

MS BGD

MDI 720 : ACP

Joseph Salmon

<http://josephsalmon.eu>

Télécom Paristech, Institut Mines-Télécom

Plan

ACP

Définition

Interprétation et récursion

Sommaire

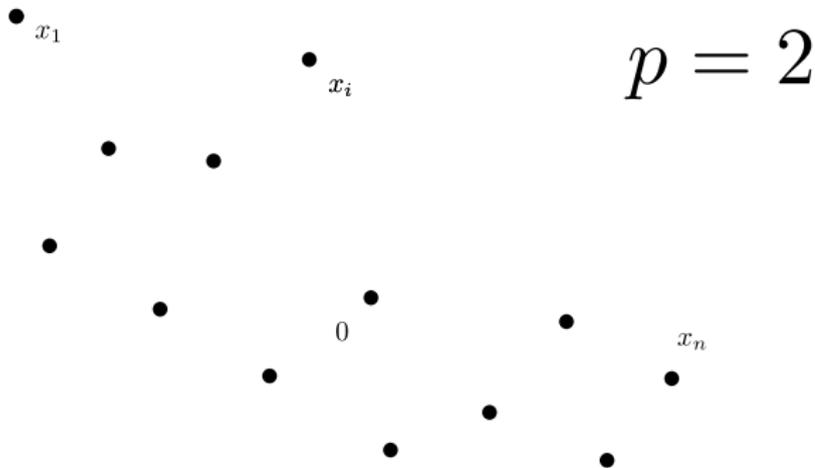
ACP

Définition

Interprétation et récursion

ACP

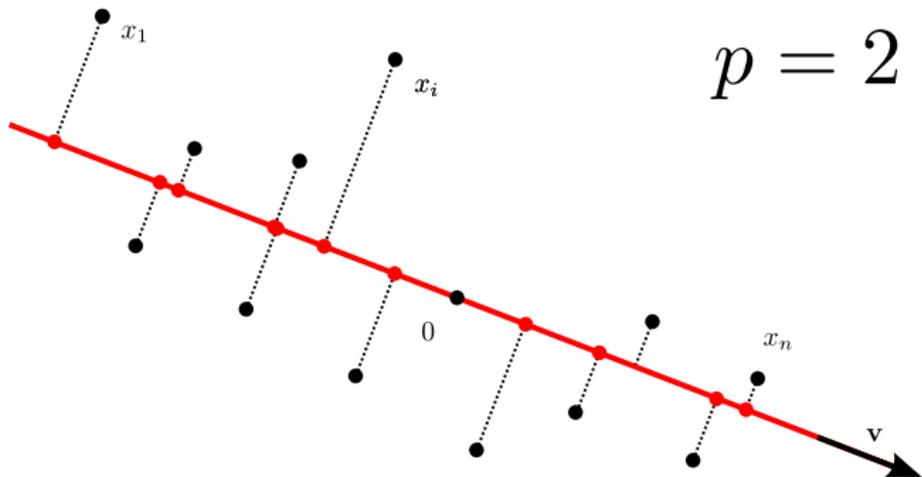
On observe n points x_1, \dots, x_n dans \mathbb{R}^p , ainsi on crée une matrice $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$, n observations (lignes), p *features* (colonnes)



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

ACP

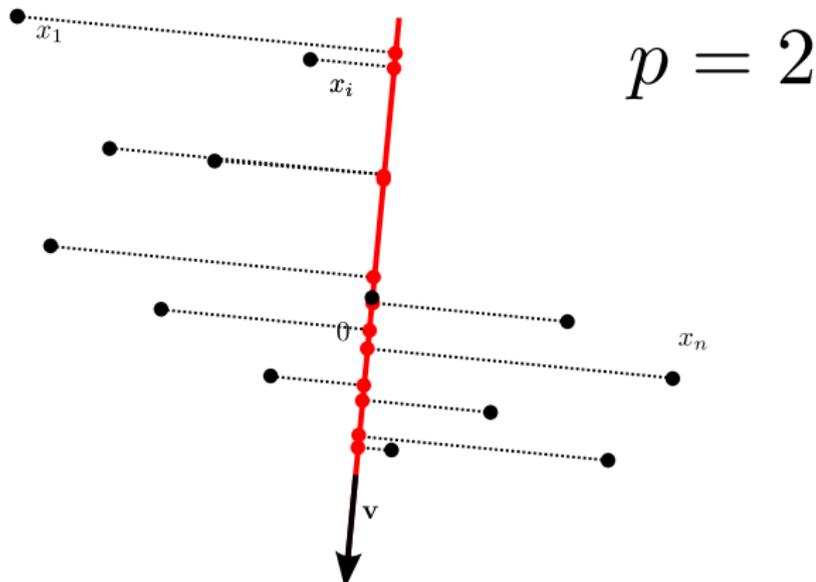
On observe n points x_1, \dots, x_n dans \mathbb{R}^p , ainsi on crée une matrice $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times p}$, n observations (lignes), p features (colonnes)



