

# Réduction de Dimensions

→ Apprentissage non supervisé  $X \in \mathbb{R}^d$

→  $\psi: \mathbb{R}^d \rightarrow \mathbb{R}^p$ ,  $p \ll d$ . 

$\psi(X)$  ?

→ Pourquoi? Interprétabilité  
Calculatoire

Statistique (ne pas tenir compte du bruit en tant que dimension supplémentaire)

## Théorème

Classification  $[0,1]^d \times \{-1, 1\}$ ,  $(X_1, Y_1) — (X_n, Y_n)$ .

On le classifieur kNN. On note  $g^*$  le classifieur de Bayes.

Sat  $L > 0$  tq  $x \mapsto \eta(x) = \mathbb{P}[Y=1 | X=x]$  soit  $L$ -Lipschitz.

Alors,

$$\mathbb{P}\{g_m(x) \neq Y | \mathcal{D}_n\} \leq \mathbb{P}\{g^*(x) \neq Y\} \left(1 + \sqrt{\frac{8}{K}}\right) + \frac{6L\sqrt{d} + k}{n^{1/d+1}}$$

Si  $L > 1$ , alors on peut trouver une

$$\leq \text{ si } n \geq \left(\frac{6Ld + k}{\epsilon}\right)^d$$

distribution pour  $(X, Y)$  avec  $\eta(x) = \mathbb{P}[Y=1 | X=x]$   $L$ -Lipschitz et

$$n \geq \frac{(L+1)^d}{2}, \quad \mathbb{P}\{g_m(x) \neq Y | \mathcal{D}_n\} \geq \frac{1}{L}.$$

→ Validité de la dimension ←

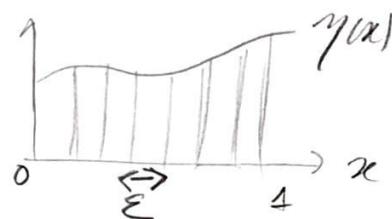
Plus la dimension augmente, plus on a besoin (exponentiellement) de données

## Classifieur de Bayes

$$g^*(x) = \begin{cases} 1 & \text{si } \gamma(x) \geq 0.5 \\ -1 & \text{sinon.} \end{cases}$$

$$\gamma(x) = P[Y=1|X=x]$$

Approximation de  $\gamma(x)$ .



En dimension  $d$ , il faut  $\varepsilon^{-d}$  points :  $\varepsilon = 10^{-2}$ ,  $d=1 \rightarrow 100$  points.  
 $d=10 \rightarrow 10^{30}$  points.

→ On cherche à construire une fonction  $\psi: \mathbb{R}^d \rightarrow \mathbb{R}^p$  de sorte que :

- ▷ Reconstruction :  $\varphi \cdot (\rho, \psi)(x) \approx x$ .
- ▷ Préservation des distances :  $d(\psi(x), \psi(x')) \approx d(x, x')$ .

## D - Méthodes linéaires

### Analyse en composantes principales

On suppose  $\psi$  et  $\varphi$  linéaires.

$$\begin{array}{ll} \text{minimiser}_{\psi, y} & \text{IE}[Ux - (\varphi \circ \psi)(x)]_2^2 \\ \text{st} & \psi(x) = Ux \\ & \varphi(y) = Vy \end{array}$$

$$\Rightarrow \begin{array}{l} \text{Trouver deux matrices } U \text{ et } V \text{ tq} \\ \min_{U, V} \text{IE}[Ux - UWx]_2^2 \quad (P) \\ U \in \mathbb{R}^{p \times d}, V \in \mathbb{R}^{d \times p} \quad \text{psd.} \end{array}$$

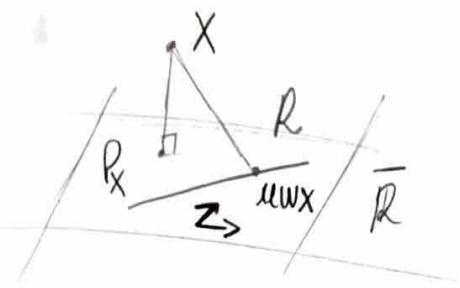
Lemme

Si  $(P)$  admet une solution, alors  $\exists V \in \mathbb{R}^{d \times p}$  tq  $\psi(x) = V^T x$  et  $\varphi(y) = Vy$  sont optimales.  
 $(U, W) = (V, V^T)$  est solution de  $(P)$ . De plus  $V^T V = I_p$ .

