Analyse en Composantes Principales

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

March 20, 2023

Résumer l'information ?

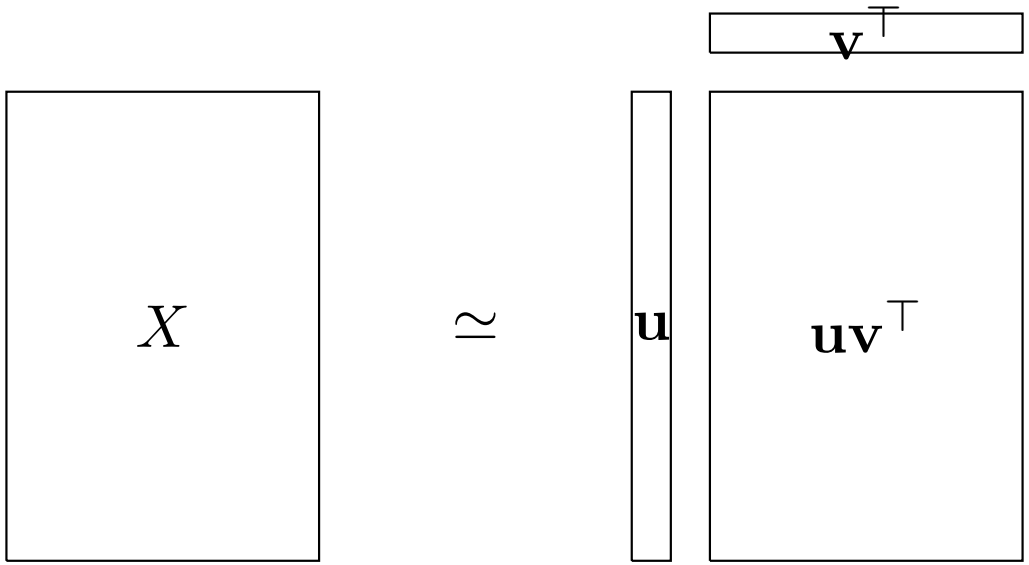
- Calculer une représentation avec moins de données mais un maximum d'informations
- Compression
- Suppression du bruit
- Visualisation en 2D ou 3D

Résumer un tableau de données

Comment résumer l'information ?

- Résumer un tableau de données par deux vecteurs u et v
- $X \in \mathbb{R}^{n \times p}$

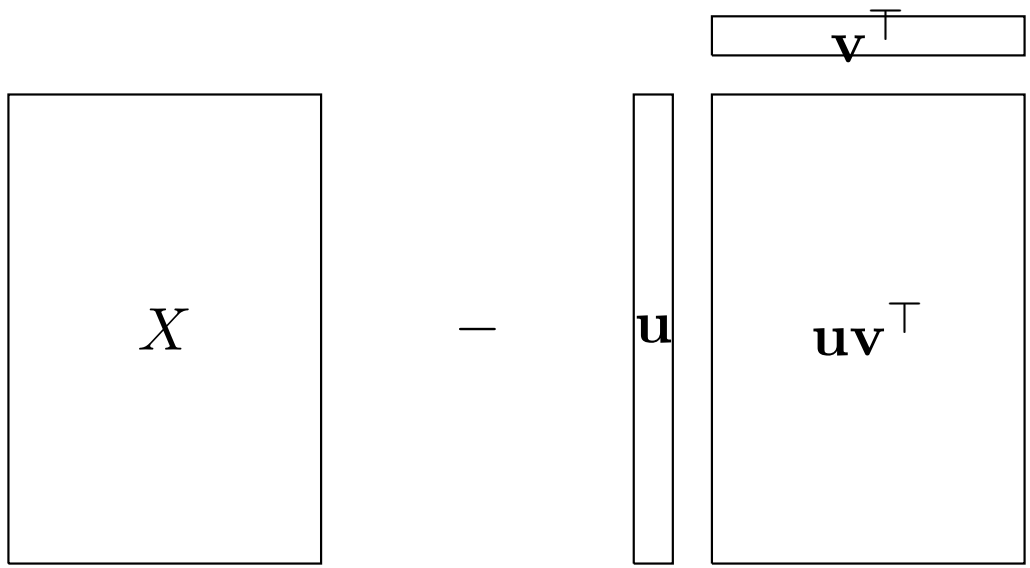
La meilleur représentation linéaire du nuage de points est donnée par le couple de vecteurs $\mathbf{u} \in \mathbb{R}^n$ et $\mathbf{v} \in \mathbb{R}^p$ permettant au mieux de reconstruire la matrice X.



