SD 204
SVD / PCA

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

Sommaire

La décomposition spectrale

Théorème spectral

Une matrice symétrique $S \in \mathbb{R}^{n \times n}$ est diagonalisable en base orthonormée, *i.e.*, il existe $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ et une matrice orthogonale $U \in \mathbb{R}^{n \times n}$ telle que :

$$S = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n) U^\top \text{ ou } SU = U \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

Rem: Si l'on écrit $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ cela signifie que :

$$S = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$$

De plus $\forall i \in [\![1, n]\!], \quad S\mathbf{u}_i = \lambda_i \mathbf{u}_i$

Rappel : une matrice orthogonale $U \in \mathbb{R}^n$ est une matrice telle que $U^\top U = UU^\top = \mathrm{Id}_n$ ou $\forall i, j = 1, \ldots, n, \mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$

Vocabulaire : les λ_i sont les **valeurs propres** de S et les $\mathbf{u}_i \in \mathbb{R}^n$ sont les **vecteurs propres** associés

La décomposition en valeurs singulières (: *Singular Value Decomposition, SVD*)

Théorème

Pour toute matrice $X \in \mathbb{R}^{n \times p}$, il existe une matrice orthogonale $U \in \mathbb{R}^{n \times n}$ et une matrice orthogonale $V \in \mathbb{R}^{p \times p}$, telles que

$$U^\top X V = \operatorname{diag}(s_1, \ldots, s_{\min(n,p)}) = \Sigma \in \mathbb{R}^{n \times p}$$

avec $s_1 \geqslant s_2 \geqslant \ldots \geqslant s_{\min(n,p)} \geqslant 0$, ou encore :

$$\boxed{X = U \Sigma V^\top}$$

avec $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ et $V = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$

Rappel : $\begin{cases} \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}, & \forall i, j \in [\![1, n]\!] \\ \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{i,j}, & \forall i, j \in [\![1, p]\!] \end{cases}$

Démonstration : diagonaliser $X^\top X$ Golub et Van Loan (1996)

SVD la suite

Vocabulaire : les s_j sont les **valeurs singulières** de X ; les $\mathbf{u}_j$ (resp. $\mathbf{v}_j$) sont les **vecteurs singuliers** à gauche (resp. droite)

Propriété variationnelle de la plus grande valeur singulière

$$s_1 = \begin{cases} \max\limits_{\mathbf{u}\in\mathbb{R}^n, \mathbf{v}\in\mathbb{R}^p} \mathbf{u}^\top X \mathbf{v} \\ \text{s.c.} \|\mathbf{u}\|^2 = 1 \text{ et } \|\mathbf{v}\|^2 = 1 \end{cases}$$

Lagrangien : $\mathcal{L}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top X \mathbf{v} - \lambda_1(\|\mathbf{u}\|^2 - 1) - \lambda_2(\|\mathbf{v}\|^2 - 1)$

CNO : $\begin{cases} \nabla_{\mathbf{u}} \mathcal{L} = X\mathbf{v} - 2\lambda_1 \mathbf{u} = 0 \\ \nabla_{\mathbf{v}} \mathcal{L} = X^\top \mathbf{u} - 2\lambda_2 \mathbf{v} = 0 \end{cases} \Longleftrightarrow \begin{cases} X\mathbf{v} = 2\lambda_1 \mathbf{u} \\ X^\top \mathbf{u} = 2\lambda_2 \mathbf{v} \end{cases} \Rightarrow \begin{cases} X^\top X \mathbf{v} = \alpha \mathbf{v} \\ X X^\top \mathbf{u} = \alpha \mathbf{u} \end{cases}$

avec $\alpha = 2\lambda_1\lambda_2$, et donc $\mathbf{v}$ et $\mathbf{u}$ sont des vecteurs propres de $X^\top X$ et de $XX^\top$

La SVD toujours et encore

SVD compacte

On ne garde que les éléments non-nuls de la diagonale

$$\boxed{X = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^\top = U_r \operatorname{diag}(s_1, \dots, s_r) V_r^\top}$$

avec $s_i > 0, \forall i \in [\![1, r]\!]$ et $U_r = [\mathbf{u}_1, \dots, \mathbf{u}_r]$, $V_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$