Preuve

Soit  $(U, V) \in \mathbb{R}^{p \times d} \times \mathbb{R}^{d \times p}$  solution de  $(P)$ .

$R = \cancel{UVW^T}$  de dimension  $n \leq p$  car  $\text{rg}(W) \leq p$ , sao de  $\mathbb{R}^d$  de dimension  $n$ .



Soit  $(v_1, \dots, v_n)$  base de  $R$ .  $\underbrace{(v_1, \dots, v_n, v_{n+1}, \dots, v_d)}_{\text{vecteurs}} \text{ base de } R^d$ .

$$v = [v_1 \dots v_p]$$

$$\bar{R} = \text{Vect}_d v \bar{q} \quad \left\{ \text{en } a \in R \subset \bar{R} \right.$$

On appelle  $P$  le projecteur orthogonal sur  $\bar{R}$ , puisque  $uWX \in R \subset \bar{R}$ , on a :

$$\|X - uWX\| \geq \|X - PX\|$$

$$\mathbb{E}[\|X - uWX\|^2] \geq \mathbb{E}[\|X - PX\|^2]$$

Par définition de  $P$ ,  $P = \underbrace{V(V^\top V)^{-1}V^\top}_{P} V^\top$

$$\text{On se retrouve donc avec : } \min_{V \in R^{d \times p}} \mathbb{E}[\|X - VV^\top X\|^2]$$

$$\text{st } V^\top V = I_p.$$

Soit  $V \in R^{d \times p}$  telle que  $V^\top V = I_p$ ,

$$\begin{aligned} \mathbb{E}[\|X - VV^\top X\|^2] &= \mathbb{E}[VV^\top X]^2 + \mathbb{E}[X^\top V V^\top V V^\top X] - 2\mathbb{E}[X^\top V V^\top X] \\ &= \mathbb{E}[VV^\top X]^2 - \mathbb{E}[X^\top V V^\top X] \\ &= \text{tr}(\mathbb{E}[V^\top X X^\top V]) = \mathbb{E}[\text{tr}(V^\top X X^\top V)] \\ &= \mathbb{E}[\text{tr}(V^\top X X^\top V)] \end{aligned}$$

$$\max_{V \in R^{d \times p}} \text{tr}(V^\top \mathbb{E}[XX^\top] V)$$

$$\text{st } V^\top V = I_p$$

□

## Théorème

C est R<sup>dxd</sup> SDP et symétrique.

$C = V \Lambda V^T$  décomposition en les pôles de C avec  $\Lambda = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_d \end{pmatrix}$

et  $V \in \mathbb{R}^{dxd}$  orthogonale.

k entier non nul tq k < d, alors :

$$\max_{\substack{U \in \mathbb{R}^{d \times k} \\ U^T U = I_k}} \text{tr}(U^T C U) = \text{tr}(V_+^T C V_+)$$

$$\min_{\substack{U \in \mathbb{R}^{d \times k} \\ U^T U = I_k}} \text{tr}(U^T C U) = \text{tr}(V_-^T C V_-)$$

$$V = \begin{bmatrix} V_+ & & V_- \end{bmatrix}$$

$\xleftarrow{k} \quad \xleftarrow{k}$

Preuve

Montrer que : ①  $\forall U \in \mathbb{R}^{d \times k}$  tq  $U^T U = I_k$  :  $\text{tr}(U^T C U) \geq \inf_{\beta \in \mathcal{B}} \sum_{i=1}^k d_i \beta_i$

avec  $\mathcal{B} = \{\beta \in [0, 1]^d \mid \pi^T \beta = k\}$ .

$$\text{② } \inf_{\beta \in \mathcal{B}} \sum_{i=1}^k d_i \beta_i = \sum_{i=1}^k d_i$$

$$\text{③ } \sum_{i=1}^k d_i = \text{tr}(V^T C V_+)$$

④  $\forall U \in \mathbb{R}^{d \times k}$  avec  $U^T U = I_k$ .

$$\text{tr}(U^T C U) = \text{tr}(U^T V \Lambda V^T U) \quad \text{avec } \Lambda = \text{diag}(d_i)$$

$\vdots \quad \vdots \quad \vdots$

$$= \text{tr}(B^T \Lambda B) = \sum_{j=1}^k b_j^T \Lambda b_j, \quad B = [b_1 \ | \ \dots \ | \ b_k]$$

$$= \sum_{j=1}^k \sum_{i=1}^d d_i (b_j)_i^2$$

$$= \sum_{i=1}^d d_i \underbrace{\sum_{j=1}^k (b_j)_i^2}_{:= \beta_i^2} = \sum_{i=1}^d d_i \beta_i^2$$

$$B = \begin{bmatrix} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & \end{bmatrix}$$