Reconstruction de X

$$\min_{\mathbf{u},\mathbf{v}} \; J(\mathbf{u},\mathbf{v}) \quad \text{avec} \quad J(\mathbf{u},\mathbf{v}) \; = \sum_{i}^{n}\sum_{j}^{p}(x_{ij}-u_i v_j)^2$$

Aussi noté : $J(\mathbf{u},\mathbf{v}) \; = \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$



Fonction de coût

La fonction coût $\sum_i^n \sum_j^p (x_{ij} - u_i v_j)^2$ peut se réécrire :

$$\begin{aligned} J(\mathbf{u}, \mathbf{v}) &= \sum_i^n \sum_j^p x_{ij}^2 - 2 \sum_i^n \sum_j^p x_{ij} u_i v_j + \sum_i^n \sum_j^p (u_i v_j)^2 \\ &= \sum_i^n \sum_j^p x_{ij}^2 - 2 \sum_i^n \left(\sum_j^p x_{ij} v_j \right) u_i + \sum_i^n u_i^2 \sum_j^p v_j^2 \\ &= \sum_i^n \sum_j^p x_{ij}^2 - 2 (X\mathbf{v})^\top \mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \end{aligned}$$

et donc

$$\min_{\mathbf{u},\mathbf{v}} \underbrace{\|X - \mathbf{u}\mathbf{v}^\top\|_F^2}_{J(\mathbf{u},\mathbf{v})} \Leftrightarrow \min_{\mathbf{u},\mathbf{v}} \underbrace{-2(X\mathbf{v})^\top \mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v}\|^2}_{\mathcal{J}(\mathbf{u},\mathbf{v})}$$

Comment résoudre

$$\min_{\mathbf{u},\mathbf{v}} \mathcal{J}(\mathbf{u},\mathbf{v})?$$

Comment résoudre $\min_{\mathbf{u},\mathbf{v}} \mathcal{J}(\mathbf{u},\mathbf{v})$?

$\Rightarrow$ La méthode du gradient

Méthode du gradient

Minimisation d'une fonction de plusieurs variables

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \;=\; \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$$

Définition : Gradient

soit F une fonction de plusieurs (d) variables :

$$\begin{array}{rccc} F: & \mathbb{R}^d & \longmapsto & \mathbb{R} \\ & \mathbf{x} & \longrightarrow & F(\mathbf{x}) \end{array}$$

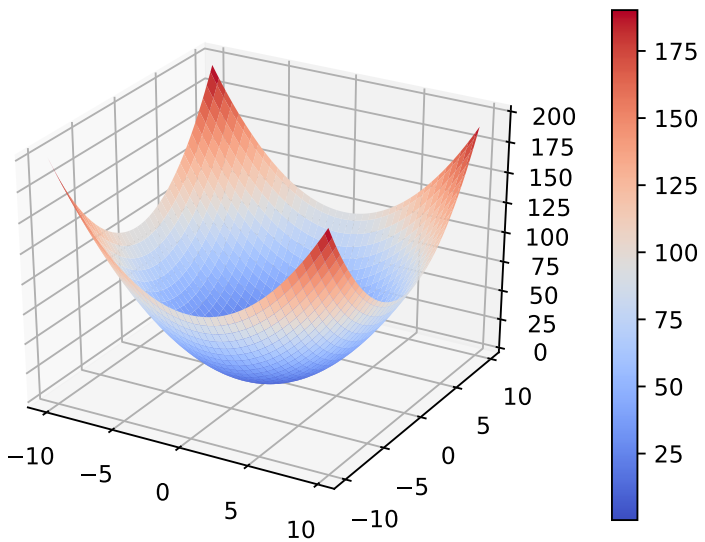
On appelle gradient de F au point $\mathbf{x}$ la fonction des dérivées partielles

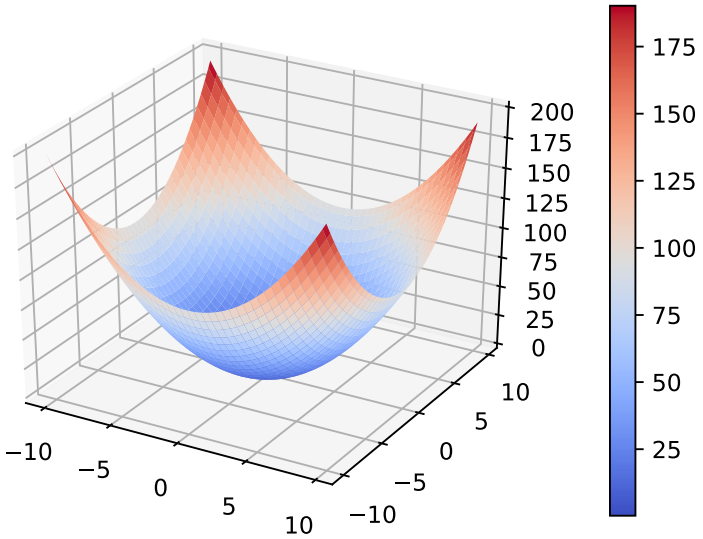
$$\begin{array}{rccc} \nabla_{\mathbf{x}} F: & \mathbb{R}^d & \longmapsto & \mathbb{R}^d \\ & \mathbf{x} & \longrightarrow & \nabla_{\mathbf{x}} F(\mathbf{x}) = \left(\frac{\partial F}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial F}{\partial x_i}(\mathbf{x}), \ldots, \frac{\partial F}{\partial x_d}(\mathbf{x})\right)^\top \end{array}$$

