

Analyse en Composantes Principales

Annie Chateau

Université de Montpellier

*annie.chateau@umontpellier.fr – noura.faraj@umontpellier.fr –
mountaz.hascoet@umontpellier.fr – violaine.prince@umontpellier.fr*

HAI702I Algèbre, géométrie, transformation, calcul numérique

- 1 Matrices symétriques
- 2 Décomposition en valeurs singulières (SVD)
Valeurs singulières et vecteurs singuliers
La SVD réduite
- 3 Meilleure approximation
- 4 Pseudo-inverse
- 5 Analyse en composante principale (ACP ou PCA)
Interprétation et régression

Théorème spectral

Théorème (Théorème spectral)

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

$$S = U \text{Diag}(\lambda_1, \dots, \lambda_n) U^T \text{ ou } SU = U \text{Diag}(\lambda_1, \dots, \lambda_n)$$

On peut donc écrire la matrice S comme la somme pondérée de matrice de projection. On a $U = [u_1, \dots, u_n]$ qui donne :

$$S = \sum_{i=1}^n \lambda_i u_i u_i^T, \quad \text{avec } \forall i \in \{1, n\}, \quad Su_i = \lambda_i u_i.$$

Remarque

- 1 une matrice $U \in \mathbb{R}^{n \times n}$ est dite orthogonale si elle vérifie $U^T U = U U^T = I_n$ ou $\forall (i, j) \in \{1, n\}^2 \quad u_i^T u_j = \langle u_i, u_j \rangle = \delta_{i,j}$.
- 2 les λ_i sont les valeurs propres de S et les $u_i \in \mathbb{R}^n$ sont les vecteurs propres associés.
- 3 les matrices $u_i u_i^T$ sont de rang 1 et correspondent aux projections sur $\text{Vect}(u_i)$.

Décomposition en valeurs singulières (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

$$U^T X V = \text{Diag}(s_1, \dots, s_{\min(n,p)}) = \Sigma \in \mathbb{R}^{n \times p}$$

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

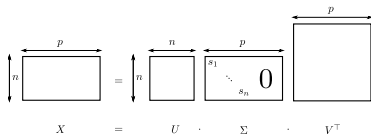
$$X = U \Sigma V^T$$

avec $U = [u_1, \dots, u_n]$ et $V = [v_1, \dots, v_p]$.

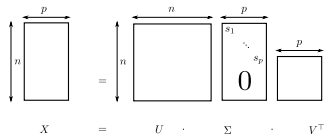
Décomposition en valeurs singulières (SVD)

Les matrices U et V étant orthogonales, on a bien

$$\begin{cases} \langle u_i, u_j \rangle = \delta_{i,j}, & \forall (i,j) \in \{1, n\}^2 \\ \langle v_i, v_j \rangle = \delta_{i,j}, & \forall (i,j) \in \{1, p\}^2 \end{cases}$$



(a) $n \leq p$



(b) $n \geq p$

Figure – Visualisation de la SVD.

Décomposition en valeurs singulières (SVD)

Les nombres réels s_j sont les *valeurs singulières* de X et les u_j (resp. v_j) sont les *vecteurs singuliers* à gauche (resp. droite).

Remarque Attention, les matrices U et V ne sont pas uniques. En particulier, les vecteurs u_j (resp. v_j) sont définis à orientation près (multiplication par -1).

Propriété variationnelle de la plus grande valeur singulière

La plus grande valeur singulière satisfait :

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

Pour résoudre ce problème d'optimisation sous contrainte, on écrit le Lagrangien :

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

qui donne les conditions nécessaire d'optimalité suivante :

Propriété variationnelle de la plus grande valeur singulière

$$\begin{cases} \nabla_u \mathcal{L} = Xv - 2\lambda_1 u = 0 \\ \nabla_v \mathcal{L} = X^\top u - 2\lambda_2 v = 0 \end{cases} \iff \begin{cases} Xv = 2\lambda_1 u \\ X^\top u = 2\lambda_2 v \end{cases} \Rightarrow \begin{cases} X^\top Xv = \alpha v \\ XX^\top u = \alpha u \end{cases}$$

avec $\alpha = 4\lambda_1 \lambda_2$, et donc v est vecteur propre de $X^\top X$ et u est vecteur propre de XX^\top .