<u>Rem</u>: $r = \operatorname{rg}(X)$ nombre de valeurs singulières (non-nulles)
<u>Rem</u>: les matrices $\mathbf{u}_i \mathbf{v}_i^\top$ sont toutes de rang 1
<u>Rem</u>: les vecteurs $\mathbf{u}_i$ (resp. les vecteurs $\mathbf{v}_i^\top$) sont des vecteurs orthonormaux qui engendrent le même espace que celui engendré par les colonnes (resp. les lignes) de X

$$\boxed{\operatorname{vect}(\mathbf{x}_1, \dots, \mathbf{x}_p) = \operatorname{vect}(\mathbf{u}_1, \dots, \mathbf{u}_r)}$$

SVD et meilleure approximation

Théorème (meilleure approximation de rang k)

Prenons la SVD de $X \in \mathbb{R}^{n \times p}$ donnée par $X = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^\top$ (*i.e.*, $r = \mathrm{rg}(X)$). Si $k < r$ et si $X_k = \sum_{i=1}^{k} s_i \mathbf{u}_i \mathbf{v}_i^\top$ alors

$$\min_{Z \in \mathbb{R}^{n \times p} \,:\, \mathrm{rg}(Z) = k} |||X - Z|||_2 = |||X - X_k|||_2 = s_{k+1}$$

Rem: la norme spectrale de X est définie par

$$|||X|||_2 = \sup_{u \in \mathbb{R}^p, \|u\| = 1} \|Xu\| = s_1(X)$$

Rem: ce théorème est aussi crucial pour l'analyse en composante principale (ACP)

Sommaire

Pseudo-inverse

Définition

Si $X \in \mathbb{R}^{n \times p}$ admet pour SVD $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$ alors sa **pseudo-inverse** $X^+ \in \mathbb{R}^{p \times n}$ est définie par :

$$X^+ = \sum_{i=1}^{r} \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$$

Rem: Si $X \in \mathbb{R}^{n \times n}$ est inversible (*i.e.*, de rang n) alors $X = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{v}_i^\top$ et alors $X^+ = X^{-1}$

Démonstration :

$$\begin{aligned} XX^+ &= \sum_{j=1}^{n} s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^{n} \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} s_j \frac{1}{s_i} \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^{n} \sum_{i=1}^{n} s_j \frac{1}{s_i} \delta_{i,j} \mathbf{u}_j \mathbf{u}_i^\top = \sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top = \mathrm{Id}_n \end{aligned}$$

SVD et numérique

Les fonctions SVD et pseudo-inverse sont disponibles dans toutes librairies numériques, par exemple `Numpy`

- Pseudo-inverse : `U, s, V = np.linalg.svd(X)`
  Attention dans ce cas :
  `X=np.dot(U, np.dot(np.diag(S), V))`
  Il y a aussi plusieurs variantes matrice pleine ou non
  cf. `full_matrices=True/False`
- Pseudo-inverse : `Xinv = np.linalg.pinv(X)`

Exo: Vérifier numériquement le théorème de meilleure approximation de rang fixé pour une matrice tirée aléatoirement selon une loi gaussienne (*e.g.*, de taille 9×6, pour $k = 3$)

Sommaire

Quelques mots de stabilité numérique

Prenons $\hat{\boldsymbol{\theta}} = X^{+}\mathbf{y}$ comme solution des moindres carrés.
Supposons qu'on observe maintenant non plus $\mathbf{y}$ mais $\mathbf{y} + \Delta$ où Δ est une erreur très petite : $\|\Delta\| \ll \|\mathbf{y}\|$.
Alors l'estimateur des moindres carrés pour $\mathbf{y} + \Delta$ par X donne