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

ACP

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



Rem: on doit recentrer les points pour qu'ils aient une moyenne nulle $X \leftarrow [x_1 - \bar{x}_n, \dots, x_n - \bar{x}_n]^\top = X - \mathbf{1}_n \bar{x}_n^\top$ (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, \dots, 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 $\mathbf{v}_1, \dots, \mathbf{v}_k$, et en général $k \ll p$ (e.g., $k = 2$, pour une visualisation planaire)

Nouvelle représentation des données

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

Nouvelle représentation (ordre k) :

- ▶ La matrice XV_k (avec $V_k = [\mathbf{v}_1, \dots, \mathbf{v}_k]$) 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 : $\hat{\mathbf{x}} = \sum_{j=1}^k (\mathbf{x}^\top \mathbf{v}_j) \mathbf{v}_j$

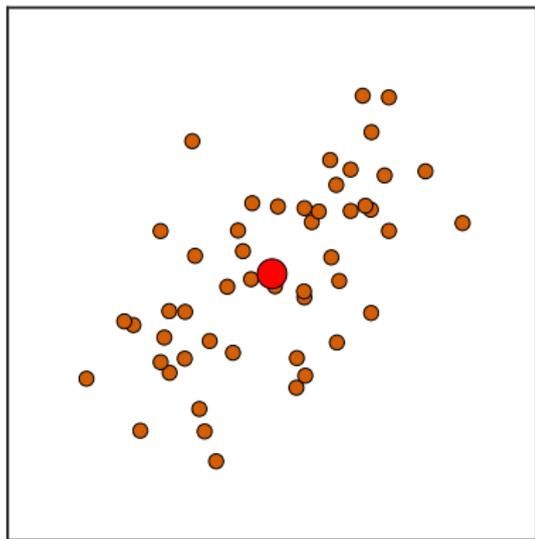
Sommaire

ACP

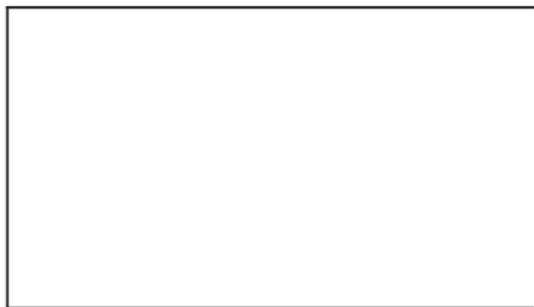
Définition

Interprétation et récursion

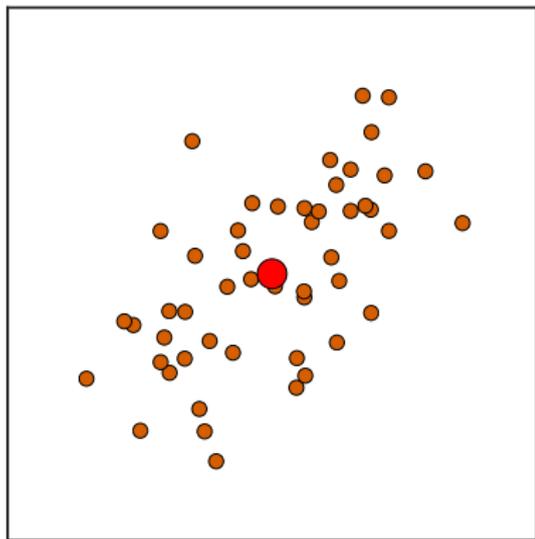
Axe principal : maximisation de la variance



Data and mean



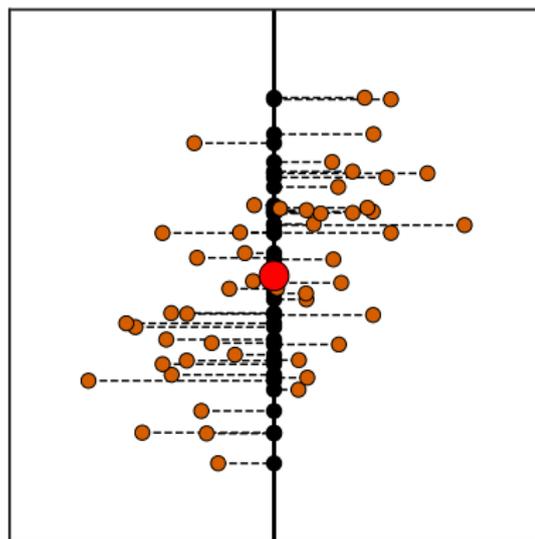
Axe principal : maximisation de la variance



