Semaine 4 : Apprentissage non-supervisé et réduction de dimension

1 : Algèbre linéaire,SVD et matrices de covariances

Laurent Risser
Ingénieur de Recherche CNRS
Institut de Mathématiques de Toulouse / 3IA ANITI

lrisser@math.univ-toulouse.fr

0 : Introduction

La **factorisation de la matrice** est essentielle en *algèbre linéaire* dans et est ainsi utilisée dans de nombreux domaines dont l'*apprentissage automatique* :

- Cadre général → accélération de la résolution des problèmes de systèmes linéaires (il n'est pas toujours optimal d'inverser explicitement une matrice ou de calculer un déterminant).
- Apprentissage statistique → *identification de structures* au sein de données ou *réduction de dimension* des observations.

Différentes méthodes de factorisation de matrice existent :

- Racine carrée Cholesky d'une matrice
- Factorisation LU ou QR
- Décomposition de valeurs singulières (SVD)

Nous allons nous intéresser à la **SVD**, qui est la plus utile en apprentissage automatique.

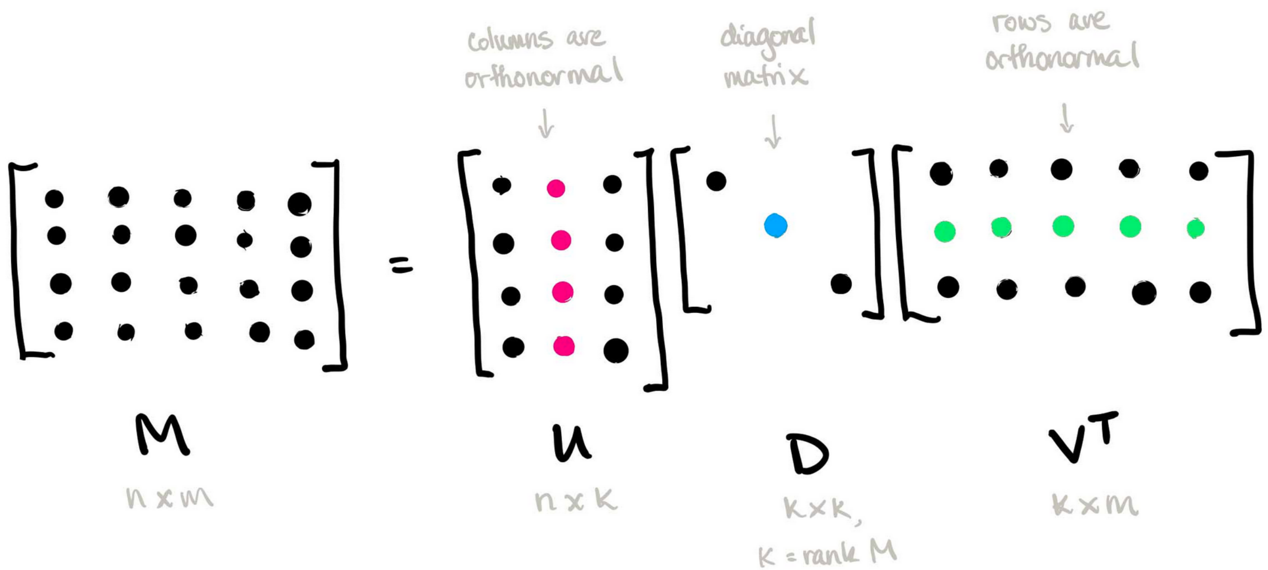


Image : https://www.math3ma.com/blog/understanding-entanglement-with-svd

Partie inspirée de :
- Jean Fresnel : *Algèbre des matrices*. 2013 (lien)
- Lloyd N Trefethen and David Bau III : Numerical linear algebra, volume 50. Siam 1997 (lien)
- Cours de Samuel Bernard (CNRS)

Définition 1. Soit K un corps commutatif ($\mathbb{R}$ ou $\mathbb{C}$). Un ensemble E muni d'une somme $+$ (application de $E \times E \to E$ commutative et associative avec élément 0 neutre et élément inverse) et d'une multiplication $\cdot$ par un scalaire appartenant à K (application de $K \times E \to E$) est un **espace vectoriel** sur K, $(E, K, +, \cdot)$, s'il possède les propriétés suivantes:

Pour tout $\mu, \lambda \in K$ et $x, y \in E$,

$$\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y, \tag{1}$$

$$(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x, \tag{2}$$

$$1 \cdot x = x, \tag{3}$$

$$\lambda \cdot (\mu \cdot x) = (\lambda\mu) \cdot x. \tag{4}$$

Définition 2. Soit E un espace vectoriel sur K. Une partie F non-vide de E est appelée **sous-espace vectoriel** de E, si pour tout $x, y \in F$ on a $x - y \in F$ et si pour tout $\lambda \in K$ et $x \in F$, on a $\lambda \cdot x \in F$. On utilise l'abréviation s.e.v. pour un sous-espace vectoriel.

Définition 3. Soit E et F deux e.v. sur K. Une application $f : E \to F$ est appelée **application linéaire** si

$$f(x+y) = f(x) + f(y) \qquad \text{pour tout } x, y \in E \tag{5}$$

$$f(\lambda x) = \lambda f(x) \qquad \text{pour tout } x \in E, \lambda \in K \tag{6}$$

Un **isomorphisme** est une application linéaire bijective. L'inverse d'un isomorphisme est aussi une application linéaire. Un **endomorphisme** de l'e.v. E est une application linéaire de E dans E. Un **automorphisme** est un endomorphisme bijectif.

Exemple : Matrice de rotation 2D

Définition 5. Soit E un e.v. sur K, on appelle **famille de E indexée par un ensemble I** une application de I dans E. Cette application associe $i \to x_i$. On note cette famille sous la forme $(x_i)_{i \in I}$. La notion de famille généralise celle d'une suite, où l'index n'est pas nécessairement entier.

Une famille $(x_i)_{i \in I}$ est **libre** si pour toute famille $(\lambda_i)_{i \in I}$, $\lambda_i \in K$ à support fini (c.-à-d. $\lambda_i \neq 0$ pour un nombre fini de valeurs $i \in I$), l'égalité

$$0 = \sum_{i \in I} \lambda_i x_i \tag{10}$$

implique $\lambda_i = 0$ pour tout $i \in I$.