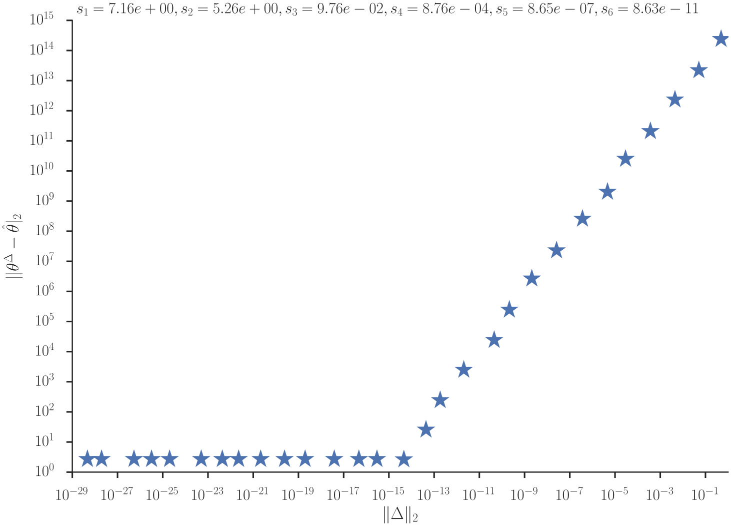
$$\hat{\boldsymbol{\theta}}^{\Delta} = X^{+}(\mathbf{y} + \Delta)$$

$$\hat{\boldsymbol{\theta}}^{\Delta} = \hat{\boldsymbol{\theta}} + X^{+}\Delta$$

$$\hat{\boldsymbol{\theta}}^{\Delta} = \hat{\boldsymbol{\theta}} + \sum_{i=1}^{r} \frac{1}{s_i}\mathbf{v}_i\mathbf{u}_i^{\top}\Delta$$

Exemple de problème de conditionnement

$X \in \mathbb{R}^{10\times 6}$ dont les valeurs singulières sont ci-dessous :



Prochains cours : remèdes possibles

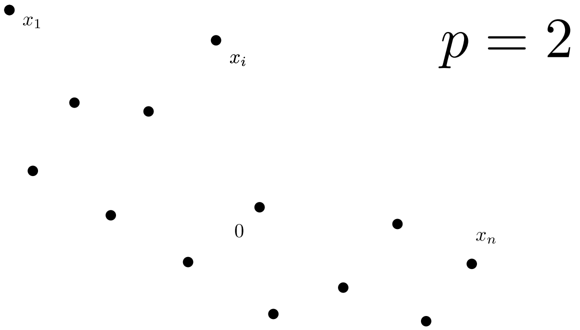
- Régulariser le spectre / les valeurs singulières
- Contraindre les coefficients de $\hat{\boldsymbol{\theta}}$ à n'être pas trop grands

Une solution rendant ces deux points de vue équivalents : *Ridge Regression* / Régularisation de Tychonoff

Sommaire

ACP

On observe n points $x_1, \ldots, x_n$ dans $\mathbb{R}^p$, ainsi on créé une matrice $X = [x_1, \ldots, x_n]^\top$ matrice $n \times p$: n observations (lignes), p *features* (colonnes)



<u>Rem</u>: on doit recentrer les points pour qu'ils aient une moyenne nulle $X \leftarrow [x_1 - \overline{x}_n, \ldots, x_n - \overline{x}_n]^\top = X - \mathbf{1}_n \overline{x}_n^\top$ (on peut aussi mettre à l'échelle pour avoir un écart-type similaire par *feature*)

Analyse en Composante Principale, ACP (: *Principal Component Analysis, PCA*)

Paramètre k : nombre d'axes pour représenter un nuage de n points $(x_1, \ldots, x_n)$, représentés par les lignes de $X \in \mathbb{R}^{n \times p}$.

Cette méthode **compresse** le nuage de points de dimension p en un nuage de dimension k

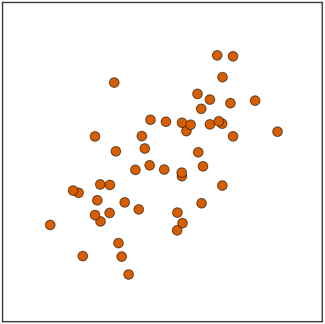
L'ACP (de niveau k) consiste à effectuer la SVD de X, et à ne garder que les k axes principaux pour représenter le nuage.

$$X = \sum_{i=1}^{r} s_i \mathbf{u}_i \mathbf{v}_i^\top \longrightarrow \sum_{i=1}^{k} s_i \mathbf{u}_i \mathbf{v}_i^\top$$

