# Recap PCA

**Objectif** : Passer d'un espace de dimension $p$ des features à un espace de dimension $k<p$.

# Cela peut-être à des fins de visualisation ($k=2$ ou $3$), computationnelles (complexité) ou encore pour éviter l'overfitting (mais ce n'est pas "une bonne pratique" : mieux vaut régulariser).
+Utilisation annexe importante : avoir des nouvelles features décorrélées.

**Principe** : On va projeter nos données $x_1, .., x_n \in \mathbb{R}^p$ dans un espace de dimension $k<p$ telle que la "variance totale" résultante est maximisée.

. On écrit $X \in \mathbb{R}^{n\times p}$, $n$ le nombre de données, $p$ le nombre features :

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = (X_1 | \dots | X_p)$$

↳ En pratique : on centrera et réduira nos données au préalable.

$$x \longrightarrow \frac{x-\bar{x}}{\sigma}$$

# Rappel de Maths préliminaires

## I/ SVD : Singular Value Decomposition

. Pour $X \in \mathbb{R}^{n\times p}$, on peut écrire $\boxed{X = UDV^T}$, $U \in O_n(\mathbb{R})$ (i.e. orthogonale), $V \in O_p(\mathbb{R})$, $D = \begin{pmatrix} D_{11} & & 0 \\ & \ddots & \\ 0 & & D_{mm} \end{pmatrix}$, $m = \min(n,p)$ et $D_{11} \geq \dots \geq D_{mm} \geq 0$.

⊕ Si $n>p$ : On peut réécrire cela comme $\boxed{X = UDV^T}$, $U \in \mathbb{R}^{n\times p}$, $D \in \mathbb{R}^{p\times p}$, $V \in O_p(\mathbb{R})$

où $U = (U_1 | \dots | U_n) \xrightarrow{\text{devient}} (U_1 | \dots | U_p)$.

(On supprime les dernières colonnes de $U$) et $D$ diagonale

En effet :

$$UDV^T = (U_1 | \dots | U_n) \begin{pmatrix} D_{11} & & 0 \\ & \ddots & \\ 0 & & D_{pp} \\ & & \end{pmatrix} \cdot \begin{pmatrix} V_1^T \\ \vdots \\ V_p^T \end{pmatrix}$$

$$= (U_1 | \dots | U_n) \begin{pmatrix} D_{11} V_1^T \\ \vdots \\ D_{pp} V_p^T \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sum_{i=1}^{p} U_i D_{ii} V_i^T$$

# On peut donc supprimer les dernières colonnes

⊕ Si $n<p$ : On réécrit $X = UDV^T$, $V^T = \begin{pmatrix} V_1^T \\ \vdots \\ V_n^T \end{pmatrix}$, $U \in O_n(\mathbb{R})$, $D \in \mathbb{R}^{n\times n}$.

↳ Cette réécriture s'appelle thin SVD (en anglais)

· On a alors : $X^TX = (UDV^T)^T(UDV^T)$

$X^TX = VD^TDV^T$

$\boxed{X^TX = V\Lambda V^T}$, $\Lambda = D^TD$, $V \in O_p(\mathbb{R})$

· $\Lambda$ → contient les valeurs propres de $X^TX$ (valeurs singulières) qui sont réelles $\geq 0$.

· $V$ → contient les vecteurs propres qui forment une base orthonormée de $\mathbb{R}^p$ $(v_1, \dots, v_p)$

## II | Produit scalaire et projection en 2D


$\cos(\alpha) = \frac{p_u(x)}{\|x\|_2}$ ← projection orthogonale de $x$ sur la droite avec vecteur directeur $u$

Donc $\langle x,u\rangle = \|x\|_2 \cdot \|u\|_2 \cdot \cos(\alpha) = \|u\|_2 \cdot p_u(x)$

Si $\|u\|_2 = 1 \Rightarrow \langle x,u\rangle = p_u(x)$

## III | Caractérisation projection orthogonale sur un plan dans $\mathbb{R}^n$, $n \geq 3$

Rappel : $\Pi_P(x) = \underset{y\in P}{\arg\min} \|y-x\|_2^2$ pour $P$ un plan. → projection orthogonale de $x$ sur $P$.

· Prenons une base orthonormée de $P$ $(v_1,v_2)$.