Une famille $(x_i)_{i \in I}$ est **génératrice** si pour tout élément $x \in E$, il existe une famille $(\lambda_i)_{i \in I}$, $\lambda_i \in K$ à support fini telle que

$$x = \sum_{i \in I} \lambda_i x_i. \tag{11}$$

Une famille $(x_i)_{i \in I}$ de E est une **base** de E si elle est libre et génératrice. Ça veut dire que pour tout $x \in E$, il existe une famille *unique* $(\lambda_i)_{i \in I}$, $\lambda_i \in K$ à support fini telle que $x = \sum_{i \in I} \lambda_i x_i$. L'élément λ_i est la **coordonnée de x** dans la base $(x_i)_{i \in I}$.

Proposition 1.3. *Soit E un e.v. de type fini, c.-à-d. que E admet une famille génératrice finie. Alors*

i) E admet une base $\mathcal{B}$ indexée sur un ensemble fini. Le cardinal (nombre d'éléments) de cet ensemble est noté $\operatorname{card} \mathcal{B}$.

ii) Soit $\mathcal{B}'$ une base de E, alors $\mathcal{B}'$ est une famille indexée sur un ensemble fini et $\operatorname{card} \mathcal{B}' = \operatorname{card} \mathcal{B}$.

Le $\operatorname{card} \mathcal{B}$ *s'appelle dimension de E sur K et se note* $\dim_K E$. $\dim_K E = 0$ *si* $E = \{0\}$.

Définition 6. Soit E, F deux e.v. sur K de dimensions $n \geq 1$, $p \geq 1$, $(e_i)_n$, $(f_j)_p$ des bases de E et F, et $f : E \to F$ une application linéaire telle que

$$f(e_i) = \sum_{j=1}^{p} a_{ji} f_j. \tag{13}$$

Alors la **matrice de f dans les bases $(e_i)_n$ et $(f_j)_p$**, notée $\mathrm{Mat}(f, (e_i)_n, (f_j)_p)$ est la famille $(a_{ji})_{(j,i)\in[1,p]\times[1,n]}$.

$$\mathrm{Mat}(f, (e_i)_n, (f_j)_p) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & & & \\ \vdots & & & \\ a_{p1} & a_{p2} & \dots & a_{pn} \end{pmatrix} \tag{14}$$

Cette matrice est une **matrice à** n **colonne et à** p **lignes** et prend ses **entrées** $a_{ji} \in K$.

Définition 7. On définit le **produit matriciel** entre deux matrices $A \in M_{p,n}(K)$ et $B \in M_{n,\ell}(K)$ comme la matrice $C \in M_{p,\ell}(K)$ avec

$$c_{jk} = \sum_{i=1}^{n} a_{ji} b_{ik}. \tag{16}$$

On note le produit matriciel de A et B par AB.

Définition 8. Soit $A \in M_{p,n}(K)$ une matrice $p \times n$. On appelle **transposée** de A la matrice $[b_{ij}]_{ij} \in M_{n,p}(K)$ définie par $b_{ij} = a_{ji}$, et est notée ${}^tA$ ou $\bar{A}$.

On a ${}^t(A+B) = {}^tA + {}^tB$, ${}^t(\lambda A) = \lambda {}^tA$, ${}^t(AC) = {}^tC {}^tA$, et ${}^t({}^tA) = A$. On dit qu'une matrice carrée $A \in M_n K$ est **symétrique** si ${}^tA = A$.

Définition 9. Déterminant d'une matrice carrée. Soit $A = [a_{ij}]_{ij} \in M_n(K)$ une matrice $n \times n$. On appelle **déterminant de** A, noté $\det A$ le scalaire défini par

$$\det A = \sum_{\sigma \in \mathrm{perm}(n)} \mathrm{sgn}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \ldots a_{\sigma(n)n}. \tag{17}$$

On note aussi

$$\det A = \begin{vmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} \end{vmatrix}. \tag{18}$$

La **signature** d'une permutation $\mathrm{sgn}(\sigma)$ vaut 1 si la permutation σ est paire et -1 si la permutation est impaire. Une permutation est **paire** (respectivement **impaire**) si elle est le produit d'un nombre pair (impair) de transpositions (permutation de deux éléments entre eux).

Exemple : perm(2) = { (1,2) , (2,1) }. Retrouvez alors le determinant d'une matrice 2*2 ?

Le déterminant possède les propriétés suivantes:

1. $\det I = 1$.
2. $\det A = \det {}^t A$.
3. $\det[C_1 ... \sum \lambda_i C_j ... C_n] = \lambda_i \det[C_1 ... C - n] = \lambda_i \det C$, où C_i sont les colonnes de C.
4. $$\det \begin{bmatrix} L_i \\ \sum \lambda_i L_j \\ L_n \end{bmatrix} = \lambda_i \begin{bmatrix} L_i \\ \vdots \\ L_n \end{bmatrix}. \tag{19}$$
5. $\det[C_{\sigma(1)} ... C_{\sigma(n)}] = \mathrm{sgn}(\sigma) \det C$.
6. idem pour les lignes.
7. Si une ligne (colonnes) est une combinaison linéaire d'autres lignes (colonnes), alors $\det = 0$.
8. Soit A avec $a_{i1} = 0$ pour $i \geq 2$ et $B = [b_{ij}]$ avec $b_{ij} = a_{i+1,j+1}$. Alors $\det A = a_{11} \det B$.
9. $\det AB = \det BA = \det A \det B$.
10. A est inversible si et seulement si $\det A \neq 0$.

Projection orthogonale sur une droite

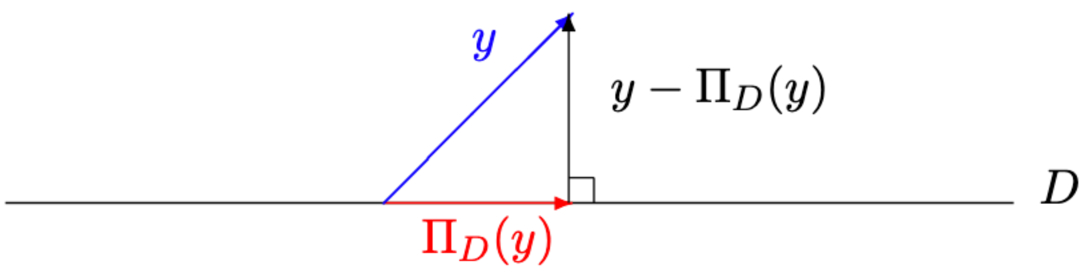
Soit x un vecteur non nul de E et soit D la droite vectorielle engendrée par x, i.e. $D = \{\lambda x \mid \lambda \in \mathbb{R}\}$. Soit $y \in E$. On veut projeter le vecteur y orthogonalement sur la droite D. On note $\Pi_D(y)$ la projection orthogonale de y sur D. Elle est caractérisée par les deux propriétés suivantes :

(i) $\Pi_D(y) \in D$, i.e. $\Pi_D(y)$ est colinéaire à x

(ii) $y - \Pi_D(y)$ est orthogonal à x.