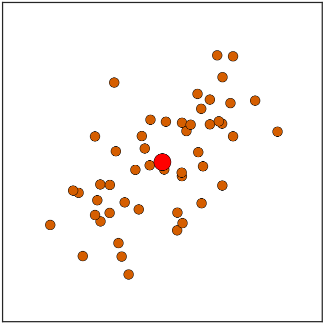
On appelle **axes principaux** les k vecteurs $\mathbf{v}_1, \ldots, \mathbf{v}_k$, et en général $k \ll p$ (*e.g.*, $k = 2$, pour une visualisation planaire)

Sommaire

Axe principal : maximisation de la variance

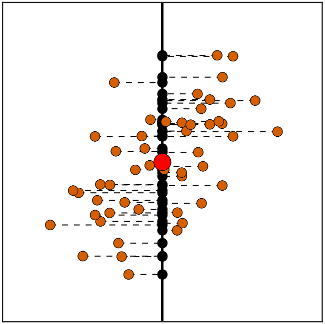


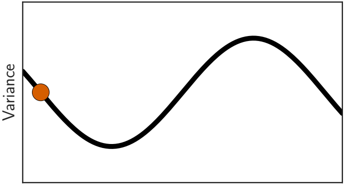
Axe principal : maximisation de la variance



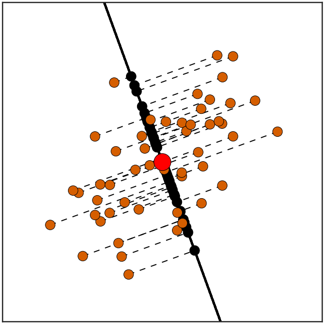
Data and mean

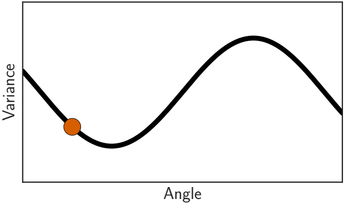
Axe principal : maximisation de la variance



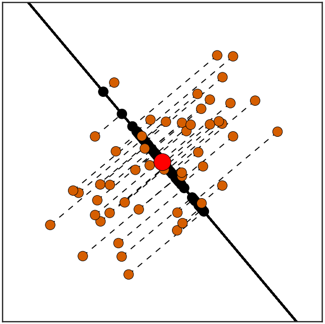


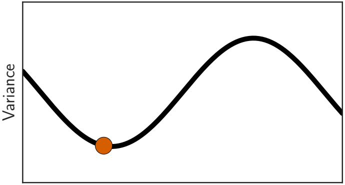
Axe principal : maximisation de la variance


Data, mean and projection

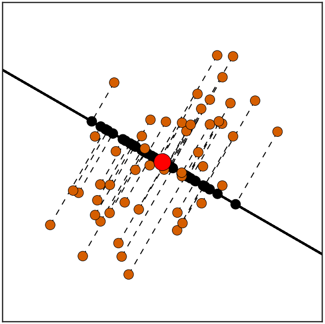


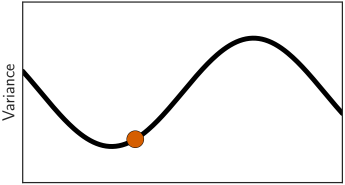
Axe principal : maximisation de la variance



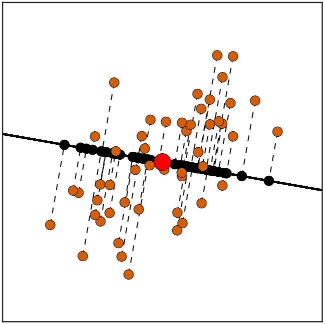


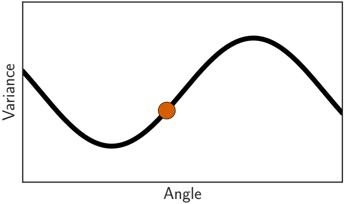
Axe principal : maximisation de la variance