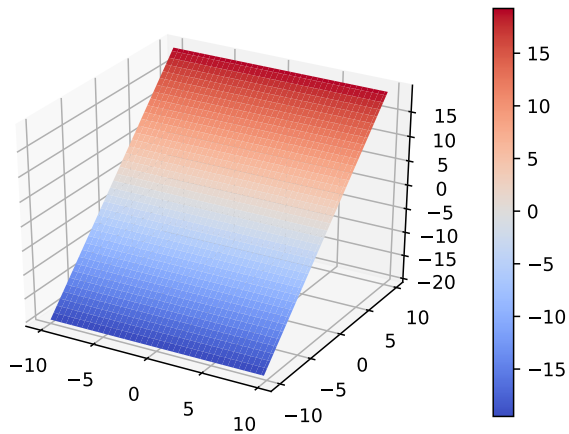
Minimisation d'une fonction de plusieurs variables

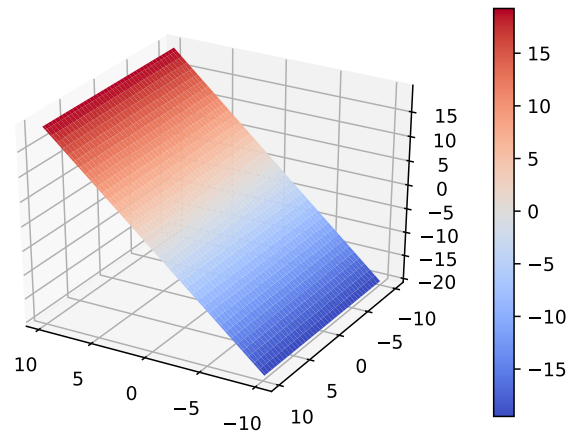
Condition d'optimalité

- $(u^\star, v^\star)$ est solution du problème de minimisation ssi le gradient de la fonction J s'annule en ce point
- $$\begin{cases} \nabla_{\mathbf{u}} J(u^\star, v^\star) = 0 \\ \nabla_{\mathbf{v}} J(u^\star, v^\star) = 0 \end{cases}$$
- F doit être convexe et différentiable
- Si F est strictement convexe : solution unique




200
175
150
125
100
75
50
25
0
−10
−5
0
5
10
175
150
125
100
75
50
25


15
10
5
0
−5
−10
−15
−20
−10
−5
0
5
10
15
10
5
0
−5
−10
−15


15
10
5
0
−5
−10
−15
−20
10
5
0
−5
−10
15
10
5
0
−5
−10
−15

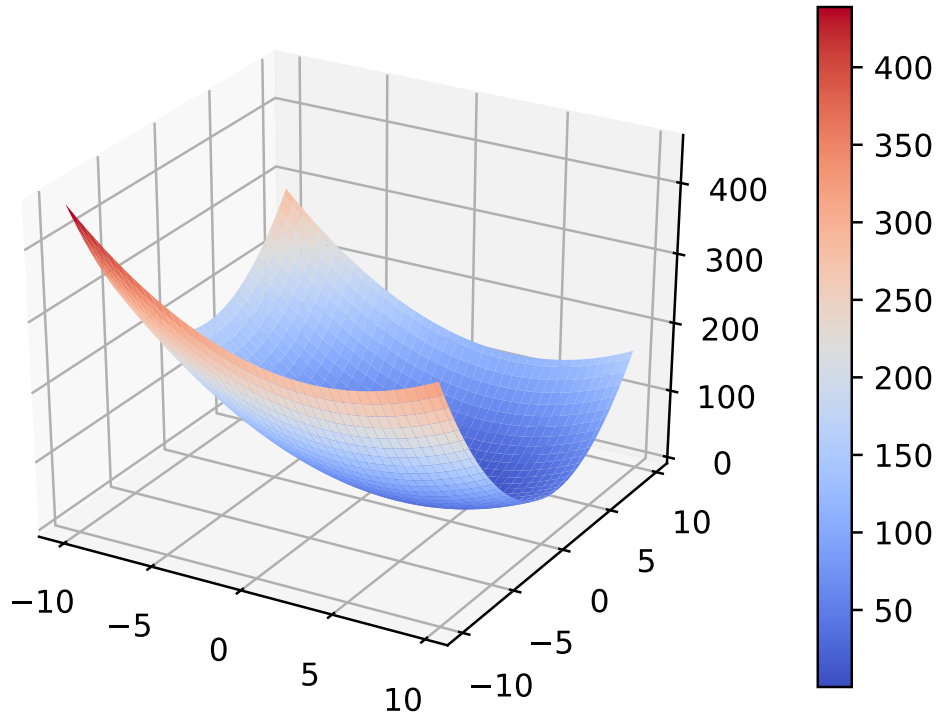
Minimisation d'une fonction de plusieurs variables

Exemple

$$\min_{x,y} J(x,y) \ = 2(x-a)^2 + (y-b)^2$$

Méthode de résolution du problème

1. Calcul du gradient $\nabla_{\mathbf{x}} F(x,y)$ et $\nabla_{\mathbf{y}} F(x,y)$
   - $\nabla_x = 4(x-a)$
   - $\nabla_y = 2(y-b)$
2. Résolution des équations $\nabla_{\mathbf{x}} F(x^\star, y^\star) = 0$ (2 équations à 2 inconnues)
   - $\nabla_x = 0 \Leftrightarrow x^\star = a$
   - $\nabla_y = 0 \Leftrightarrow y^\star = b$