On en déduit la solution :

$$\Pi_D(y) = \frac{< x, y >}{\|x\|^2} x \ .$$



Projection orthogonale sur un sous-espace vectoriel

Soit A une matrice de taille $m \times n$. Soit $A_1, \ldots, A_n$ ses colonnes. Ce sont des vecteurs de $\mathbb{R}^m$. Par exemple, si

$$A = \begin{pmatrix} 1 & 2 \\ 3 & -4 \\ 2 & 0 \end{pmatrix} ,$$

ses colonnes sont

$$A_1 = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix} , \quad A_2 = \begin{pmatrix} 2 \\ -4 \\ 0 \end{pmatrix}$$

On cherche à déterminer la projection orthogonale d'un vecteur y sur le sous espace vectoriel de $\mathbb{R}^m$ engendré par $A_1, \cdots, A_n$. Cette projection, notée Π_A est caractérisée par les propriétés

(i) $\Pi_A(y)$ est une combinaison linéaire des colonnes de A ;

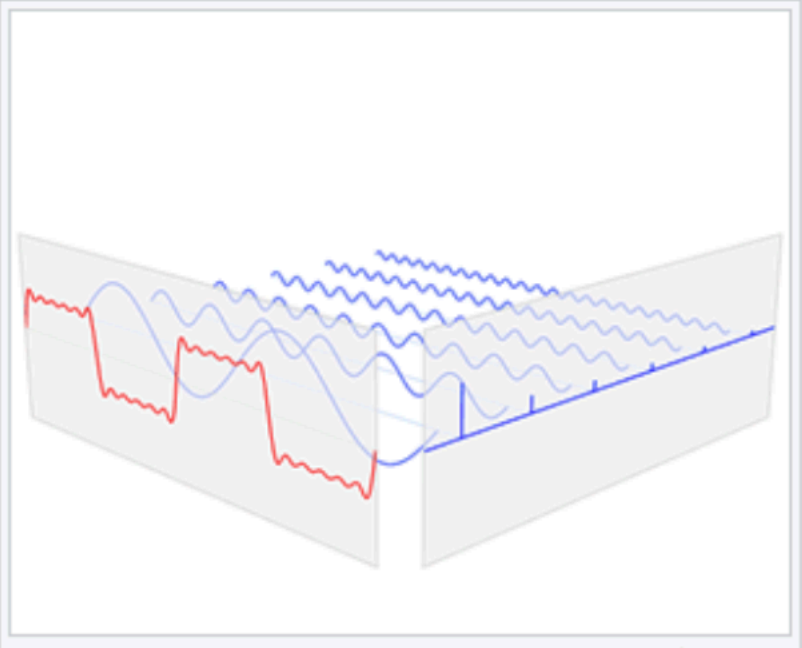
(ii) $y - \Pi_A(y)$ est orthogonal aux colonnes de A.

Projection orthogonale sur une base

Théorème *Soit A une matrice de taille $m \times n$ et soit $y \in \mathbb{R}^m$. Si la matrice ${}^tAA$ est inversible, alors la projection orthogonale de y sur le sous-espace de $\mathbb{R}^m$ engendré par les colonnes de A est*