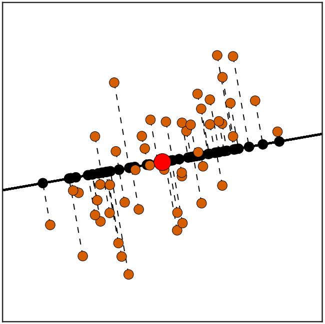


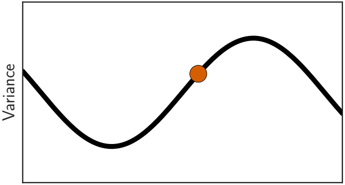
Axe principal : maximisation de la variance



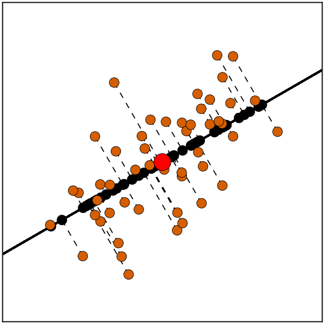


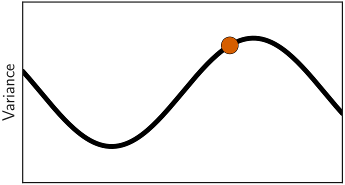
Axe principal : maximisation de la variance





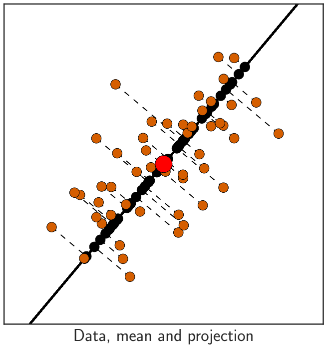
Axe principal : maximisation de la variance

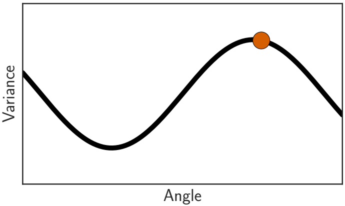

Data, mean and projection



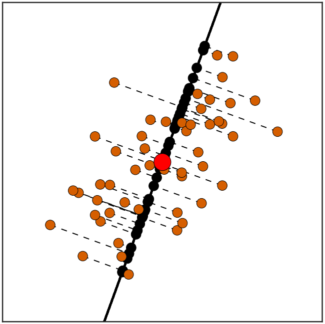
Angle

Axe principal : maximisation de la variance

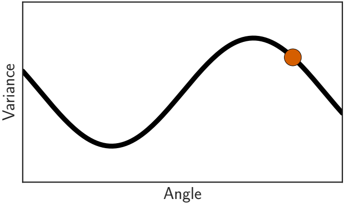




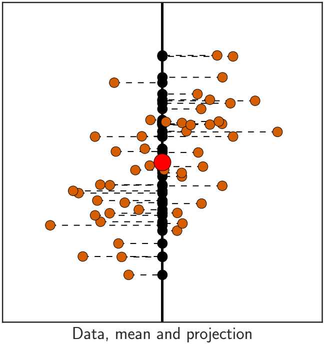
Axe principal : maximisation de la variance

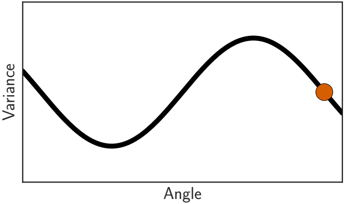


Data, mean and projection

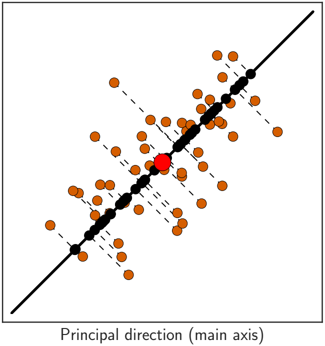


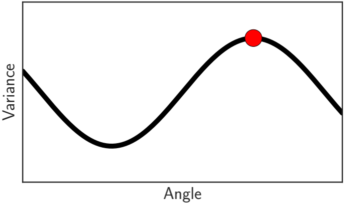
Axe principal : maximisation de la variance