La SVD réduite

On part de la SVD $X = U\Sigma V^\top$. On ne garde que les éléments utiles avec $r = \min(n, p)$:

$$X = \sum_{i=1}^r s_i u_i v_i^\top = U_r \text{Diag}(s_1, \dots, s_r) V_r^\top$$

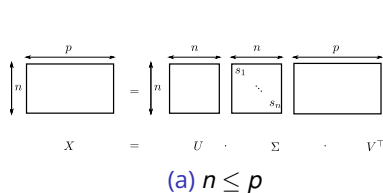
avec $s_i > 0, \forall i \in \{1, r\}$ et $U_r = [u_1, \dots, u_r]$, $V_r = [v_1, \dots, v_r]$. Quand on en garde que les $r = \text{rank}(X)$ valeurs singulières non-nulles, on parle alors de *SVD compacte*.

Remarque

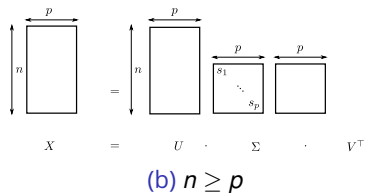
- 1 Les matrices $u_i v_i^\top$ sont toutes de rang 1.
- 2 Les u_i (resp. les v_i^\top) sont orthonormés et engendrent le même espace que celui engendré par les colonnes (resp. les lignes) de X :

$$\text{Vect}(x_1, \dots, x_p) = \text{Vect}(u_1, \dots, u_r)$$

La SVD réduite


$$X = U \cdot \Sigma \cdot V^T$$

(a) $n \leq p$


$$X = U \cdot \Sigma \cdot V^T$$

(b) $n \geq p$

Figure – Visualisation de la SVD réduite.

Meilleure approximation

Etant donnée une matrice X que l'on suppose de rang $r \in \mathbb{N}^*$, on se pose la question de trouver une matrice de rang $1 \leq k \leq r$ qui approxime le mieux X . On a donc besoin d'un critère de proximité (en fait, une norme). On sait résoudre le problème pour la norme spectrale :

Définition

La norme spectrale de $X \in \mathbb{R}^{n \times p}$ est définie par

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

C'est donc le module de la plus grande valeur propre.

Meilleure approximation

Dans ce cas, la réponse au problème de meilleure approximation est donnée par la SVD tronquée :

Théorème (Meilleure approximation de rang k)

Soit $X = \sum_{i=1}^r s_i u_i v_i^\top$ la SVD compacte de $X \in \mathbb{R}^{n \times p}$. On note

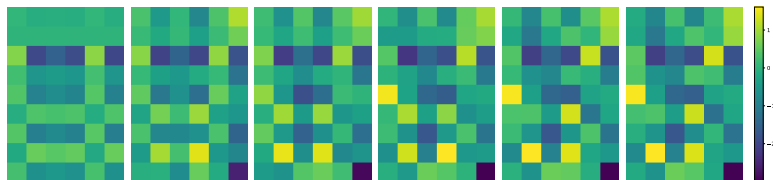
$$X_k = \sum_{i=1}^k s_i u_i v_i^\top, \quad \text{pour tout } k \in \{1, r\}.$$

On a alors,

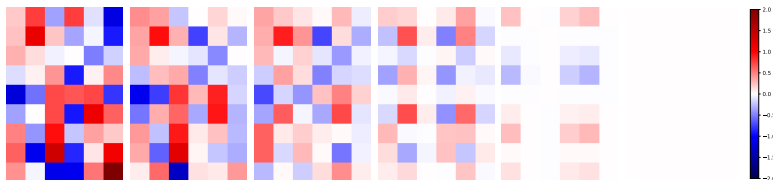
$$\min_{Z \in \mathbb{R}^{n \times p} : \text{rank}(Z)=k} \|X - Z\|_2 = \|X - X_k\|_2 = s_{k+1}$$

Cette propriété est cruciale pour l'analyse en composante principale (ACP).

Meilleure approximation



(a) $k = 1$ (b) $k = 2$ (c) $k = 3$ (d) $k = 4$ (e) $k = 5$ (f) A



(g) $k = 1$ (h) $k = 2$ (i) $k = 3$ (j) $k = 4$ (k) $k = 5$ (l) A

Figure – Visualisation de l'approximation de rang k . Première ligne : les matrices A_k . Deuxième ligne : la différence $A_k - A$.

Définition

Si $X \in \mathbb{R}^{n \times p}$ admet pour SVD $X = \sum_{i=1}^r s_i u_i v_i^\top$ avec $r = \text{rank}(X)$, alors sa pseudo-inverse $X^+ \in \mathbb{R}^{p \times n}$ est définie par :

$$X^+ = \sum_{i=1}^r \frac{1}{s_i} v_i u_i^\top$$

Si $X = \sum_{i=1}^n s_i u_i v_i^\top \in \mathbb{R}^{n \times n}$ est inversible alors $X^+ = X^{-1}$

Analyse en composante principale (ACP ou PCA)

On observe n points x_1, \dots, x_n dans \mathbb{R}^p . On peut donc créer une matrice

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times p},$$

de n observations (lignes), p features (colonnes).