Par ex.  $B^T B = \underbrace{U V V^T U}_U = U^T U = I_k$  D'où B est orthonormée.  
car V est orthogonale

$A = [B | e]$  de sorte que A est orthogonale (Théorème de la base incomplète)

Alors,  $A^T A = I_d = AA^T$ . En particulier, si  $i \in \{1, d\}$ ,

$$\sum_{j=1}^d A_{ij}^2 = 1 = \underbrace{\sum_{j=1}^k \beta_{ij}^2}_{\leq 1} + \underbrace{\sum_{j=k+1}^d \alpha_{ij}^2}_{\geq 0} = 1 \quad \text{car } A = [B \mid C]$$

Ainsi  $\beta_{ij} \in [0, 1]$ .

Par ailleurs,  $\sum_{i=1}^d \beta_i^2 = \sum_{i=1}^d \sum_{j=1}^k \beta_{ij}^2 = \text{Tr}(B^T B) = \text{Tr}(I_k) = k$  et  $B^T B \in \mathcal{J}$ .

On en déduit que  $\text{tr}(B^T B) \geq \inf_{\beta \in \mathcal{J}} \sum_{i=1}^d d_i \beta_i^2$ .

① Soit  $\bar{\beta} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathcal{J}$ . On a bien  $\bar{\beta} \in \mathcal{J}$ .

$$\forall \beta \in \mathcal{J}, \sum_{i=1}^d d_i \beta_i^2 - \sum_{i=1}^d d_i \bar{\beta}_i^2 = \underbrace{\sum_{i=1}^k d_i (\bar{\beta}_i - 1)}_{\geq 0} + \underbrace{\sum_{i=k+1}^d d_i \bar{\beta}_i^2}_{\geq 0} \quad (\text{def } \bar{\beta})$$

$$= dk \sum_{i=1}^k (\bar{\beta}_i - 1) + dk \sum_{i=k+1}^d \bar{\beta}_i^2$$

$$= dk \left( \sum_{i=1}^d \bar{\beta}_i - k \right) \quad \text{or} \quad \sum_{i=1}^d \bar{\beta}_i = k \text{ par construction.}$$

$$= 0$$

$$\text{Et donc } \min_{\beta \in \mathcal{J}} \sum_{i=1}^d d_i \beta_i^2 = \sum_{i=1}^d d_i \quad \begin{pmatrix} d & 0 \\ 0 & \dots \\ 0 & d \end{pmatrix}$$

$$\textcircled{2} \quad V = \boxed{\begin{array}{c|c} V_- & \\ \hline & \end{array}} \quad \text{tr}(V^T C V) = \text{tr}\left(V_-^T V_- \underbrace{V_+^T V_+}_{J} \right)$$

$$= \sum_{i=1}^k d_i$$

$$J = d \boxed{\begin{array}{cccc} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{array}}$$

$$\begin{cases} \text{maximise } b \in \mathbb{R} \text{ tel que } \\ \|u\| = 1 \\ \text{et } u^T u = b \end{cases}$$

$E[X^T X]$  est bien SPD et symétrique  $\rightarrow$  solution :  $V = [V_{dp} | -V_d]$   
 vecteurs propres majoritairement de  $E[X^T X]$   
 (liés associés aux grandes  $v_p$ )

Fonction de compression :  $V(x) = V^T x = \sum v_i v_i^T x$

Erreur de reconstruction :  $E[(X - VV^T X)^2] = E[\|X\|^2] - b(E[X^T X])V^T V$   
 $\rightarrow$  On peut choisir la dimension de compression selon l'erreur de reconstruction que l'on souhaite permettre

$$= b(E[X^T X]) - b$$

$$= \sum_{i=1}^d d_i \quad \text{rang} : d_1 \leq \dots \leq d_d$$

lien avec la maximisation de la variance :

$V$ : linéaire       $\mathbb{E}$ : linéaire

$\Leftrightarrow$  Rechercher  $X$  sur un sous-espace de  $\mathbb{R}^d$ .  $\rightarrow$  Sous-espace affine ?

$$\min_{U, W, V \in \mathbb{R}^d} E[\|X - (UW + V)\|^2]$$

$U$  et  $W$  fixés,  $\mu = E[X] - UW E[X]$ , on obtient :

$$E[\|X - E[X] - UW(X - E[X])\|^2]$$

$$E[(X - E[X])(X - E[X])^T] = \text{Var}(X)$$

Sélection  $\rightarrow$  vecteurs propres de  $\text{Var}(X)$ .