400
300
200
100
0
−10
−5
0
5
10
−10
−5
0
5
10
400
350
300
250
200
150
100
50

Introduction à l'ACP

Reconstruction de X

Rappel du problème

$$\min_{\mathbf{u},\mathbf{v}} \underbrace{\|X - \mathbf{u}\mathbf{v}^\top\|_F^2}_{J(\mathbf{u},\mathbf{v})} \Leftrightarrow \min_{\mathbf{u},\mathbf{v}} \underbrace{-2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2}_{\mathcal{J}(\mathbf{u},\mathbf{v})}$$

Résolution

Minimiser le cout c'est trouver $\mathbf{u}$ et $\mathbf{v}$ qui annulent le gradient :

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) = \|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \Leftrightarrow \begin{cases} \nabla_{\mathbf{u}}(\mathbf{u}) = 0 \\ \nabla_{\mathbf{v}}(\mathbf{v}) = 0 \end{cases}$$

Calcul du gradient

$$J(\mathbf{u},\mathbf{v}) = \sum_{i}^{n}\sum_{j}^{p} x_{ij}^2 - 2\sum_{i}^{n}\left(\sum_{j}^{p} x_{ij}v_j\right)u_i + \sum_{i}^{n} u_i^2 \sum_{j}^{p} v_j^2$$

Gradient

$$\begin{cases} \dfrac{\partial J(\mathbf{u},\mathbf{v})}{\partial u_i} = -2\displaystyle\sum_{j}^{p} x_{ij}v_j + 2u_i \sum_{j}^{p} v_j^2 \\ \dfrac{\partial J(\mathbf{u},\mathbf{v})}{\partial v_j} = -2\displaystyle\sum_{i}^{n} x_{ij}u_i + 2v_j \sum_{i}^{n} u_i^2 \end{cases}$$

Écriture matricielle du gradient[1]

$$\begin{cases} \nabla_{\mathbf{u}} J(\mathbf{u}) = -2X\mathbf{v} + 2\|\mathbf{v}\|^2\mathbf{u} \\ \nabla_{\mathbf{v}} J(\mathbf{v}) = -2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v} \end{cases}$$

[1] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

$\min_{\mathbf{u},\mathbf{v}} \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$

Conditions d'optimalité

$$\begin{cases} \nabla_{\mathbf{u}} \mathcal{J}(\mathbf{u}) = 0 & \Leftrightarrow \quad -X\mathbf{v} + \|\mathbf{v}\|^2 \mathbf{u} = 0 \\ \nabla_{\mathbf{v}} \mathcal{J}(\mathbf{v}) = 0 & \Leftrightarrow \quad -X^\top \mathbf{u} + \|\mathbf{u}\|^2 \mathbf{v} = 0 \end{cases}$$

Solutions

$$\begin{cases} X\mathbf{v} = \|\mathbf{v}\|^2 \mathbf{u} \\ X^\top \mathbf{u} = \|\mathbf{u}\|^2 \mathbf{v} \end{cases} \Rightarrow X^\top X \mathbf{v} = \|\mathbf{v}\|^2 X^\top \mathbf{u} = \underbrace{\|\mathbf{v}\|^2 \|\mathbf{u}\|^2}_{\lambda} \mathbf{v}$$

- Solution $A\mathbf{v} = \lambda \mathbf{v}$
- p solutions $(\mathbf{v}_i, \lambda_i)$.

Quel vecteur propre choisir ?

À l'optimum

$$\mathcal{J}(\mathbf{u},\mathbf{v}) = -2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \qquad \text{et} \qquad X\mathbf{v} = \|\mathbf{v}\|^2\mathbf{u}$$

$$\begin{aligned} \Rightarrow \mathcal{J}(\mathbf{u},\mathbf{v}) &= -2\|\mathbf{v}\|^2\mathbf{u}^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \\ &= -2\|\mathbf{v}\|^2\|\mathbf{u}\|^2 + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \\ &= -\|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \qquad = -\lambda \end{aligned}$$

$$\Rightarrow J(\mathbf{u},\mathbf{v}) = \|X\|_F^2 - \lambda$$

Solution

Couple vecteur propre/valeur propre associé à la plus grande valeur propre de $X^\top X$:

$$\nabla_J(\mathbf{v}) = 0 \quad \Leftrightarrow \quad X^\top X\mathbf{v} - \lambda\mathbf{v} = 0 \quad \text{avec} \quad J(\mathbf{u},\mathbf{v}) = \|X\|^2 - \lambda$$

NB: $X^\top X$ est la matrice de covariance des données centrées à $\frac{1}{n}$ près. Toutes les v.p. de $X^\top X$ (sdp.) sont positives.

Résumé d'un tableau de données : résultat principal

Théorème : (Eckart & Young, 1936)

La solution unique du problème d'optimisation

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \qquad \text{avec} \quad J(\mathbf{u},\mathbf{v}) = \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$$

avec $\|\mathbf{v}^\star\| = 1$, est donnée par

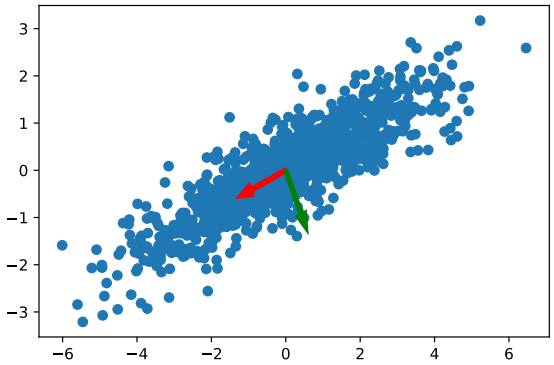
$$\mathbf{v}^\star \qquad \text{et} \qquad \mathbf{u}^\star = X\,\mathbf{v}^\star,$$