$\forall y \in P,\ y = \lambda_1 v_1 + \lambda_2 v_2$

$\|x-y\|_2^2 = \|x\|_2^2 - 2\langle x,y\rangle + \|y\|_2^2$.

$$\boxed{\|x-y\|_2^2 = \|x\|_2^2 + \lambda_1^2+\lambda_2^2 - 2\lambda_1\langle x,v_1\rangle - 2\lambda_2\langle x,v_2\rangle}$$

Posons $\varphi(\lambda_1,\lambda_2) = -2\lambda_1\langle x,v_1\rangle - 2\lambda_2\langle x,v_2\rangle + \lambda_1^2+\lambda_2^2$

$$\begin{cases} \frac{\partial\varphi}{\partial\lambda_i} = -2\langle x,v_i\rangle + 2\lambda_i, & i=1,2 \\ \frac{\partial^2\varphi}{\partial\lambda_i^2} = +2, & i=1,2 \\ \frac{\partial^2\varphi}{\partial\lambda_1\partial\lambda_2} = 0 \end{cases}$$

↳ $\frac{\partial^2\varphi}{\partial\lambda^2} > 0$ positive définie (Jacobienne)

donc la fonction est convexe (strictement)

⇒ Minimum atteint pour $\frac{\partial\varphi}{\partial\lambda} = 0$ ie

$$\boxed{\lambda_i = \langle x,v_i\rangle,\ i=1,2}$$

↳ $\boxed{y = \langle x,v_1\rangle v_1 + \langle x,v_2\rangle v_2} = \Pi_P(x)$.

· On généralise facilement à un espace $E_k$ de dimension $k$ en prenant $(v_1,\dots,v_k)$ une base orthonormée de $E_k$ :

On obtient
$$\boxed{\begin{aligned} \Pi_{E_k}(x) &= \sum_{i=1}^{k} \langle x,v_i\rangle v_i \\ \Pi_{E_k}(x) &= \sum_{i=1}^{k} (x^Tv_i)v_i \end{aligned}}$$

# PCA · Introduction

· la "variance totale" est définie comme : $\frac{1}{n}\sum_{i=1}^{n}\|x_i-\bar{x}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|x_i\|_2^2$

données centrées

$= \frac{1}{n}\sum_{i=1}^{n}\|X_i\|_2^2$

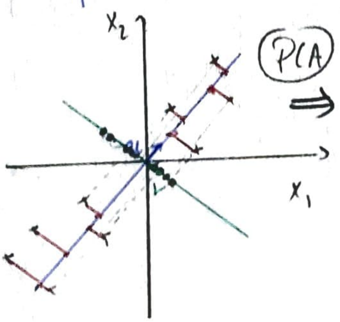
(# le facteur $\frac{1}{n}$ n'est pas important pour la suite).

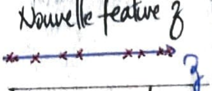
↳ On cherche donc un espace de dimension k tq la projection des $x_i$ sur cet espace donne des données (et nouvelles features associées) qui gardent le maximum de variance totale possible.

ie : on aura une nouvelle matrice $Y = \begin{pmatrix} y_1^T \\ y_n^T \end{pmatrix}$, $y_i \in \mathbb{R}^k$ telle que

$\sum \|y_i\|_2^2$ est "maximale" dans le sens défini ci-dessus.

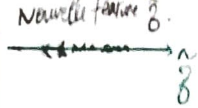
Example : $\boxed{2D \longrightarrow 1D}$ (ou $\boxed{pD \rightsquigarrow 1D}$)



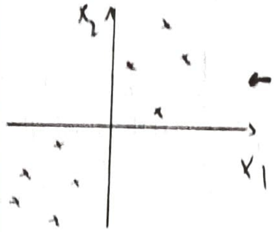

↳ On veut garder une "dispersion" maximale

Bon choix de projection




# Mauvais choix de projection

↳ les nouvelles données sont très rapprochées.




← Données originelles.

— Bon choix de projection

— Mauvais choix

· Prenons le cas où on a des données avec 2 features qu'on veut "compresser" en une seule. On cherche $\vec{u}$ ($\|u\|_2=1$) tq les données projetées sur $\vec{u}$ ($x_i^Tu$, $\forall i=1,\dots,n$) gardent une variance maximale.

ie $Xu = \begin{pmatrix} x_1^Tu \\ \vdots \\ x_n^Tu \end{pmatrix}$ et on a vu que

$x_i^Tu = p_u(x_i)$ pour $u$ de norme 1.

On cherche $u \in \mathbb{R}^2$ tq : $\boxed{u = \underset{v\in\mathbb{R}^2,\ \|v\|_2=1}{\operatorname{argmax}} \|Xv\|_2^2}$

ou, plus généralement : $\boxed{u = \underset{v\in\mathbb{R}^p,\ \|v\|_2=1}{\operatorname{argmax}} \|Xv\|_2^2}$

↳ Si on part de $\mathbb{R}^p$, $p \geq 2$.

· On a : $\|Xv\|_2^2 = v^TX^TXv = v^TVD^TDV^Tv$ (avec $\|v\|_2=1$)

Ecrivons $a = V^Tv$, $\|a\|_2 = \|v\|_2 = 1$, $\|Xv\|_2^2 = a^TD^TDa$

$\|Xv\|_2^2 = \sum_{i=1}^{m} a_i^2 D_{ii}^2 \leq D_{11}^2 \sum a_i^2 = D_{11}$

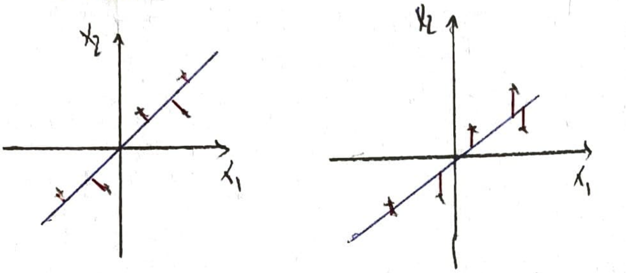
Donc $a = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ maximise $\|Xv\|_2^2$ ie $\boxed{v = Va = V_1}$

↳ la direction qui maximise $\|Xv\|_2^2$ est $V_1$, vecteur propre associé à $D_{11}^2$ : $\boxed{\|XV_1\|_2^2 = \|D_{11}U_1\|_2^2 = D_{11}^2}$.