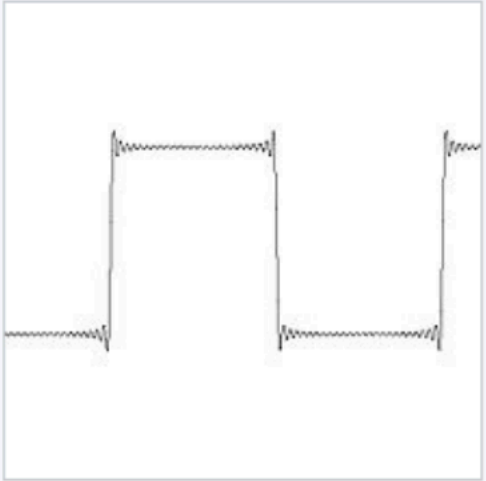
$$A\,({}^tAA)^{-1}\,{}^tA\,y\,.$$

2 : Projection sur une base - Illustrations

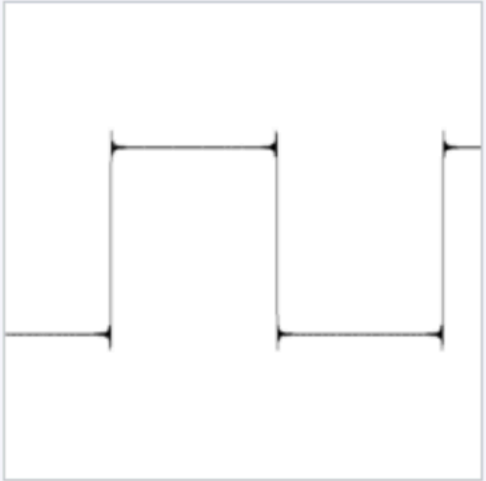


Projections sur une série de Fourier

Approximation du créneau à l'ordre 10

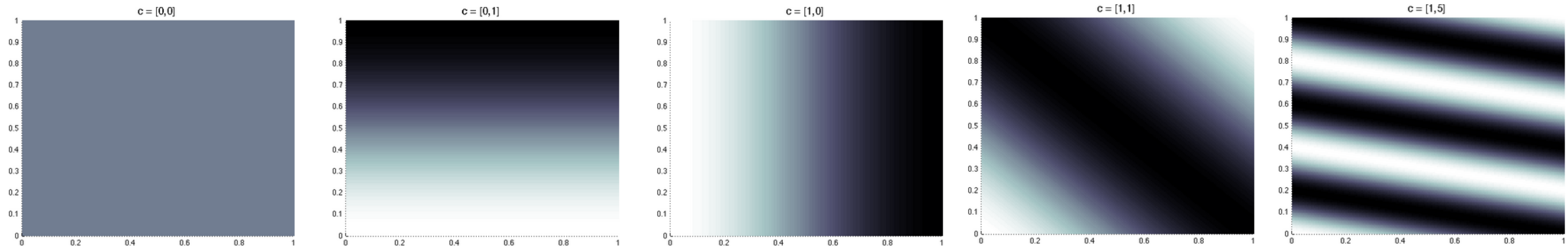


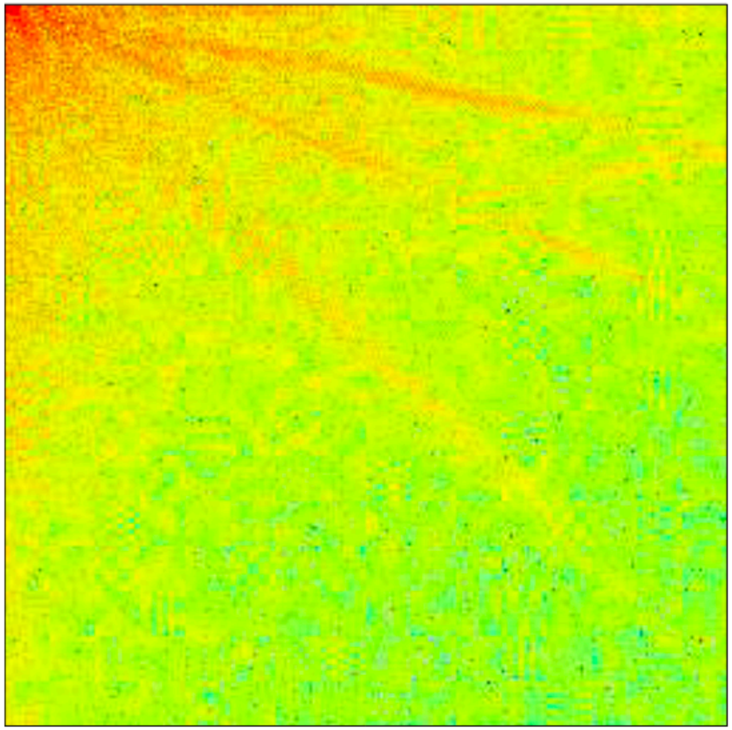
Approximation du créneau à l'ordre 50



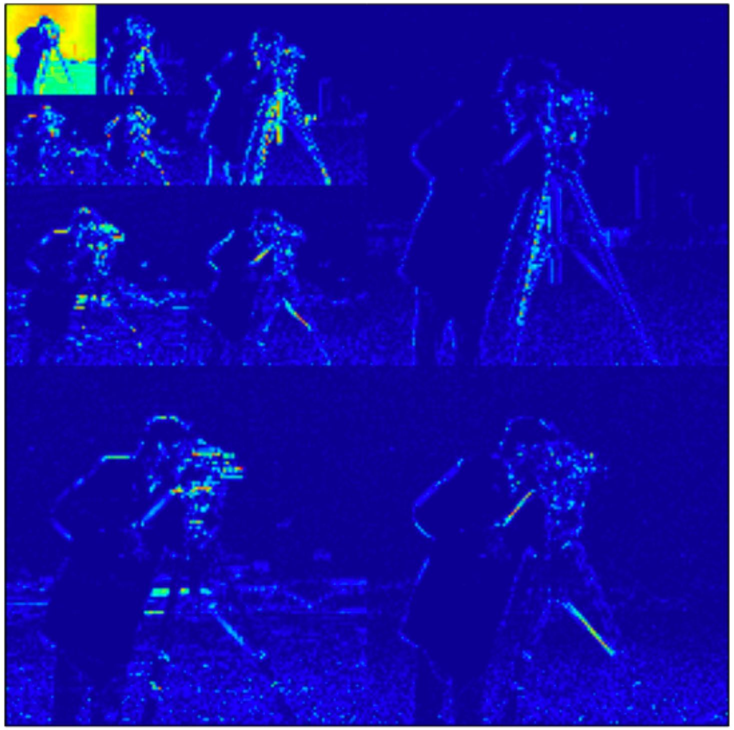
Approximation du créneau à l'ordre 250

Base de Fourier : $\varphi_j(s) = \cos(\pi c_j s)$, où c_j contrôle les propriétés de chaque projection





Base d'ondelettes: projections dépendant de l'échelle et de la localisation



2 : Projection sur une base - Illustrations

Transformations rigides d’images : projection des coordonnées des pixels à l’aide d’une matrice modélisant une rotation et une translation

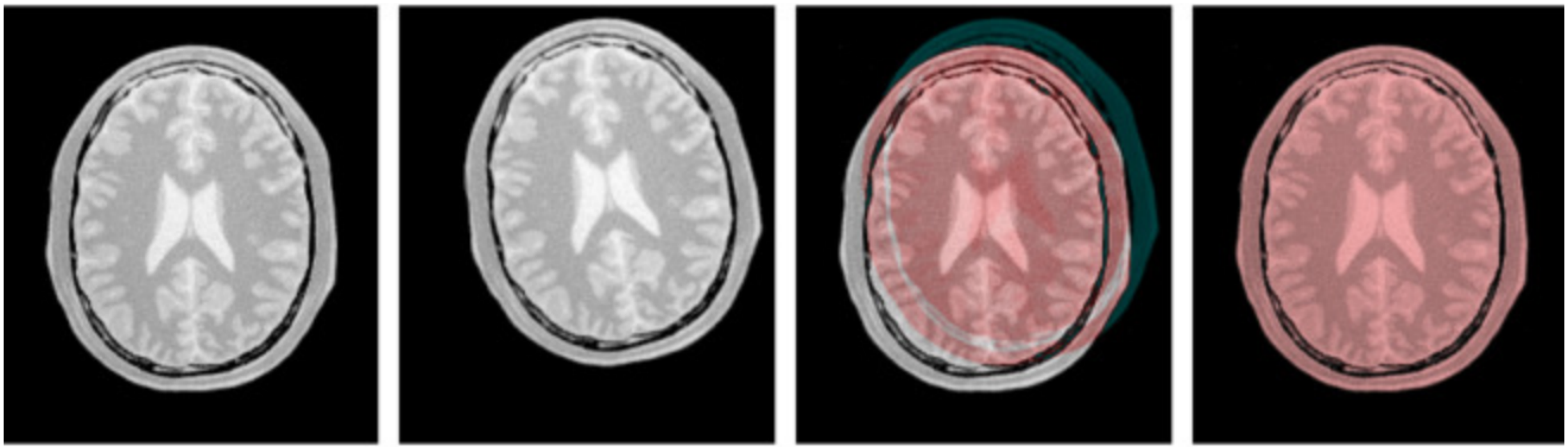


Image A | **Image B** | **Image A with overlaid Image B** | **Image A with overlaid Image B after translation-rotation**

Partie inspirée de : *Algèbre des matrices*. (Jean Fresnel, 2013, lien) et Cours de Samuel Bernard (CNRS)

Définition Une **décomposition en valeurs propres** d'une matrice carrée A est une factorisation

$$A = XDX^{-1},$$

avec X une matrice inversible et D une matrice **diagonale** ($D = [d_{ij}]$ avec $d_{ij} = 0$ si $i \neq j$).