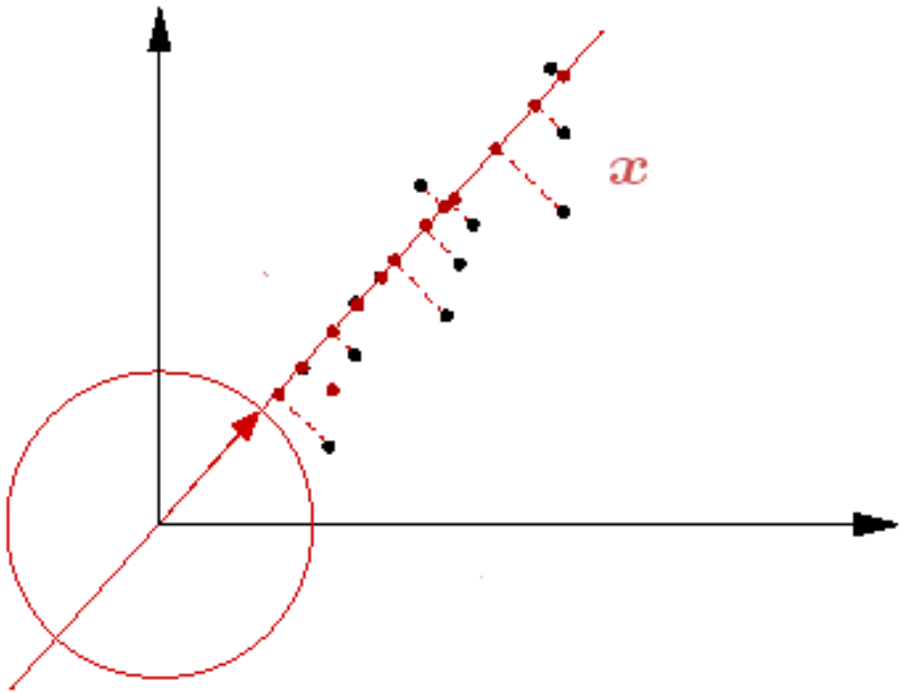
où $\mathbf{v}^\star$ est le vecteur propre normé associé à λ la plus grande valeur propre de la matrice $X^\top X$. De plus on a $\|\mathbf{u}^\star\| = \sqrt{\lambda}$.

En résumé

$\min_{\mathbf{u},\mathbf{v}} \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$

- $\mathbf{u} = X\mathbf{v}$: résumé de X en 1D
- $\mathbf{v}$: vecteur propre de λ_1 de $X^\top X$
- $\mathbf{v}$: Matrice de projection $\mathbb{R}^{n\times p} \to \mathbb{R}^{n\times 1}$
- $\mathbf{v}^\top$: Matrice de projection $\mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times p}$





Analyse de la signification de λ

Quelle est la variance de la projection $\mathbf{u}$?

- Considérons X centrée
- $\mathbf{u} = X\mathbf{v}$, $\bar{\mathbf{u}} = 0$
- Variance de $\mathbf{u}$: $s^2_{\mathbf{u}} = \|\mathbf{u}\|^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_i^2$
- $s^2_{\mathbf{u}} = \frac{1}{n}(X\mathbf{v})^\top X\mathbf{v} = \frac{1}{n}\mathbf{v}^\top X^\top X\mathbf{v}$

$$\Rightarrow s^2_{\mathbf{u}} = \frac{1}{n}\lambda, \text{ avec } X^\top X\mathbf{v} = \lambda\mathbf{v}$$

Meilleure projection ?

- $\mathbf{v}$ est le premier axe factoriel de X
- $X\mathbf{v}$ est la projection linéaire dont la variance est maximale
- Variance $\Leftrightarrow$ Information

Et ensuite ?

Augmenter la quantité d'infos

- La matrice de projection peut être étendue à plusieurs dimensions
- Reconstructions de plus en plus précise

Comment faire ?

Les Résidus I

Matrice des résidus

Construisons la matrice des résidus $R = X - \mathbf{u}\mathbf{v}^\top$

- Contient toute l'info "inexpliquée" par $\mathbf{u}$

Spectre de R

- $R^\top R$ admet les mêmes valeurs/vecteurs propres que $X^\top X$

$$\begin{aligned} R^\top R\mathbf{v}_i &= (X - \mathbf{u}\mathbf{v}^\top)^\top (X - \mathbf{u}\mathbf{v}^\top)\mathbf{v}_i \\ &= X^\top X\mathbf{v}_i - 2X^\top \mathbf{u}\mathbf{v}^\top \mathbf{v}_i + \mathbf{v}\mathbf{u}^\top \mathbf{u}\mathbf{v}^\top \mathbf{v}_i \\ &= \lambda_i \mathbf{v}_i \end{aligned}$$

NB: $\mathbf{v}^\top \mathbf{v}_i = 0$ puisque les vecteurs propres sont orthogonaux entre eux.