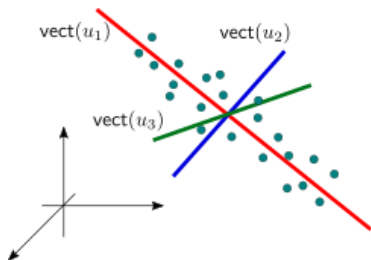


Figure – Le but de l'ACP est de trouver les directions de plus grande dispersion d'un nuage de points (ici $p = 3$)

Définition

- ① *La moyenne (ou le barycentre, ou encore le centre de gravité) d'un nuage de points X est*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \in \mathbb{R}^p.$$

- ② *La variance (ou l'inertie) d'un nuage de points est*

$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^n \|x_i - \bar{X}\|^2 = \frac{1}{n} \sum_{i=1}^n \|x_i\|^2 - \left\| \frac{1}{n} \sum_{i=1}^n x_i \right\|^2.$$

- ③ *L'écart-type d'un nuage de points est $\sigma_X = \sqrt{\sigma_X^2}$.*

Analyse en composante principale (ACP ou PCA)

Remarque La variance d'un nuage de point est nulle si et seulement si $x_i = x_j$, pour tout $i, j = 1, \dots, n$.

On considère ici des nuages de points centrés. À l'aide d'une translation, on peut toujours recentrer le nuage de points pour qu'il ait une moyenne nulle :

$$X \leftarrow \begin{pmatrix} x_1^\top - \bar{x}^\top \\ \vdots \\ x_n^\top - \bar{x}^\top \end{pmatrix} = X - \mathbb{1}_n \bar{x}^\top.$$

où l'on a noté $\mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ et le vecteur colonne $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \in \mathbb{R}^p$

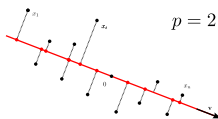
contenant la moyenne de chaque coordonnée des x_i , $i = 1, \dots, n$.
On peut aussi mettre à l'échelle pour avoir un écart-type similaire par colonne (feature).

Analyse en composante principale (ACP ou PCA)

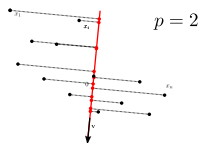
Soit le paramètre $1 \leq k \leq p$. On cherche à projeter le nuage de points X sur un sous espace vectoriel $E_k \subset \mathbb{R}^p$ de dimension k . Cette méthode *compresse* donc le nuage de points de dimension p en un nuage de dimension k . L'idée est de choisir le sev E_k de manière à ce que le nuage projeté, noté X_k ressemble le plus à X : la compression doit faire perdre le moins possible d'information. On utilise un critère basé sur l'inertie.



(a) Données brutes



(b) Données projetées sur l'axe optimal



(c) Données projetées sur un autre axe

Figure – Projection d'un nuage de points de dimension $p = 2$ sur un sev de dimension $k = 1$.

Analyse en composante principale (ACP ou PCA)

Definition

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

$$X = \sum_{i=1}^r s_i u_i v_i^{\top} \longrightarrow \sum_{i=1}^k s_i u_i v_i^{\top}.$$

Les sous espaces $E_k = \text{Vect}(v_1, \dots, v_k)$ sont appelés *sous espaces principaux*.

Analyse en composante principale (ACP ou PCA)

Cela nous donne une nouvelle manière de représenter les données. On a projeté le nuage de point original sur un sous-espace vectoriel. Un peu de vocabulaire :

- Les axes (de direction) $v_1, \dots, v_p \in \mathbb{R}^p$ sont appelés *axes principaux* ou *axes factoriels*, les nouvelles variables

$$c_j = Xv_j, j = 1, \dots, p$$

sont appelées *composantes principales* (coordonnées des x_i projetés sur l'axe $\text{Vect}(v_j)$).

- La matrice $XV_k \in \mathbb{R}^{n \times k}$ (avec $V_k = [v_1, \dots, v_k] \in \mathbb{R}^{p \times k}$) est la matrice représentant les coordonnées des données dans la base des k premiers axes principaux (nouvelle représentation d'ordre k).