Remarques : les données projetées $x_i^T v$ sont centrées ($\sum x_i^T v = \underbrace{(\sum x_i^T)}_{=0} v = 0$), donc la variance totale est bien $\sum_{i=1}^{n} (x_i^T v)^2 = \|Xv\|_2^2$

(# facteur $\frac{1}{n}$ pas important pour trouver le maximum)

• Ne pas confondre PCA et regression linéaire.




PCA — Regression linéaire

↳ les "traits rouge" (ce qu'on cherche à minimiser sont différents).

## Formulation du problème

Notre problème peut se résumer à :

$$\boxed{\underset{(v_1,\dots,v_k)\in(\mathbb{R}^p)^k}{\operatorname{argmax}} \left\{\sum_{l=1}^{k} \|Xv_l\|_2^2\right\} \text{ tels que } v_i^T v_j = \delta_{i,j}}$$

En effet, on a, pour un espace $E_k$ de dimension $k$ et une base orthonormée $(v_1,\dots,v_k)$ :

$$\Pi_{E_k}(x_i) = \sum_{l=1}^{k} (x_i^T v_l) \cdot v_l$$ (projection orthogonale sur $E_k$)

N.B. $\sum_{i=1}^{n} \|\Pi_{E_k}(x_i)\|_2^2 = \sum_{i=1}^{n} \sum_{l=1}^{k} (x_i^T v_l)^2$

Dans ce nouvel espace, on peut réécrire les coordonnées des $x_i$ projetées dans la base orthonormée :

$$Y = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} = \left(\begin{array}{c|c|c|c} x_1^T v_1 & x_1^T v_2 & \dots & x_1^T v_k \\ \vdots & \vdots & & \vdots \\ x_n^T v_1 & x_n^T v_2 & & x_n^T v_k \end{array}\right)$$

nouvelles données projetées

$$\boxed{Y = (Xv_1|\dots|Xv_k)} = (Y_1|\dots|Y_k)$$

La variance totale est :

↳ $\boxed{\sum_{i,l} x_i^T v_l = \sum_{l=1}^{k} \|Xv_l\|_2^2}$

## Résolution

On peut commencer par $v_1$ pour la 1ère feature puis chercher $u$ tq $\|Xu\|_2^2$ est maximale et $\|u\|_2 = 1$ et $u^T v_1 = 0$ (pb séparable).

On peut montrer facilement que $\boxed{u = v_2}$.

Récursivement, on obtient $\boxed{(v_1,\dots,v_k)}$ les vecteurs propres de $X^T X$ associées aux valeurs propres $D_{11}^2 \geq D_{22}^2 \geq \dots \geq D_{kk}^2 \geq 0$ :

On a alors : $Y = (Y_1|\dots|Y_k) = (Xv_1|\dots|Xv_k)$

$Y = (D_{11}u_1|\dots|D_{kk}u_k)$

↳ la variance totale est alors : $\boxed{\sum_{l=1}^{k} \|D_{ll} v_l\|_2^2 = \sum_{l=1}^{k} D_{ll}^2}$

A noter que la variance totale avant projection est : $\boxed{\sum_{l=1}^{m} D_{ll}^2}$.

Une façon de choisir $k$ (dimension de l'espace sur lequel on projète) :

$$\boxed{\min\left\{k \,\middle|\, \frac{\sum_{l=1}^{m} D_{ll}^2}{\sum_{l=1}^{m} D_{ll}^2} \geq 0.95\right\}}$$

↳ ou autre : 0.99, 0.90, 0.80, ...

- Remarques :

  - les données sont centrées

$$XV_1 = \begin{pmatrix} x_1^T V_1 \\ \vdots \\ x_n^T V_1 \end{pmatrix} \rightarrow \left(\sum x_i^T V\right) = \left(\sum x_i^T\right) V = 0$$

  - les features créées sont décorrélées :

$$i \neq j, \ (XV_i)^T (XV_j) = V_i^T X^T X V_j = \underbrace{V_i^T V \Lambda V^T V_j}_{=0}$$

⇒ D'ailleurs, on utilise cette méthode pour "décorréler des features" (ie avoir de nouvelles features décorrélées)

  - les features créées perdent en interprétabilité/explicabilité

$\boxed{\text{∘ Idée générale : garder le plus de "dispersion" possible entre les données.}}$