Les Résidus II

Spectre de R

- $R^\top R$ admet les mêmes valeurs/vecteurs propres que $X^\top X$
- . . . Sauf la plus grande λ
- $X^\top X\mathbf{v} = \lambda\mathbf{v}, X^\top X\mathbf{v}_2 = \lambda_2\mathbf{v}_2$ et $\lambda_2 < \lambda$
- On a :

$$\begin{aligned} R^\top R\mathbf{v} &= (X - \mathbf{u}\mathbf{v}^\top)^\top(X - \mathbf{u}\mathbf{v}^\top)\mathbf{v} \\ &= X^\top X\mathbf{v} - 2\mathbf{v}\mathbf{u}^\top X\mathbf{v} + \mathbf{v}\mathbf{u}^\top\mathbf{u}\mathbf{v}^\top\mathbf{v} \\ &= \lambda\mathbf{v} - 2\|\mathbf{u}\|^2\mathbf{v} + \|\mathbf{u}\|^2\|\mathbf{v}\|^2\mathbf{v} \\ &= \lambda\mathbf{v} - \|\mathbf{u}\|^2\mathbf{v} \qquad = 0 \end{aligned}$$

- NB: $\mathbf{u} = X\mathbf{v}$, $\|\mathbf{v}\|^2 = 1$ et $\|\mathbf{u}\|^2 = \|X\mathbf{v}\|^2 = \mathbf{v}^\top X^\top X\mathbf{v} = \lambda$

2ème axe factoriel

$\mathbf{v}_2$

- $X^\top X\mathbf{v}_2 = \lambda_2\mathbf{v}_2$, $\lambda_1 > \lambda_2 > \lambda_3 \ldots$
- $R\mathbf{v}_2$ est la projection linéaire de R qui maximise la variance

Matrice de projection

- $P = [\mathbf{v}_1, \mathbf{v}_2]$
- XP : projection de X en 2D
- Maximise l'information encodée
- et ainsi de suite ...

$$\underset{n}{X}\left(\begin{array}{c}\overset{p}{\\ \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_i^\top \\ \vdots \\ \boldsymbol{x}_n^\top \\ \\}\end{array}\right)\left(\begin{array}{c}\overset{2}{\\ XP \\ \\}\end{array}\right)_{n}$$

$$\underset{p}{P}\left(\begin{array}{c}\overset{2}{\\ \boldsymbol{v}_1\ \boldsymbol{v}_2 \\ \\}\end{array}\right)_{p} = P$$

k premiers axes factoriels

Lemme

Le sous-espace de dimension k maximisant la variance des projections contient nécessairement le sous-espace de dimension $k-1$.

k-ième axe factoriel

- $X^\top X \mathbf{v}_k = \lambda_k \mathbf{v}_k$, λ_k : k-ième valeur propre.
- Matrice de projection $P_k : \mathbb{R}^p \to \mathbb{R}^k$:

$$P_k = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_k \end{bmatrix}$$

- Les k vecteurs propres liées aux k plus grandes valeurs

Algorithme

1. Centrer les données : $\{\mathbf{x}_i \in \mathbb{R}^p\}_{i=1}^N \longrightarrow \{\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}} \in \mathbb{R}^p\}_{i=1}^N$
2. Calculer la matrice de covariance $\mathbf{\Sigma} = \frac{1}{N}\tilde{X}^\top \tilde{X}$ avec $\tilde{X}^\top = \begin{pmatrix} \tilde{\mathbf{x}}_1 & \cdots & \tilde{\mathbf{x}}_N \end{pmatrix}$
3. Calculer la décomposition en valeurs propres $\{\mathbf{v}_j \in \mathbb{R}^p, \lambda_j \in \mathbb{R}\}_{j=1}^p$ de $\mathbf{\Sigma}$
4. Ordonner les valeurs propres λ_j par ordre décroissant
5. Nouvelle base de représentation des données :

   $$P = (\mathbf{v}_1, \cdots, \mathbf{v}_k) \in \mathbb{R}^{p \times k}$$

   $\{\mathbf{v}_1, \cdots, \mathbf{v}_k\}$ sont les k vecteurs propres associés aux k plus grandes valeurs propres λ_j.
6. Projection de tous les points sur P :

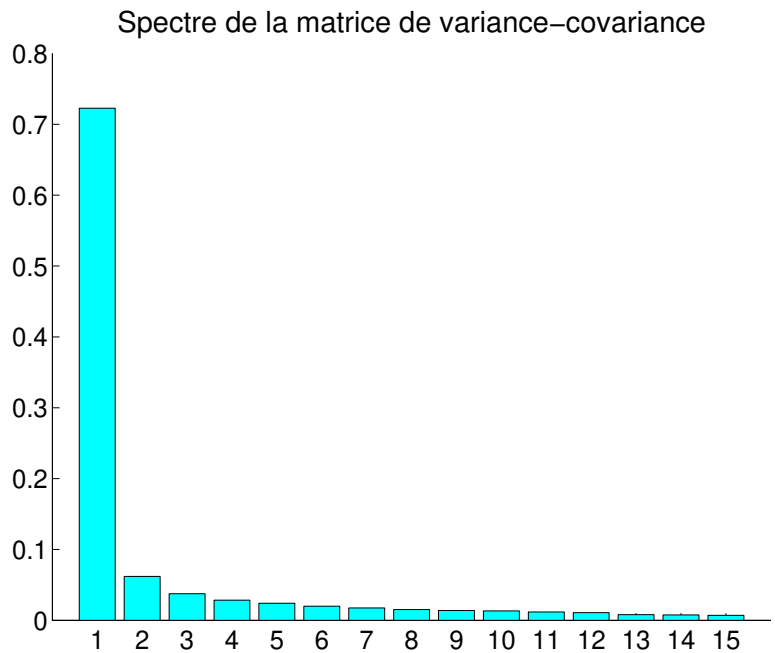
   $$C = XP \text{ avec } C = \begin{pmatrix} \mathrm{t}_1^\top \\ \vdots \\ \mathrm{t}_N^\top \end{pmatrix} \quad , \quad \mathrm{t}_i^\top = \tilde{X}(i,:)P$$