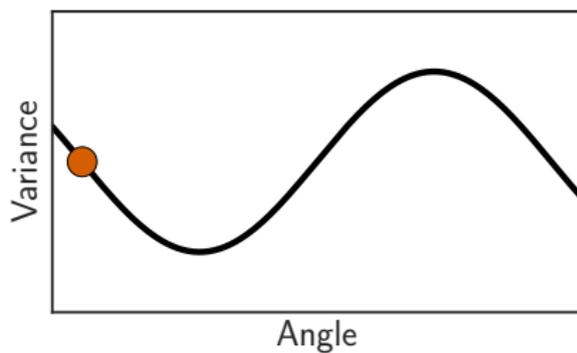
Data and mean



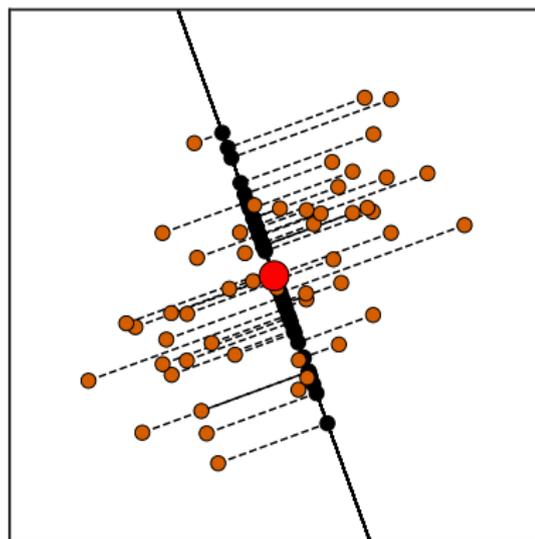
Axe principal : maximisation de la variance



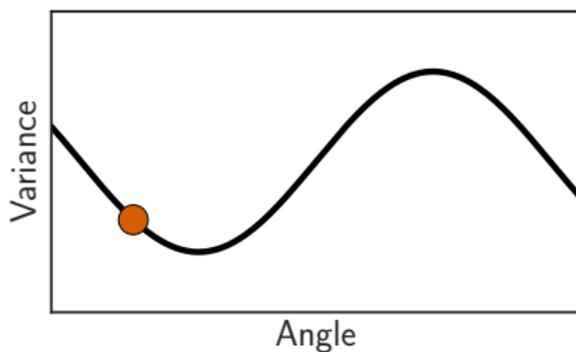
Data, mean and projection



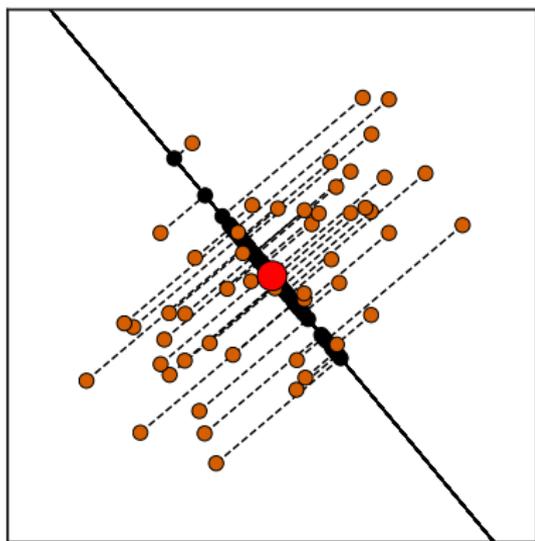
Axe principal : maximisation de la variance



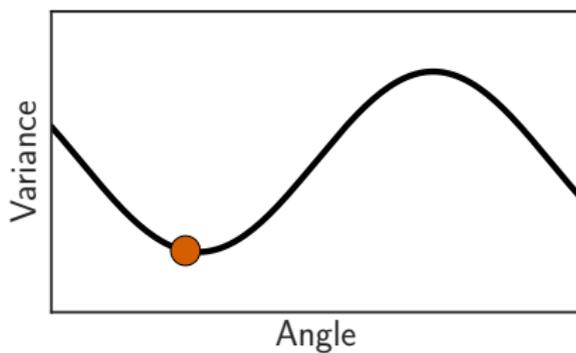
Data, mean and projection



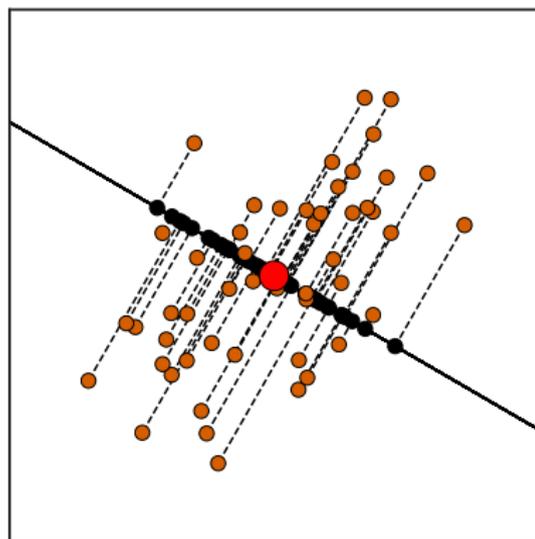
Axe principal : maximisation de la variance