Par définition de la multiplication de deux matrices, si x_i est la $i$ème colonne de X et λ_i la $i$ème entrée de D, alors

$$Ax_i = \lambda_i x_i. \tag{6}$$

Le vecteur x_i est donc un vecteur propre de A, et λ_i est sa valeur propre associée. La décomposition en valeurs propre est un changement de base vers les coordonnées des vecteurs propres. Si $Av = b$ et la matrice A se décompose $A = XDX^{-1}$, on a

$$D(X^{-1}v) = X^{-1}b \tag{7}$$

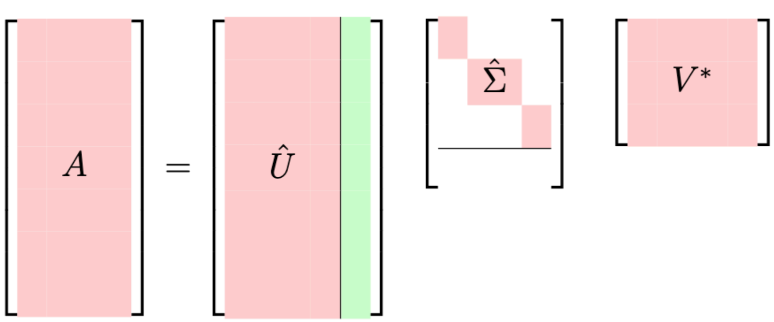
Les vecteurs $X^{-1}v$ et $X^{-1}b$ sont les vecteurs v et b exprimés dans la base des vecteurs propres que sont les colonnes de X. Dans cette base, la matrice A s'exprime comme une matrice diagonale D, qui agit sur chaque vecteur par multiplications scalaires. (La

4 : Décomposition en valeurs singulières (SVD)

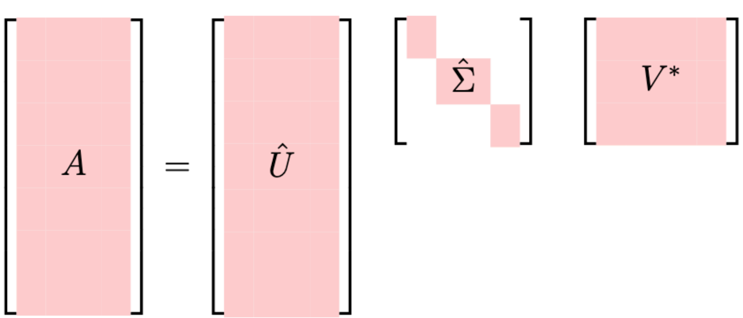
Soit A une matrice de taille $m \times n$

Définition Une **décomposition en valeurs singulières** (SVD – **Singular Value Decomposition**) est une factorisation $A = U\Sigma V^*$ où

- $U \in \mathbb{C}^{m\times m}$ est unitaire
- $V \in \mathbb{C}^{n\times n}$ est unitaire
- $\Sigma \in \mathbb{C}^{m\times n}$ est diagonale et les coefficients $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$, où $p = \min(m, n)$.



Décomposition complète



Décomposition réduite

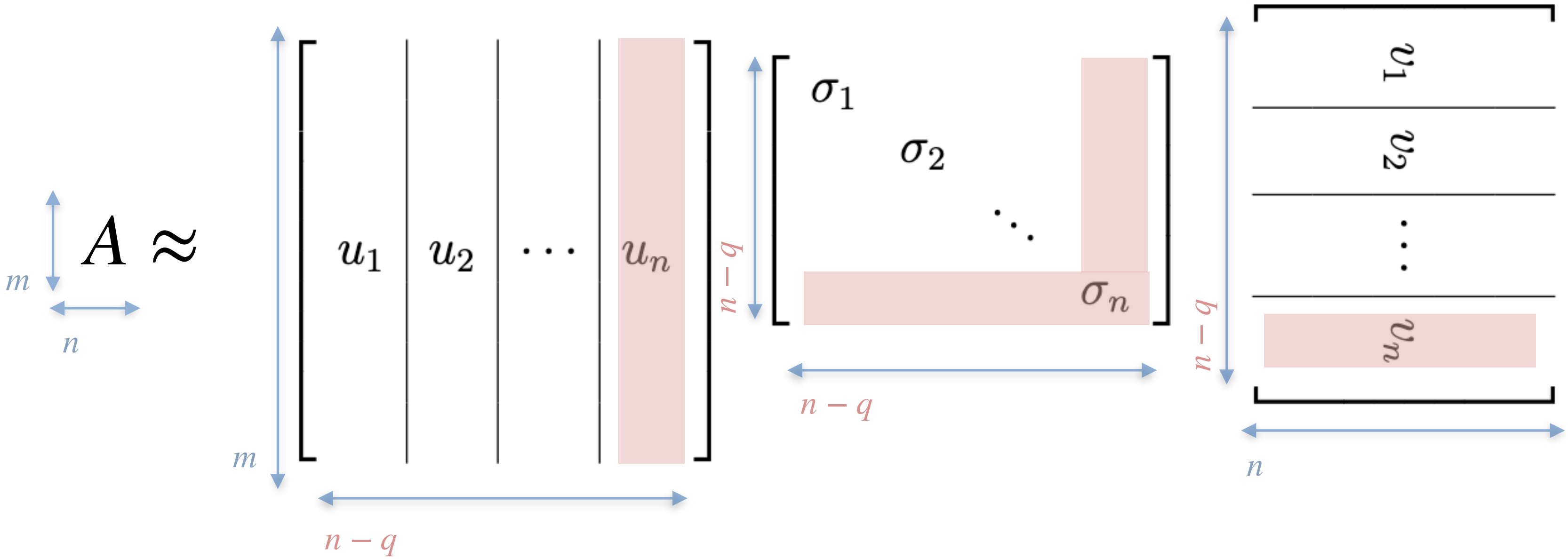
4 : Décomposition en valeurs singulières (SVD)