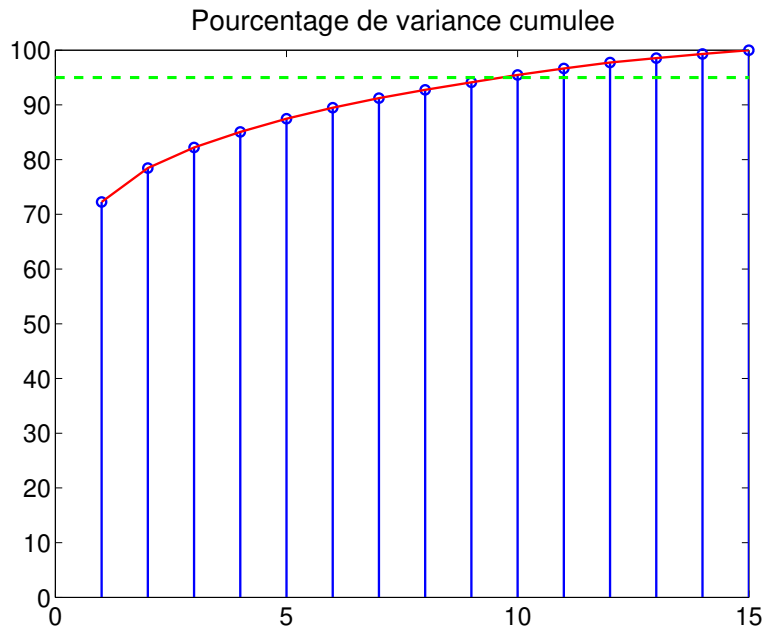
Propriétés des axes factoriels

- Les valeurs propres de $\mathbf{\Sigma}$ sont positives car $\mathbf{\Sigma}$ est une matrice semi-définie positive
- Le nombre d'axes factoriels est égal au nombre de valeurs propres non-nulles de $\mathbf{\Sigma}$.
- Variance expliquée par l'axe factoriel $\mathbf{v}_k$: $I_k = \mathbf{v}_k^\top \mathbf{\Sigma} \mathbf{v}_k = \lambda_k$.
- Variance totale des axes factoriels : $I = \sum_{k=1}^{p} \lambda_k$
- Pourcentage de variance expliquée par les d premiers axes

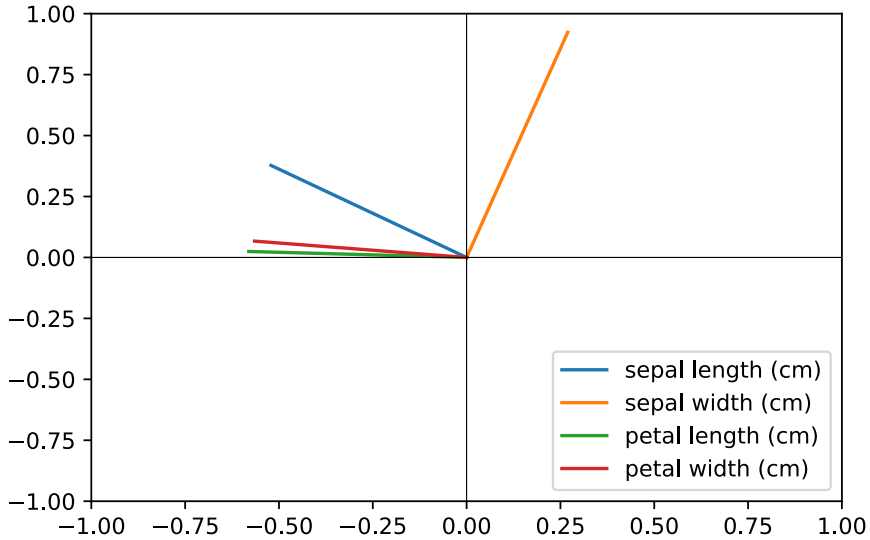
$$\frac{\sum_{k=1}^{d} \lambda_k}{\sum_{k=1}^{D} \lambda_k} \cdot 100$$