Composantes principales "maximisent la variance de  $X$ " ?  $b(u^T \text{Var}(X) u)$

- On définit  $u_1 \in \mathbb{R}^d$  tel que  $\|u_1\|=1$ ,  $u_1$  telle que  $\max_{u \in \mathbb{R}^d} b(u^T \text{Var}(X) u)$  et  $\|u\|=1$

- $(u_1, \dots, u_k)$ ,  $u_{k+1} \in \arg \max_{u \in \mathbb{R}^d} b(u^T \text{Var}(X) u)$

et  $\|u\|=1$

$u \perp u_1, \dots, u \perp u_k$

→ Les composantes principales  $v_{d+1} \dots v_d$  correspont aux ( $u_1 - u_d$ ) .

▷ Par le théorème de fact à l'heure,  $u_d$  le  $u_p$  de  $X$  associé à la plus grande cp de  $\text{Var}(X)$ .

$\|u_d\|=1$  et  $u_d \in \arg\max_{u \in \text{Lat}(\text{Var}(X))} u^T u$

$$\text{st } u^T u = 1$$

On peut donc choisir  $u_d = v_d$ .

▷  $(u_1 \dots u_d) = (v_d \dots v_{d-k+1}) , u_{d+1} ?$

$u_{d+1} \in \arg\max_{\substack{\text{st } u^T u = 1 \\ u \perp u_1 \dots u_d}} \underbrace{\text{Var}(u^T X)}_{\text{cots}} + \underbrace{\text{Var}(u_1^T X u_1)}_{\text{cots}} + \dots + \underbrace{\text{Var}(u_k^T X u_k)}_{\text{cots}}$

$u_{d+1} \in \arg\max_{u \in \text{Lat}(\text{Var}(X))} u^T u$

st  $u = (u_1 / \dots / u_k / u)$   
 $u^T u = 1$

$V = [v_1 | \dots | v_{d-k+1} | v_{d-k}]$  est solution du pb d'optimisation.

Donc on peut choisir  $u_{d+1} = v_{d-k}$ .

▷ Dire avec la matrice de Gram.

• linéaire compression  $\Leftrightarrow$  espace affine  
• affine décompression

Diagonaliser  $\text{Var}(X)$

$x_1, \dots, x_n$  iid  $\rightarrow X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}$

$C = X^T X \in \mathbb{R}^{d \times d}$ .  
 $d \gg 1, p$  petit.  
 $d \gg n$ .

$$K = XX^T = (\langle x_i, x_j \rangle)_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$$

Propriété

$$\boxed{K} / C$$

Remarque

$K = \sum_{i=1}^m d_i v_i v_i^T$  la décomposition en les propres de  $K$ . Alors,  $\exists d_1, \dots, d_{d-n} \in \mathbb{R}^+$

et  $(v_i)_{i=1}^m, \dots, v_{d-n}^T$  est la décomposition en les propres de  $C$ .

Preuve

- $C(X^T v_i) = d_i(X^T v_i)$
- $\text{rg}(C) = \text{rg}(X^T X) = \text{rg}(X^T X) = \text{rg}(K) \leq n$ .

Au plus  $n$  op non nulles de la décomposition en les propres de  $C$ .

Mais dans  $d_1, \dots, d_n$  il y a déjà au plus  $n$  op non nulles ( $\text{rg}(K)$ ).  
Donc les autres op de  $C = 0$ .

Pour construire  $\Psi(x) = V^T K$ ,  $v_i - v_p$  les majorantes de  $K$ .

$$V = \begin{bmatrix} K^T v_1 \\ \|K^T v_1\|_2 \\ \vdots \\ K^T v_p \\ \|K^T v_p\|_2 \end{bmatrix} \in \mathbb{R}^{d \times p}$$

Remarque

Matrice des données réduites  $\in \mathbb{R}^{n \times p}$ .