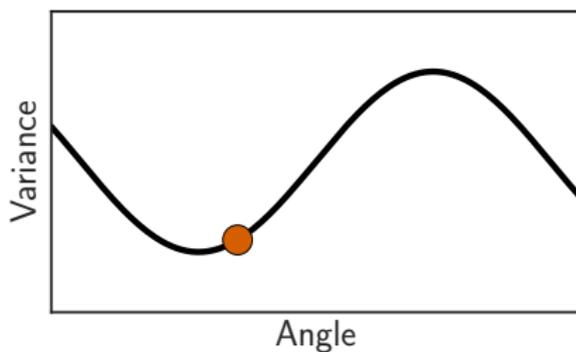
Data, mean and projection



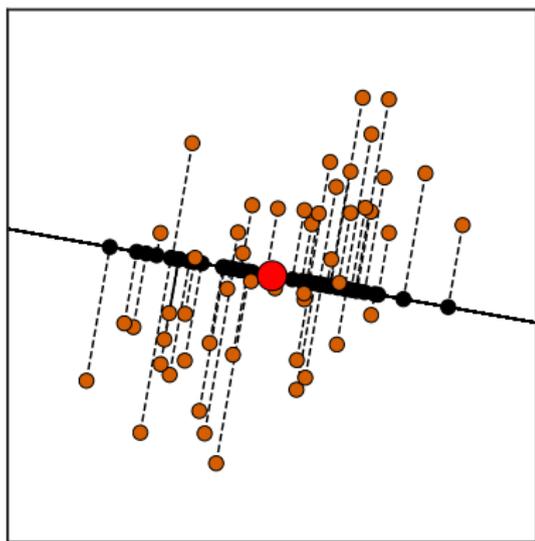
Axe principal : maximisation de la variance



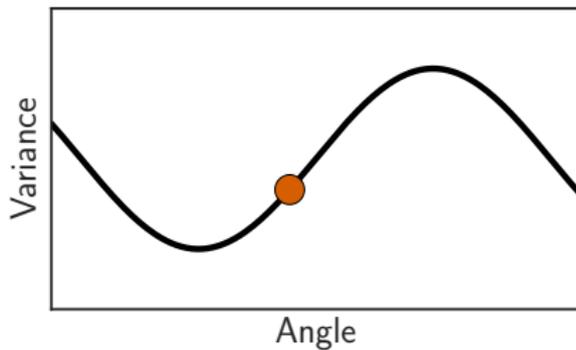
Data, mean and projection



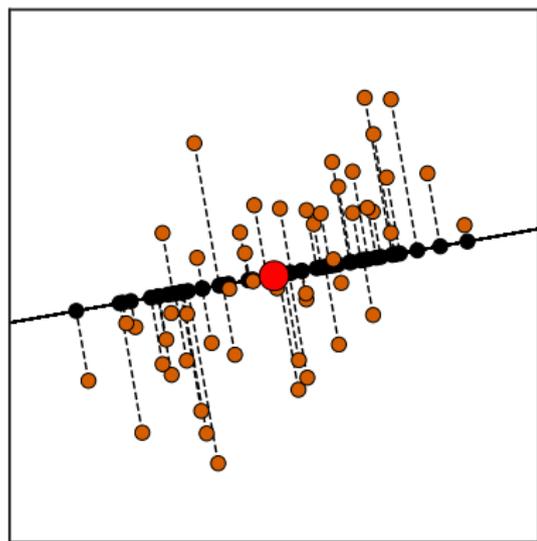
Axe principal : maximisation de la variance