Analyse en composante principale (ACP ou PCA)

Souvent, on cherche à reconstruire les données dans l'espace original \mathbb{R}^p (e.g. pour débruiter) :

- Reconstruction "parfaite" pour $x \in \mathbb{R}^p$:

$$x = \sum_{j=1}^p \langle x, v_j \rangle v_j = \sum_{j=1}^p (x^\top v_j) v_j$$

qui donne $X = XVV^\top$.

- Reconstruction avec perte d'information pour $x \in \mathbb{R}^p$:

$$\hat{x} = \sum_{j=1}^k \langle x, v_j \rangle v_j = \sum_{j=1}^k (x^\top v_j) v_j \in \mathbb{R}^p$$

qui donne le nouveau nuage de points $\hat{X}_k = XV_k V_k^\top$ projeté de X sur E_k dans \mathbb{R}^p .

Analyse en composante principale (ACP ou PCA)

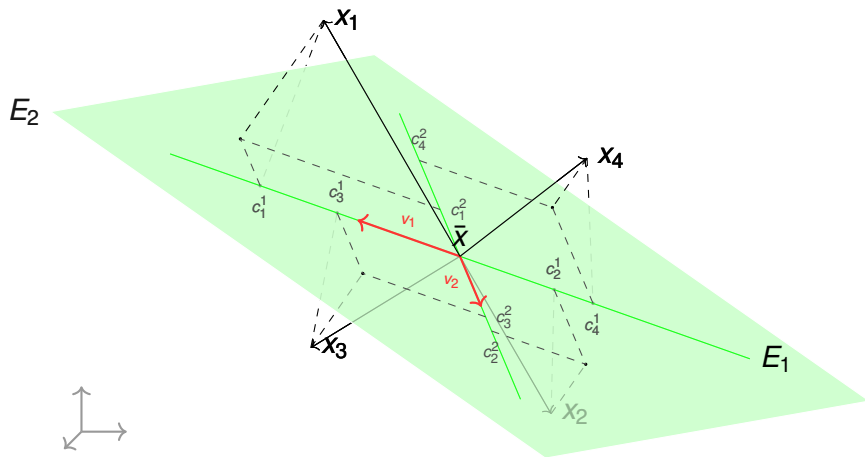


Figure – Espaces factoriels et composantes principales ($n = 4$ et $p = 3$)

Interprétation et récursion

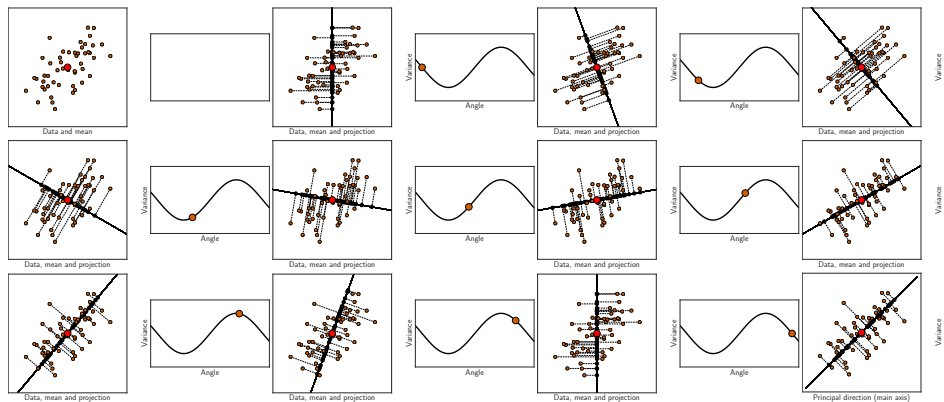


Figure – Voir aussi la vidéo :

<https://twitter.com/salmonjsph/status/1177564363261194240>

Axe principal : maximisation de la variance projetée

L'axe principal (normalisé) v_1 est la solution du problème :

$$v_1 \in \underset{v \in \mathbb{R}^p, \|v\|=1}{\operatorname{Argmax}} v^\top X^\top X v = \underset{v \in \mathbb{R}^p, \|v\|=1}{\operatorname{Argmax}} \|Xv\|^2 = \underset{v \in \mathbb{R}^p, \|v\|=1}{\operatorname{Argmax}} \sum_{i=1}^n (x_i^\top v)^2$$

Après recentrage le dernier terme est la variance du nuage de points projeté sur l'axe v .