$$\begin{bmatrix} \frac{\Psi(x_1)^T}{\|\Psi(x_1)\|_2} \\ \vdots \\ \frac{\Psi(x_n)^T}{\|\Psi(x_n)\|_2} \end{bmatrix} = (V^T X)^T = X V = \begin{bmatrix} X X^T v_1 \\ \|X X^T v_1\|_2 \\ \vdots \\ X X^T v_p \\ \|X X^T v_p\|_2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{d_1 v_1}{\sqrt{d_1}} \\ \vdots \\ \frac{d_p v_p}{\sqrt{d_p}} \end{bmatrix} = \begin{bmatrix} \sqrt{d_1} v_1 \\ \vdots \\ \sqrt{d_p} v_p \end{bmatrix} \in \mathbb{R}^{n \times p}$$

$\Psi(x_i)^T$

Remarque : Lien avec le clustering spectral

- $W$  = matrice d'adjacence / similarité.
- $L = D - W$ ,  $D = \begin{pmatrix} d_1 & 0 \\ 0 & \ddots & 0 \\ & \ddots & d_n \end{pmatrix} = d_1 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^*$
- $\lambda_1, \dots, \lambda_p$  v.p de  $L$  associés aux plus petites  $v_p$ .
- algorithme des  $k$ -moyennes sur les représentations.
- On prend  $W = K$ ,  $L = d_K - K$

### Lien avec la SVD

$$A \in \mathbb{R}^{m \times n}$$

$$A \begin{array}{c} n \\ \diagdown \\ m \end{array}$$

$$A = UDV^T \text{ où } r = \text{rg}(A) \leq \min\{m, n\}.$$

$$= \begin{array}{c} n \\ \diagdown \\ U \end{array} \begin{array}{c} n \\ \diagup \\ V \end{array} \begin{array}{c} n \\ \diagdown \\ V \end{array} \begin{array}{c} m \\ \diagup \\ \sqrt{r} \end{array}$$

$$U^T U = I_r$$

$$V^T V = I_r$$

$$D = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \ddots & 0 \\ & \ddots & \sigma_r \end{pmatrix} \quad \sigma_i > 0$$

Theorem SVD

$$\rightarrow AA^T \in \mathbb{R}^{m \times m}$$

$u_i$  est v.p de  $AA^T$  avec v.p  $\sigma_i^2$ .

$$\rightarrow A^T A \in \mathbb{R}^{n \times n}$$

$v_i$  \_\_\_\_\_  $A^T A$  \_\_\_\_\_  $\sigma_i^2$ .

AP : 3 façons de faire :

1. Diagonaliser  $C$ .

2. Diagonaliser  $K$ .

3. SVD de  $X$ .  $\rightarrow$  pas de calcul de  $X^T X$ .  
 $\rightarrow$  pas besoin de calculer tous les v.p.  $\sigma_{i,n}$  / v.p de  $K = X^T X$ .

Matrice de représentation réduite des  $(X_1, \dots, X_m)$  est  $(\sqrt{\sigma_1} v_1, \dots, \sqrt{\sigma_p} v_p)$

## Kernell PCA

Astuce du noyau:  $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$

et RKHS.

$\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}}, \phi: \mathbb{R}^d \mapsto \mathcal{G}$

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$$

$$(\phi(x_1), \dots, \phi(x_n)) \rightarrow (\underbrace{\phi(x_1) - \bar{\phi}_n, \dots, \phi(x_n) - \bar{\phi}_n}_{z_i \in \mathcal{G}}) \text{ où } \bar{\phi}_n = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$$

$$z = \begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix} \quad y^T z = \langle x, y \rangle_{\mathcal{G}}$$

$$K_2 = z z^T = H K_2 H \text{ où } K_2 = (k(x_i, x_j))$$

$$H = \left( I_n - \frac{1}{n} u u^T \right) \text{ si } j \neq n$$

$$\psi(\phi(x) - \bar{\phi}_n) = \sum_{i=1}^n \alpha_i k(x, x_i) - f^*(k(x, x_i)) \rightarrow \text{On peut mg l'estime du noyau s'applique.}$$

Pt de vue RKHS:  $\psi(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_p(x) \end{pmatrix} \in \mathbb{R}^p$ ,  $f_i \in \mathcal{H}$  construit tq

$$\begin{cases} \max_{f \in \mathcal{F}} \sum_{i=1}^n \text{Var}_m(f_i(x)) \\ \text{st } \langle f_i, f_j \rangle_{\mathcal{G}} = S_{ij} \end{cases}$$

$n \gg 1$ ,  $w_{ij} \sim \mathcal{N}(0, 1)$  iid,  $\nu = w_x$ , alors:

$$\|D\| - \gamma \geq 1 - \delta, \quad \|w_x - w_x^*\|_2^2 - \|x - x^*\|_2^2 \leq \epsilon.$$

26/10/2021