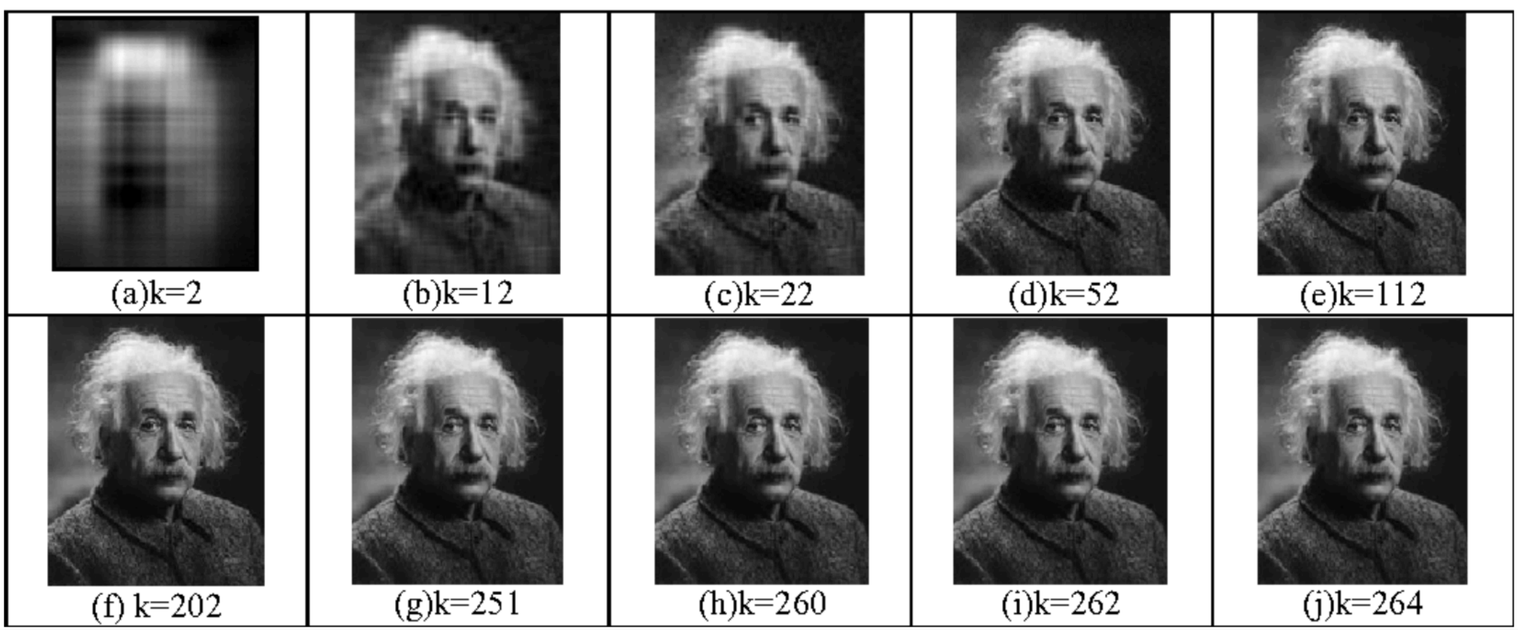
Interpretation

A une matrice de taille $m \times n$ $\rightarrow$ $A = U\Sigma V^*$

$$A \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix}$$

n vecteurs singuliers d'entrée de A $\qquad$ n vecteurs singuliers de sortie de A $\qquad$ n valeur singulières de A





(a)k=2 (b)k=12 (c)k=22 (d)k=52 (e)k=112

(f) k=202 (g)k=251 (h)k=260 (i)k=262 (j)k=264

4 : Décomposition en valeurs singulières (SVD)

Exemple issu de https://aaronschlegel.me/image-compression-singular-value-decomposition.html :

```
r <- lion[,,1]
g <- lion[,,2]
b <- lion[,,3]
```

```
lion.r.svd <- svd(r)
lion.g.svd <- svd(g)
lion.b.svd <- svd(b)

rgb.svds <- list(lion.r.svd, lion.g.svd, lion.b.svd)
```

```
for (j in seq.int(3, round(nrow(lion), -2), length.out = 8)) {
  a <- sapply(rgb.svds, function(i) {
    lion.compress <- i$u[,1:j] %*% diag(i$d[1:j]) %*% t(i$v[,1:j])
  }, simplify = 'array')
}
```