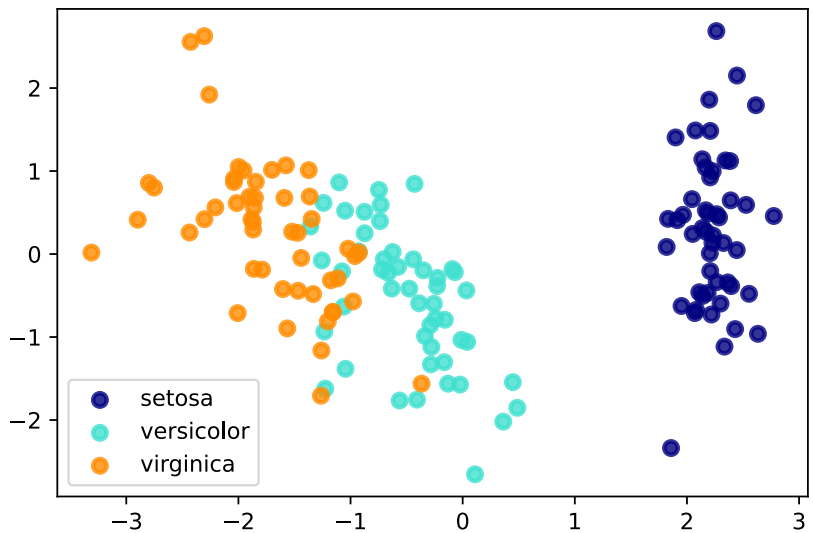
Propriétés des axes factoriels





Exemple sur Iris





Reconstruction

Revenir dans l'espace de départ

- XP : passage de $\mathbb{R}^p \to \mathbb{R}^d$
- Comment revenir dans l'espace de départ $\mathbb{R}^p$?
- Application sur la compression d'images
- Suppression du bruit

Matrice de projection inverse

- Inverse de P : P^{-1}
- P est orthogonale : $P^{-1} = P^\top$
- $P^\top \in \mathbb{R}^{p \times d}$
- Attention à la normalisation en reconstruction

$$\underset{2}{\begin{pmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \end{pmatrix}}_{p} = P^\top$$

$$\underset{n}{\begin{pmatrix} XP \end{pmatrix}}_{2} \underset{p}{\begin{pmatrix} \widetilde{\boldsymbol{x}}_1^\top \\ \widetilde{\boldsymbol{x}}_2^\top \\ \vdots \\ \widetilde{\boldsymbol{x}}_i^\top \\ \vdots \\ \widetilde{\boldsymbol{x}}_n^\top \end{pmatrix}} = \tilde{X}$$

Reconstruction

$$
\begin{array}{c}
\quad p \\
2 \begin{pmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \end{pmatrix} = P^\top
\end{array}
$$

$$
n \underbrace{\begin{pmatrix} \\ XP \\ \\ \end{pmatrix}}_{2} \underbrace{\begin{pmatrix} \widetilde{\boldsymbol{x}}_1^\top \\ \widetilde{\boldsymbol{x}}_2^\top \\ \vdots \\ \widetilde{\boldsymbol{x}}_i^\top \\ \vdots \\ \widetilde{\boldsymbol{x}}_n^\top \end{pmatrix}}_{p} = \widetilde{X}
$$

ACP : Conclusion

Méthode de réduction de dimension

- Matrice de projection P : d premiers vecteurs propres de $\mathbf{\Sigma}$
- Centrer/réduire les données
- Projection : $X_d = XP$
- Information de l'axe $\mathbf{v}_k$: λ_k
- Reconstruction : $\tilde{X} = X_d P^\top$
- Les derniers vecteurs propres contiennent le "bruit"

Limites

- Méthode de projection linéaires
- Hypothèse de distribution gaussienne

Bonus

Un autre mode de calcul du gradient

$\mathcal{J}(\mathbf{u},\mathbf{v}) = -2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2$, le minimum est atteint lorsque le gradient s'annule.

gradient d'une fonction : $J : \mathbb{R}^n \longrightarrow \mathbb{R}$

par définition le gradient $\nabla_J(\mathbf{v})$ s'obtient en posant $\phi(t) = J(\mathbf{v} + t\mathbf{d})$ où $\mathbf{d}$ est un vecteur de $\mathbb{R}^p$ et en calculant la valeur de la dérivée de ϕ au point zéro qui vérifie $\phi'(0) = \mathbf{d}^\top\nabla_J(\mathbf{v})$.

Faisons le calcul

$$\begin{aligned}
\phi(t) &= \mathcal{J}(\mathbf{u},\mathbf{v}+t\mathbf{d}) \\
&= -2(X(\mathbf{v}+t\mathbf{d}))^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}+t\mathbf{d}\|^2 \\
&= -2(\mathbf{v}^\top X^\top\mathbf{u} + t\mathbf{d}^\top X^\top\mathbf{u}) + \|\mathbf{u}\|^2(\|\mathbf{v}\|^2 + t^2\|\mathbf{d}\|^2 + 2t\,\mathbf{v}^\top\mathbf{d})
\end{aligned}$$

$$\begin{aligned}
\phi'(t) &= -2\mathbf{d}^\top X^\top\mathbf{u} + \|\mathbf{u}\|^2(2t\|\mathbf{d}\|^2 + 2\mathbf{v}^\top\mathbf{d}) \\
\phi'(0) &= -2\mathbf{d}^\top X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\,\mathbf{v}^\top\mathbf{d}
\end{aligned}$$

$$= d^\top\left(-2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v}\right) \;\Rightarrow\; \nabla_{\mathbf{v}}\mathcal{J}(\mathbf{v}) = -2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v}$$