Axe principal : maximisation de la variance





ACP : axe principal

L'axe principal (normalisé) $\mathbf{v}_1$ est la solution du problème :

$$\mathbf{v}_1 \in \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \ \mathbf{v}^\top X^\top X \mathbf{v} = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \ \|X\mathbf{v}\|^2 = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{i=1}^{n} (x_i^\top \mathbf{v})^2$$

Rem: après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe $\mathbf{v}$

Algorithme : Méthode de la puissance itérée

Entrées : $X \in \mathbb{R}^{n\times p}$, itérations K

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

L'axe principal (normalisé) $\mathbf{v}_1$ est la solution du problème :

$$\mathbf{v}_1 \in \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \; \mathbf{v}^\top X^\top X \mathbf{v} = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \; \|X\mathbf{v}\|^2 = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{i=1}^n (x_i^\top \mathbf{v})^2$$

<u>Rem</u>: après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe $\mathbf{v}$

Algorithme : Méthode de la puissance itérée

Entrées : $X \in \mathbb{R}^{n\times p}$, itérations K
$\mathbf{v}$ tiré aléatoirement dans $\mathbb{R}^{n\times p}$ (*e.g.*, $u/\|u\|$ avec u gaussien)

<u>Rem</u>: on résout une <u>maximisation</u> sous contrainte convexe

ACP : axe principal

L'axe principal (normalisé) $\mathbf{v}_1$ est la solution du problème :

$$\mathbf{v}_1 \in \operatorname*{arg\,max}_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \mathbf{v}^\top X^\top X \mathbf{v} = \operatorname*{arg\,max}_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \|X\mathbf{v}\|^2 = \operatorname*{arg\,max}_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \sum_{i=1}^n (x_i^\top \mathbf{v})^2$$

Rem: après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe $\mathbf{v}$

Algorithme : Méthode de la puissance itérée

Entrées : $X \in \mathbb{R}^{n \times p}$, itérations K

$\mathbf{v}$ tiré aléatoirement dans $\mathbb{R}^{n \times p}$ (*e.g.*, $u/\|u\|$ avec u gaussien)

pour $k = 1, \ldots, K$ **faire**

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

L'axe principal (normalisé) $\mathbf{v}_1$ est la solution du problème :

$$\mathbf{v}_1 \in \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \; \mathbf{v}^\top X^\top X \mathbf{v} = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \; \|X\mathbf{v}\|^2 = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{i=1}^{n} (x_i^\top \mathbf{v})^2$$

<u>Rem</u>: après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe $\mathbf{v}$

Algorithme : Méthode de la puissance itérée

Entrées : $X \in \mathbb{R}^{n\times p}$, itérations K

$\mathbf{v}$ tiré aléatoirement dans $\mathbb{R}^{n\times p}$ (*e.g.*, $u/\|u\|$ avec u gaussien)

pour $k = 1, \dots, K$ **faire**

  $\mathbf{w} \leftarrow X\mathbf{v}$

<u>Rem</u>: on résout une <u>maximisation</u> sous contrainte convexe

ACP : axe principal

L'axe principal (normalisé) $\mathbf{v}_1$ est la solution du problème :

$$\mathbf{v}_1 \in \underset{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \ \mathbf{v}^\top X^\top X \mathbf{v} = \underset{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \ \|X\mathbf{v}\|^2 = \underset{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{i=1}^{n} (x_i^\top \mathbf{v})^2$$

Rem: après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe $\mathbf{v}$

Algorithme : Méthode de la puissance itérée

Entrées : $X \in \mathbb{R}^{n \times p}$, itérations K

$\mathbf{v}$ tiré aléatoirement dans $\mathbb{R}^{n \times p}$ (*e.g.*, $u/\|u\|$ avec u gaussien)

pour $k = 1, \ldots, K$ **faire**

  $\mathbf{w} \leftarrow X\mathbf{v}$

  $\mathbf{v} \leftarrow X^\top \mathbf{w}$

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

L'axe principal (normalisé) $\mathbf{v}_1$ est la solution du problème :

$$\mathbf{v}_1 \in \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \ \mathbf{v}^\top X^\top X \mathbf{v} = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \ \|X\mathbf{v}\|^2 = \underset{\mathbf{v}\in\mathbb{R}^p, \|\mathbf{v}\|=1}{\arg\max} \sum_{i=1}^{n} (x_i^\top \mathbf{v})^2$$