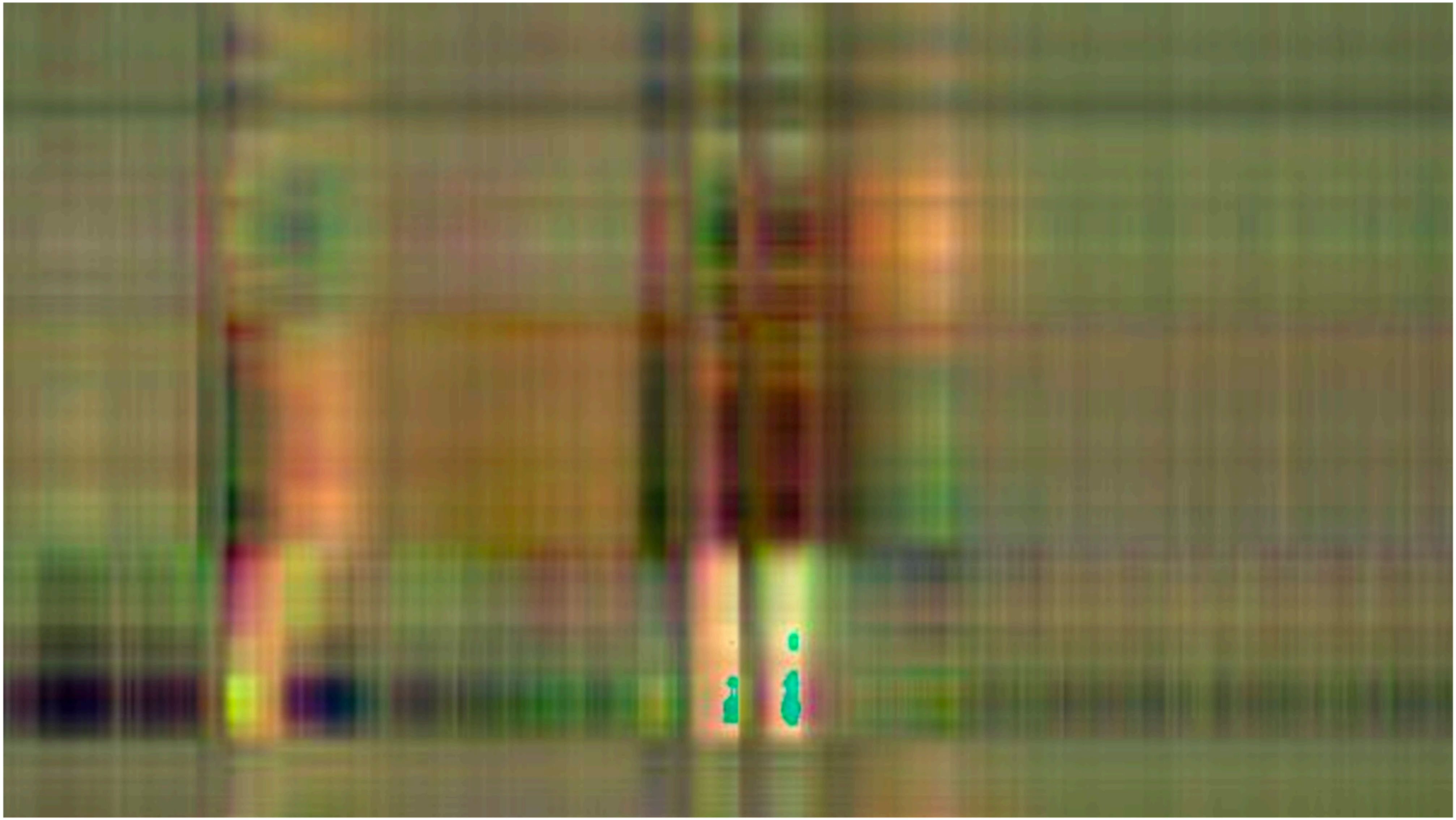
Image RGB 600x337

Rang 300

Rang 173

Rang 45

4 : Décomposition en valeurs singulières (SVD)



Rang 3

Théorème 3.1. existence et unicité de la SVD. *Toute matrice* $A \in \mathbb{C}^{n \times n}$ *possède une SVD. Les valeurs singulières* $\{\sigma_i\}$ *sont déterminées de façon unique. Si* A *est carrée* $(m = n)$ *et les valeurs singulières* σ_j *sont distinctes, les vecteurs d'entrée et de sorties* v_j *et* u_j *sont déterminés de façon unique à un facteur complexe unité près.*

Théorème 3.2. *Les valeurs singulières d'une matrice* A *sont les racines carrées des valeurs propres non-nulles de* A^*A *et* AA^*

Théorème 3.3. *Les colonnes de* U *sont les vecteurs propres orthogonaux de* AA^* *et les colonnes de* V *sont les vecteurs propres orthogonaux de* A^*A *à unité complexe près.*

5 : Matrices de covariances

Partie inspirée de :
- Cours de F. et M. Diener (U. Nice)
- Cours de L. Tournier (U. Paris 13)

Définition : Soient $x = (x_1, x_2, \ldots, x_n)$ et $y = (y_1, y_2, \ldots, y_n)$ deux échantillons de même taille n et de moyenne respective $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}$.
On appelle *covariance* (estimée[1]) de x avec y et on note $\mathrm{cov}\,(x,y)$ le nombre

$$\mathrm{cov}\,(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

On appelle *variance* (estimée) de x et on note $\mathrm{Var}\,(x)$ le nombre

$$\mathrm{Var}\,(x) = \mathrm{cov}\,(x,x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

et $\sqrt{\mathrm{Var}\,(x)}$ s'appelle *l'écart-type* de l'échantillon x. On appelle *corrélation* de x avec y et on note $\mathrm{cor}\,(x,y)$ le nombre

$$\rho = \mathrm{cor}\,(x,y) = \frac{cov(x,y)}{\sqrt{\mathrm{Var}\,(x)}\sqrt{\mathrm{Var}\,(y)}}.$$

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$. On note $X = {}^t\begin{pmatrix} X_1 & \cdots & X_d \end{pmatrix}$. Les variables aléatoires réelles $X_1, \ldots, X_d$ sont appelées les *composantes* de X et leurs lois sont les *lois marginales* de la loi de X. On définit classiquement son espérance (sa moyenne)

$$m_X = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

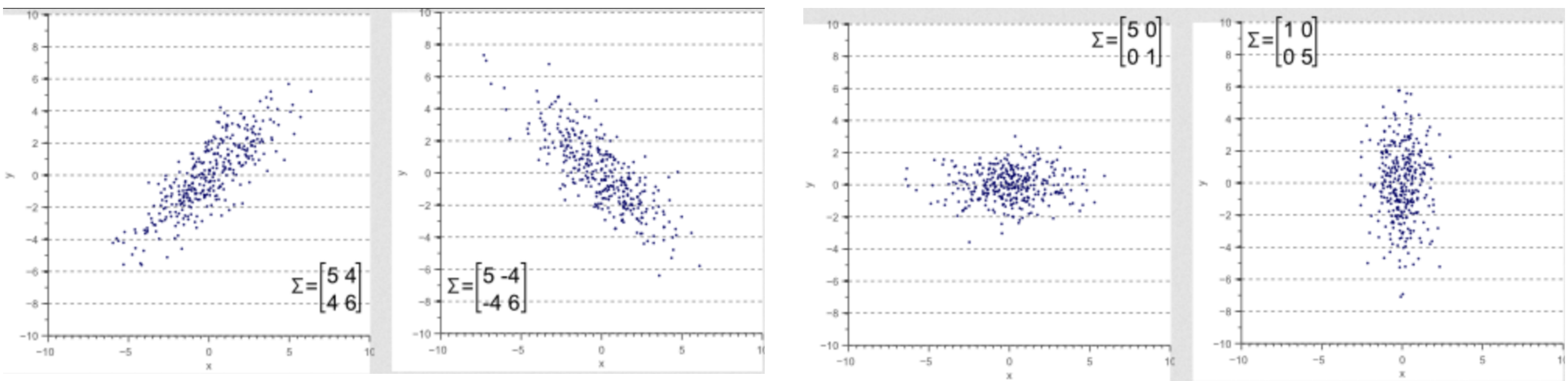
et aussi, si $X_1, \ldots, X_d$ sont de carré intégrable, sa variance

$$\begin{aligned} \Gamma_X = \mathrm{Var}(X) &= \mathbb{E}\big[(X - \mathbb{E}[X])^t(X - \mathbb{E}[X])\big] \\ &= (\mathrm{Cov}(X_i, X_j))_{1 \leq i,j \leq d} \\ &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Var}(X_d) \end{pmatrix}, \end{aligned}$$

