathcal{N}(0,\boldsymbol{I})$$

$$= \sum_{i=1}^{n} log\mathcal{N}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu},\boldsymbol{\psi})\ + log\mathcal{N}(0,\boldsymbol{I})$$

Estimation des paramètre du modèle ($\boldsymbol{W}$, $\boldsymbol{\psi}$)

- Méthode basée sur le maximum de vraisemblance

Algorithme EM

Etape E

$$\mathrm{E}[\boldsymbol{z_i}] = (\mathrm{I} + \boldsymbol{W}\boldsymbol{\psi}^{-1}\boldsymbol{W}^T)^{-1}\boldsymbol{W}^T(\boldsymbol{\psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}))$$

$$\mathrm{E}[\boldsymbol{z_i}\boldsymbol{z_i}^T] = (\mathrm{I} + \boldsymbol{W}\boldsymbol{\psi}^{-1}\boldsymbol{W}^T)^{-1} + \mathrm{E}[\boldsymbol{z_i}]\mathrm{E}[\boldsymbol{z_i}]^T$$

Etape M

$$\boldsymbol{W}_{\boldsymbol{q}} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})\mathrm{E}[\boldsymbol{z_i}]^T \sum_{i=1}^{n}\mathrm{E}[\boldsymbol{z_i}\boldsymbol{z_i}^T]$$

$$\boldsymbol{\psi}_{\boldsymbol{q}} = diag\left\{\Sigma - \boldsymbol{W}_{\boldsymbol{q}}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\boldsymbol{z_i}](\boldsymbol{x}_i - \boldsymbol{\mu})^T\right\}$$