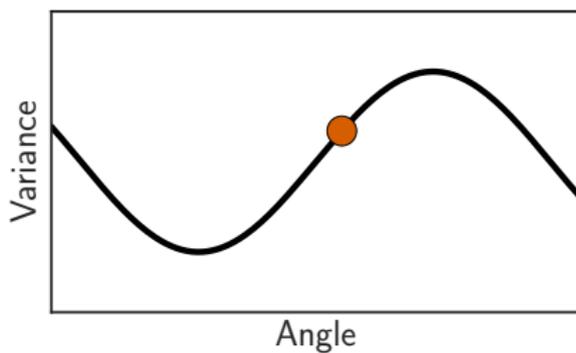
Data, mean and projection



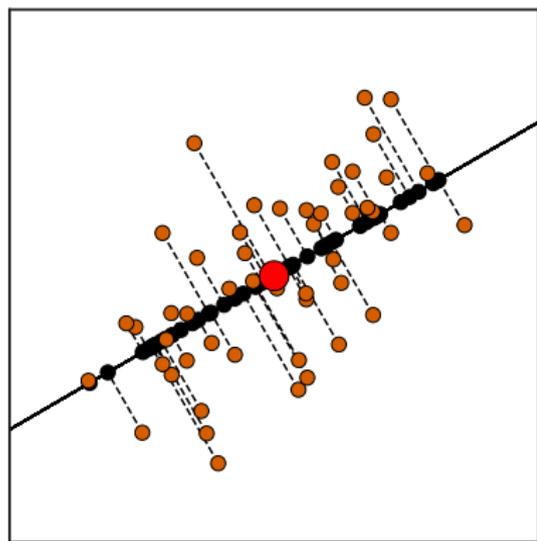
Axe principal : maximisation de la variance



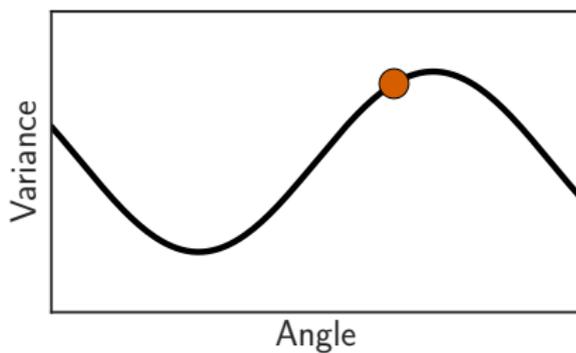
Data, mean and projection



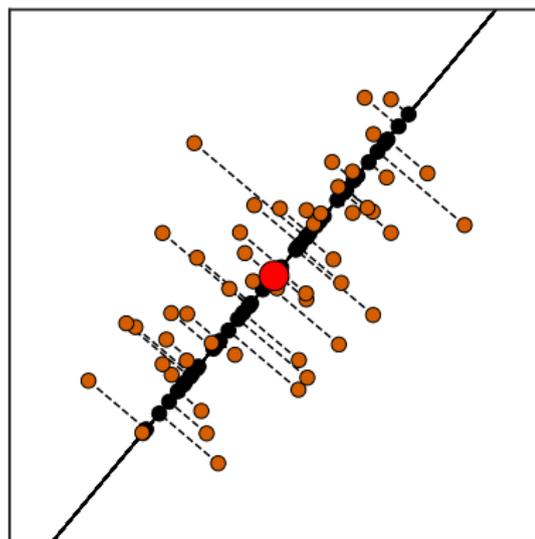
Axe principal : maximisation de la variance



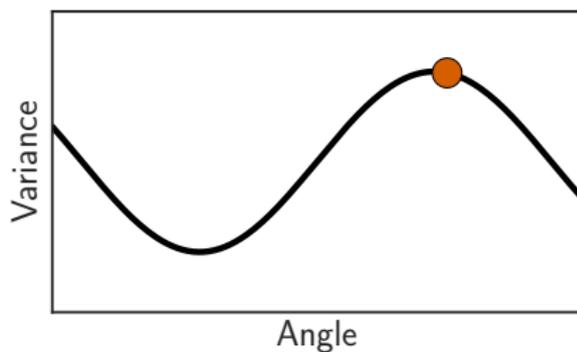
Data, mean and projection



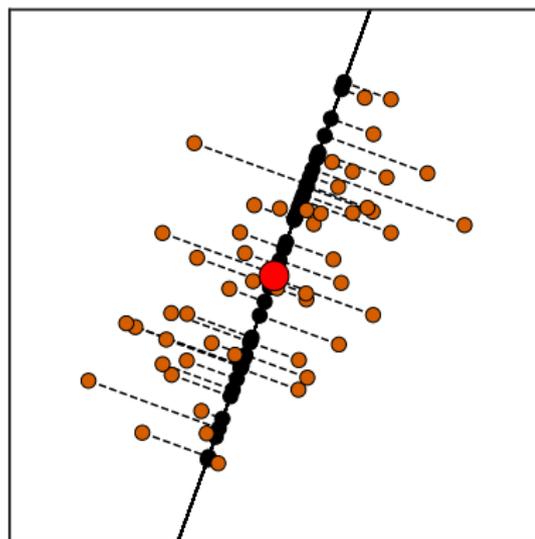
Axe principal : maximisation de la variance