En particulier, si $X_1, \ldots, X_d$ sont indépendantes deux à deux, alors Γ_X est une matrice diagonale ayant pour coefficients diagonaux $\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_d)$.
Comme $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ (ou comme ${}^t(A^tA) = A^tA$), la matrice Γ_X est symétrique. Elle est de plus positive : pour tout vecteur $u \in \mathbb{R}^d$,

$${}^t u \Gamma_X u = \sum_{i,j} u_i \,\mathrm{Cov}(X_i, X_j) u_j = \sum_{i,j} \mathrm{Cov}(u_i X_i, u_j X_j) = \mathrm{Var}(u_1 X_1 + \cdots + u_d X_d) \geq 0.$$

$$\begin{aligned}\Gamma_X = \mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^t(X - \mathbb{E}[X])\right] \\ &= (\mathrm{Cov}(X_i, X_j))_{1 \le i,j \le d} \\ &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Var}(X_d) \end{pmatrix},\end{aligned}$$



Théorèmes importants liés aux matrices de covariances :

Théorème 1 – *Théorème de Cochran.* Soit $X \sim \mathcal{N}(m, \sigma^2 I_d)$, et $\mathbb{R}^d = E_1 \overset{\perp}{\oplus} \dots \overset{\perp}{\oplus} E_r$ une décomposition de $\mathbb{R}^d$ en sous-espaces (affines) orthogonaux. Pour $k = 1, \dots, r$, on note d_k la dimension de E_k, et P_k la projection orthogonale sur E_k. Alors les variables aléatoires $Y_1 = P_1 X_1, \dots, Y_r = P_r X$ sont indépendantes, et

$$Y_k \sim \mathcal{N}(P_k m, \sigma^2 P_k) \quad \text{et} \quad \frac{1}{\sigma^2} \|Y_k - P_k m\|^2 \sim \chi^2_{d_k}.$$

Dans l'énoncé, χ^2_d est la loi du χ^2 (khi-deux) à d degrés de liberté, c'est-à-dire par définition la loi de la variable aléatoire $Z_1^2 + \cdots + Z_r^2$ où $Z_1, \dots, Z_r$ sont i.i.d. de loi $\mathcal{N}(0, 1)$.

Théorème 2 – *Théorème central limite multidimensionnel.* Soit $X_1, X_2, \dots$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, i.i.d., de carré intégrable, de moyenne m et de variance Γ. Alors

$$\frac{1}{\sqrt{n}} (X_1 + X_2 + \cdots + X_n - nm) \xrightarrow[n]{\text{(loi)}} \mathcal{N}(0, \Gamma).$$

Test de sphéricité de Bartlett : Teste l'indépendance globale des composantes d'un vecteur aléatoire

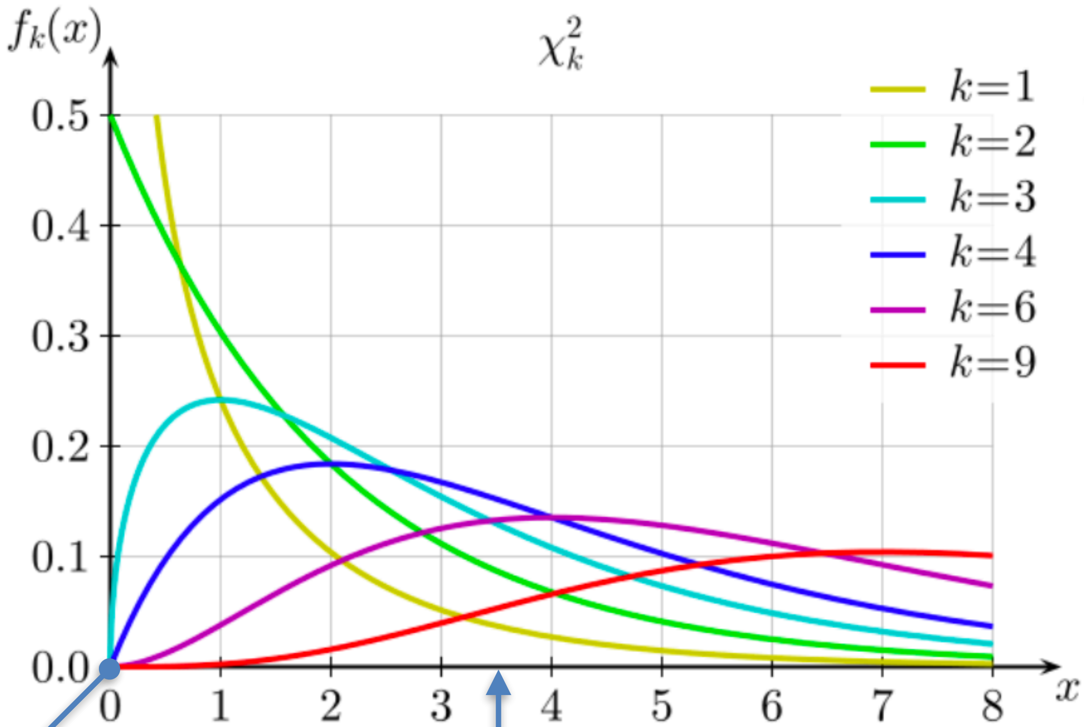
→ Echantillon de n réalisation de p v.a. réelles $X_1, X_2, \ldots, X_p$: $X = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ & & & \\ x_m^1 & x_m^2 & \cdots & x_m^p \end{pmatrix}$

→ Matrice de corrélation R (directement calculable à partir de $Var(X)$)

H_0 (hypothèse nulle) : les variables sont globalement indépendantes
H_1 : les variables sont globalement dépendantes

Le test évalue $t = -\left(n - 1 - \frac{2p+5}{6}\right) \log(|detR|)$ qui sous H_0 suit une loi du χ^2 à $k = \frac{p(p-1)}{2}$ degrés de libertés

$$\frac{p(p-1)}{2}$$



Dépendance de plus en plus grande entre les variables

Indépendance

$$t = -\left(n - 1 - \frac{2p+5}{6}\right) \log(|detR|)$$

MERCI !!!