Rem: après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe $\mathbf{v}$

Algorithme : Méthode de la puissance itérée

Entrées : $X \in \mathbb{R}^{n\times p}$, itérations K

$\mathbf{v}$ tiré aléatoirement dans $\mathbb{R}^{n\times p}$ (*e.g.*, $u/\|u\|$ avec u gaussien)

pour $k = 1, \dots, K$ **faire**

$\quad \mathbf{w} \leftarrow X\mathbf{v}$

$\quad \mathbf{v} \leftarrow X^\top \mathbf{w}$

$\quad \mathbf{v} \leftarrow \frac{\mathbf{v}}{\|\mathbf{v}\|}$

Sorties : Axe principale (approché) $\mathbf{v}_1 = \mathbf{v}$

Rem: on résout une maximisation sous contrainte convexe

Premier axe principal

Maximiser la fonction objectif suivante en $\mathbf{v}$:
$\mathcal{L}(\mathbf{v}, \lambda) = (X\mathbf{v})^\top (X\mathbf{v}) - \lambda(\mathbf{v}^\top \mathbf{v} - 1) = \mathbf{v}^\top X^\top X \mathbf{v} - \lambda(\mathbf{v}^\top \mathbf{v} - 1)$

λ : multiplicateur de Lagrange
Conditions d'optimalité du premier ordre en un extremum

$$\frac{\partial \mathcal{L}(\mathbf{v}_1, \lambda)}{\partial \mathbf{v}} = 0 \Leftrightarrow X^\top X \mathbf{v}_1 = \lambda \mathbf{v}_1$$

La matrice de Gram $X^\top X$ est diagonalisable (symétrique) donc si $\mathbf{v}_1$ est un extremum alors c'est un vecteur propre.

Rem: on normalise $\mathbf{v}_1$ pour que $\|\mathbf{v}_1\| = 1$, ainsi $\lambda = \mathbf{v}_1^\top X^\top X \mathbf{v}_1$ et $\mathbf{v}_1$ est un vecteur propre, de valeur propre λ maximale

Aspect récursif de l'ACP - Déflation

Construction récursive : définir les axes principaux en partant du plus important et en descendant

Par récurrence, on définit le k^{e} axe pour qu'il soit orthogonal aux axes principaux précédents :

$$\mathbf{v}_k = \underset{\mathbf{v}\in\mathbb{R}^p,\, \mathbf{v}^\top\mathbf{v}_1=\cdots=\mathbf{v}^\top\mathbf{v}_{k-1}=0, \|\mathbf{v}\|=1}{\arg\max} \|X\mathbf{v}\|^2$$

- le premier axe maximise la variance des données projetées sur l'axe porté par ce vecteur
- le deuxième axe est celui orthogonal au premier, de variance projetée maximale
- etc.

Nouvelle représentation des données

- Les axes (de direction) $\mathbf{v}_1, \ldots, \mathbf{v}_p \in \mathbb{R}^p$ sont appelés **axes principaux** ou **axes factoriels**, les nouvelles variables $\mathbf{c}_j = X\mathbf{v}_j, j = 1, \ldots, p$ sont appelées **composantes principales**

Nouvelle représentation :

- La matrice XV_k (avec $V_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$) est la matrice représentant les données dans la base des k premiers vecteurs propres

Reconstruction dans l'espace original (débruiter) :

- Reconstruction "parfaite" pour $\mathbf{x} \in \mathbb{R}^p$: $\mathbf{x} = \sum_{j=1}^{p} (\mathbf{x}^\top \mathbf{v}_j)\mathbf{v}_j$
- Reconstruction avec perte d'information : $\hat{\mathbf{x}} = \sum_{j=1}^{k} (\mathbf{x}^\top \mathbf{v}_j)\mathbf{v}_j$

Références I

- G. H. Golub and C. F. van Loan.
  Matrix computations.
  Johns Hopkins University Press, Baltimore, MD, third edition, 1996.