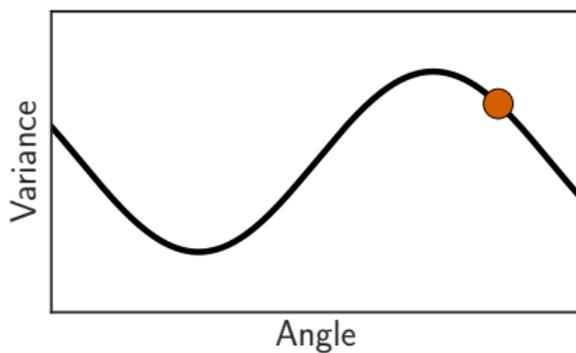
Data, mean and projection



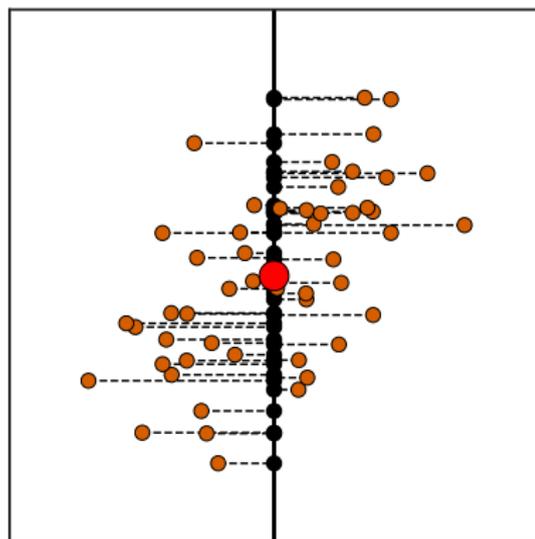
Axe principal : maximisation de la variance



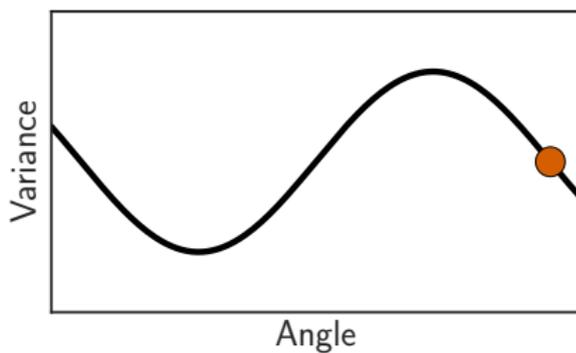
Data, mean and projection



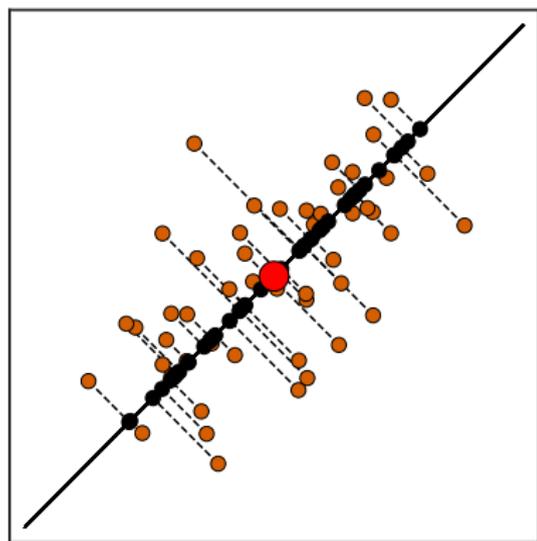
Axe principal : maximisation de la variance



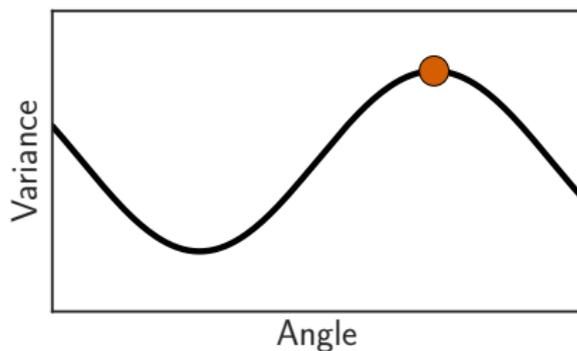
Data, mean and projection



Axe principal : maximisation de la variance



Principal direction (main axis)



ACP : axe principal

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

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

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

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

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

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

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)

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

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

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

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

|

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

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

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

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

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

Rem: on résout une maximisation sous contrainte convexe

ACP : axe principal

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

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

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 $k = 1, \dots, K$ **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 :

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

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

$$\left| \begin{array}{l} \mathbf{w} \leftarrow X \mathbf{v} \\ \mathbf{v} \leftarrow X^\top \mathbf{w} \\ \mathbf{v} \leftarrow \frac{\mathbf{v}}{\|\mathbf{v}\|} \end{array} \right.$$

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/SVD - 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 :

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

- ▶ 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.