Axe principal : maximisation de la variance projetée

On résout donc un problème d'optimisation (une maximisation) sous contrainte convexe. On se ramène à un problème de maximisation global en considérant le Lagrangien. Cela revient à maximiser la fonction objectif suivante en v :

$$\mathcal{L}(v, \lambda) = (Xv)^\top (Xv) - \lambda(v^\top v - 1) = v^\top X^\top Xv - \lambda(v^\top v - 1)$$

où λ est le multiplicateur de Lagrange.

Les conditions d'optimalité du premier ordre en un extremum sont

$$\frac{\partial \mathcal{L}(v_1, \lambda)}{\partial v} = 0 \Leftrightarrow X^\top Xv_1 = \lambda v_1$$

La matrice de Gram $X^\top X$ est diagonalisable (symétrique) donc si v_1 est un extremum alors c'est un vecteur propre.

On normalise v_1 pour que $\|v_1\| = 1$, ainsi $\lambda = v_1^\top X^\top Xv_1$ et v_1 est un vecteur propre, de valeur propre λ maximale.

Axe principal : maximisation de la variance projetée

Proposition

Dans le cas où le nuage de points est centrée, on appelle matrice de covariance (ou matrice de Gram) de $X \in \mathbb{R}^{n \times p}$ la matrice $\frac{1}{n}X^\top X \in \mathbb{R}^{p \times p}$. On a alors

$$\operatorname{tr}\left(\frac{1}{n}X^\top X\right) = \sigma_X^2 = \underbrace{\sigma_{\pi_{\operatorname{Vect}(u_1)}(X)}^2}_{=\frac{s_1}{n}} + \cdots + \underbrace{\sigma_{\pi_{\operatorname{Vect}(u_p)}(X)}^2}_{=\frac{s_p}{n}},$$

où π_E est la projection orthogonale sur le sev E .

Qualité de la représentation

La part d'inertie « expliquée » par le sous espace E_k est alors

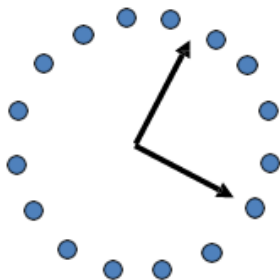
$$\frac{s_1 + \cdots + s_k}{s_1 + \cdots + s_p}.$$

Si $k = p$ la totalité de l'inertie est retrouvée. Le choix du nombre de dimension k se fait sur des critères empiriques. Cela dépend aussi de la finalité de l'ACP : pour de la description de données on a souvent $q \leq 3$ (au delà difficultés d'interprétation), pour de la compression k peut être beaucoup plus grand.

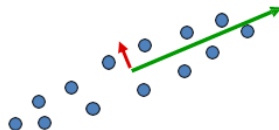
Exemple de méthodes pour l'analyse descriptive :

- Méthode de Kaiser : Ne retenir que les directions principales qui expliquent une proportion de l'inertie supérieure à $\frac{1}{p}$.
- Méthode du coude : sur "l'éboulis" des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière : sélectionner les axes avant le décrochement.
- Méthode pratique : ne retenir que les dimensions que l'on peut interpréter...

Qualité de la représentation



(a) Les valeurs singulières sont égales. Pas de direction privilégiée. Tout vecteur est vecteur principal.



(b) Les valeurs singulières sont très déséquilibrées. Il y a une direction privilégiée.

Exemple

Soit $X = \begin{pmatrix} -1 & -2 & 1 \\ 1 & 0 & 2 \\ 2 & -1 & 3 \\ 2 & -1 & 2 \end{pmatrix}$. On a donc $n = 4$ et $p = 3$. Le barycentre

des données est $\bar{X} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ et la matrice de covariance est

$$G = \frac{1}{4} \begin{pmatrix} 6 & 2 & 3 \\ 2 & 2 & 1 \\ 3 & 1 & 2 \end{pmatrix}.$$

Le polynôme caractéristique de G est

$$p_G(\lambda) = -\lambda^3 + \frac{5}{2}\lambda^2 - \frac{7}{8}\lambda + \frac{1}{16}.$$

Exemple

On trouve numériquement

$$s_1 \simeq 2.097$$

$$s_2 \simeq 0.3055$$

$$s_3 \simeq 0.09756$$

On trouve alors les vecteurs principaux $v_1 = \begin{pmatrix} -0.833 \\ -0.330 \\ -0.443 \end{pmatrix}$,

$$v_2 = \begin{pmatrix} -0.257 \\ 0.941 \\ -0.218 \end{pmatrix}, v_3 = \begin{pmatrix} -0.489 \\ -0.0675 \\ -0.870 \end{pmatrix}.$$

La part d'inertie expliquée par le premier axe principal est 0.839 et celle expliquée par le premier plan principal est 0.96.