Rem: numériquement il y a d'autres alternatives à la déflation

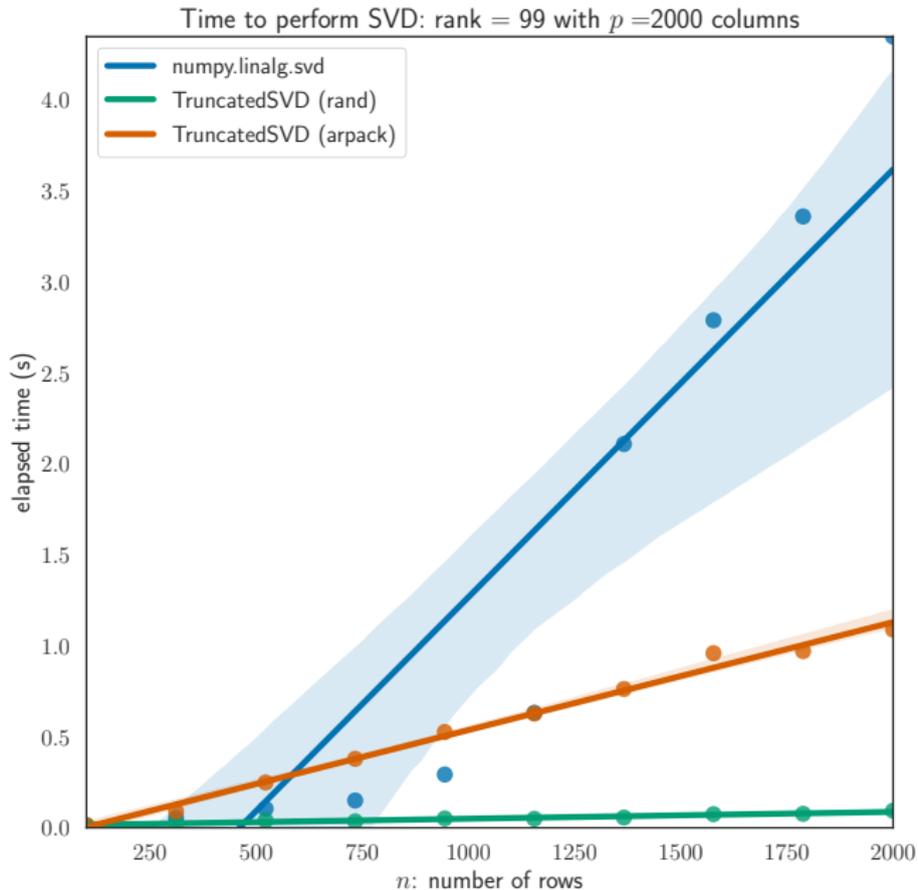
Autres méthodes numériques

- ▶ Algorithme de Lánczos / Espace de Krylov : utile quand plusieurs composantes / valeurs propres sont requises
- ▶ Itérations d'Arnoldi

cf. Golub et VanLoan (2013)

Rem: des techniques récentes ont permis des gains en rapidité en utilisant des méthodes aléatoires (cf. **sketching**), Halko *et al.* (2011)

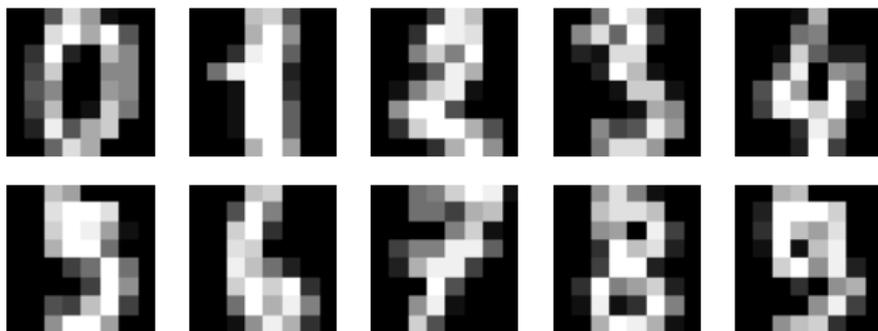
Temps de calcul pour quelques solveurs SVD



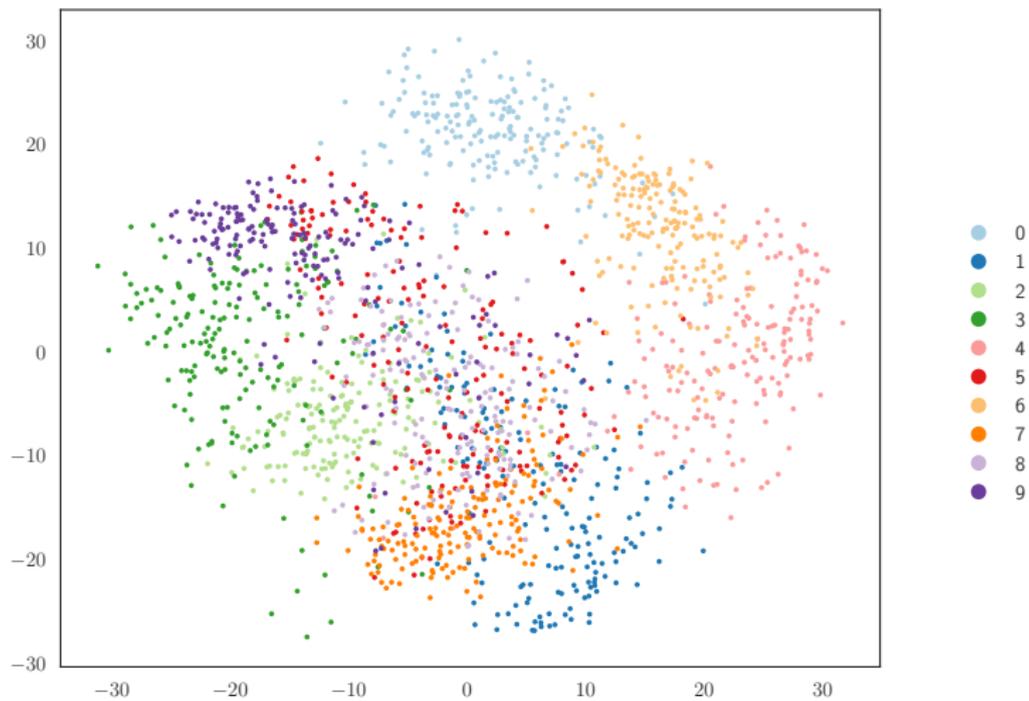
Alternatives à l'ACP

D'autres méthodes de réduction de dimension peuvent exister, e.g., **t-SNE** (t-distributed Stochastic Neighbor Embedding)

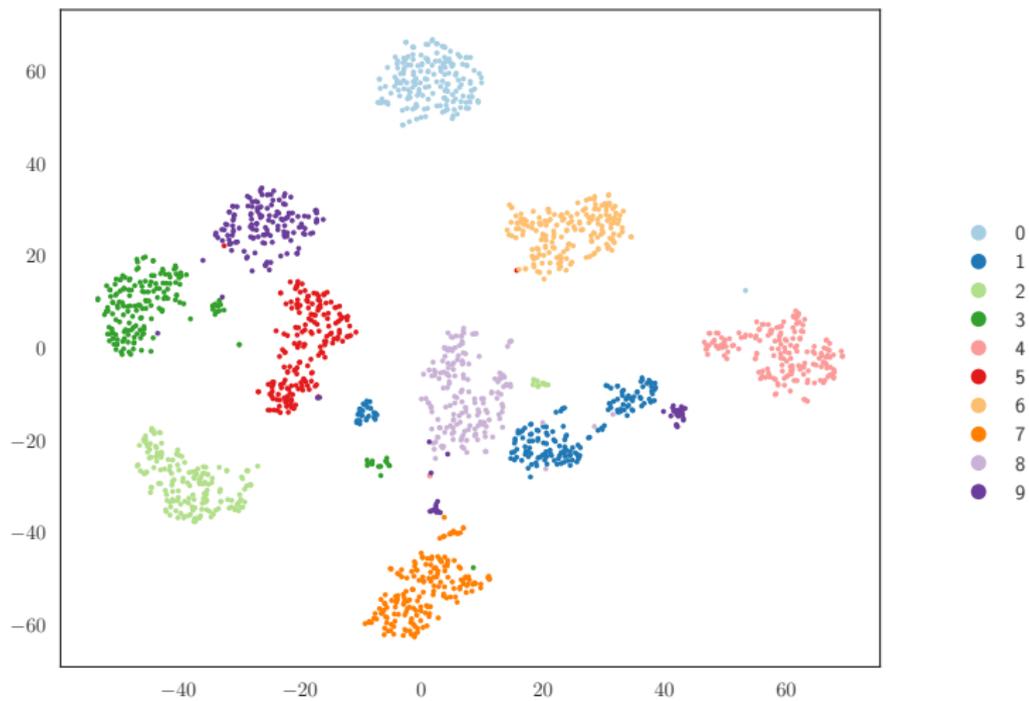
Exemple : dataset "digits" $X \in \mathbb{R}^{n \times p}$ avec $(n, p) = (1797, 64)$
(1797 chiffres numérisés d'image 8×8)



Exemple sur “digits” : PCA (2 axes)



Exemple sur “digits” : t-SNE (2 axes)



Références I

- ▶ G. H. Golub and C. F. van Loan.
Matrix computations.
Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.
- ▶ N. Halko, P. Martinsson, and J. A. Tropp.
Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions.
SIAM Review, 53 :217, 2011.