

Fondements du Machine Learning

Clément W. Royer

Notes de cours - L3 IM2D - 2024/2025

- La dernière version de ce document est accessible à l'adresse :
<https://www.lamsade.dauphine.fr/~croyer/ensdocs/FML/PolyFML.pdf>.
- Pour toute remarque, envoyer un mail à clement.royer@lamsade.dauphine.fr.
Merci à Thibault De Surrel De Saint Julien pour sa relecture attentive.
- **Historique des versions du document**
 - 2024.10.14 : Corrections mineures.
 - 2024.10.12 : Ajout du chapitre 3.
 - 2024.09.10 : Modifications de l'exemple d'illustration.
 - 2024.09.09 : Ajout du chapitre 1.
 - 2024.09.03 : Première version avec contenu du premier cours.
- **Objectifs d'apprentissage**
À l'issue de ce cours, l'étudiant(e) sera capable de
 - Donner la formule de la décomposition en valeurs singulières, et appliquer cette décomposition à des problèmes matriciels;
 - Reconnaître et formuler des problèmes aux moindres carrés linéaires;
 - Donner des solutions de ces problèmes et expliciter leur lien avec la décomposition en valeurs singulières;
 - Donner la définition d'une composante principale, et appliquer l'analyse en composantes principales à des données matricielles.

Sommaire

0	Introduction	4
0.1	Derrière le Machine Learning...	4
0.2	Contexte et objectifs du cours	5
0.3	Notations	5
I	Réduction de dimension	7
1	Décomposition en valeurs singulières	8
1.1	Rappels d'algèbre linéaire	8
1.2	Valeurs propres et décomposition spectrale	11
1.3	Décomposition en valeurs singulières	11
1.3.1	Principe de la décomposition	12
1.3.2	Décomposition tronquée et approximation	13
2	Analyse en composantes principales	16
2.1	Motivation	16
2.2	Statistique empirique et pré-traitement des données	16
2.2.1	Individu moyen et données centrées	17
2.2.2	Dispersion et dépendance	17
2.3	Principe de l'analyse en composantes principales	19
2.3.1	Analyse en une composante principale	19
2.3.2	Analyse en plusieurs composantes principales	20
2.3.3	Application : Reconnaissance de visage	23
II	Régression linéaire	24
3	Premiers pas avec le modèle linéaire	25
3.1	Introduction	25
3.2	Résolution de systèmes non linéaires	26
3.2.1	Cas d'un système carré	26
3.2.2	Cas d'un système rectangulaire	27
3.2.3	Pseudo-inverse et SVD	28
3.3	Moindres carrés linéaires	29
3.3.1	Solution au sens des moindres carrés	29

3.3.2	Résolution du problème aux moindres carrés	31
3.4	Conclusion	31
4	Régression linéaire	32

Chapitre 0

Introduction

0.1 Derrière le Machine Learning...

Le terme *machine learning*, dont les traductions varient entre apprentissage machine, apprentissage automatique et apprentissage artificiel, fait partie d'un ensemble de mots-clés qui ont récemment gagné en popularité. Parmi ceux-ci, on trouve également l'analyse de données (*data analysis*), la fouille de données (*data mining*), l'intelligence artificielle (*artificial intelligence*, ou simplement *AI*), les masses de données (*Big Data*), etc. L'utilisation de cette terminologie est parfois hasardeuse : on leur préférera donc la notion de [sciences des données](#), ou *data science*.

La notion de donnée est en effet au coeur des différents concepts sus-mentionnés, et représente un enjeu majeur dans de nombreux secteurs d'activités. Pour les entreprises de service telles que les GAFAM ¹, il s'agit de fournir une valeur ajoutée dans leur service autrement gratuit via la façon dont les données des utilisateurs sont exploitées. En recherche et développement, la quantité massive de données générées dans certains domaines (biologie, médecine) pose d'importants défis mathématiques et informatiques. Plus globalement, les approches guidées par les données (*data-driven*) deviennent de plus en plus populaires, car elles permettent de pallier le manque de modèles formels ou implémentables. C'est le cas par exemple pour la modélisation météorologique à grande échelle : nos capacités de calcul ne nous permettent pas de faire évoluer un modèle de prédiction à l'échelle du globe, mais il est possible de collecter un grand volume de données et d'en extraire les tendances majeures.

Dans ce cours, on considèrera deux approches d'analyse de données. La première approche, dite prédictive, ne pré-supposera pas de distribution sur les données, et visera à extraire de l'information des données même (on parle ainsi d'apprentissage non supervisé). L'[analyse en composantes principales](#) (ou ACP, voir Chapitre 2) sera ici l'outil-clé pour obtenir cette information. La seconde, dite fonctionnelle, supposera que les données d'apprentissage suivent une distribution de probabilité connue : il est ainsi possible de construire un modèle de ces données adapté à la distribution sous-jacente. Les approches de [régression linéaire](#) seront utilisés dans ce contexte d'apprentissage (dit supervisé).

¹Google, Apple, Facebook, Amazon et Microsoft.

0.2 Contexte et objectifs du cours

Dans ce cours, on s'intéressera à développer des techniques visant à extraire de l'information d'un jeu de données. On considèrera que l'on dispose d'une quantité importante de données, non seulement pour qu'il soit intéressant d'en extraire de l'information, mais aussi pour que ces données puissent représenter des tendances. Les techniques que nous emploierons reposent sur des algorithmes, c'est-à-dire des traitements systématiques à appliquer aux données. Comme on le verra, le développement d'un algorithme efficace repose à la fois sur des arguments mathématiques et sur une implémentation bien pensée.

Ce cours se concentre plus spécifiquement sur l'obtention de modèles **linéaires** des relations entre les données; ces données seront de plus vues comme des réalisations de variables aléatoires (généralement gaussiennes). Ce choix se justifie par la pertinence et l'efficacité de ces modèles simples dans la pratique. Il permet également d'utiliser des résultats et algorithmes issus de l'algèbre linéaire, de l'optimisation et des statistiques.

0.3 Notations

Conventions de notation

- Les scalaires seront représentés par des lettres minuscules : $a, b, c, \alpha, \beta, \gamma$.
- Les vecteurs seront représentés par des lettres minuscules **en gras** : $\mathbf{a}, \mathbf{b}, \mathbf{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$.
- Les lettres majuscules en gras seront utilisées pour les matrices : $\mathbf{A}, \mathbf{B}, \mathbf{C}$.
- En l'absence d'ambiguïté, on pourra omettre les indices de début et de fin dans une somme finie afin d'alléger les notations. On pourra de même utiliser un seul symbole de notation pour plusieurs indices et ainsi écrire de manière équivalente $\sum_{i=1}^m \sum_{j=1}^n$, $\sum_i \sum_j$ ou $\sum_{i,j}$ si le contexte le permet.

Notations vectorielles

- On notera \mathbb{R}^n l'ensemble des vecteurs à n composantes réelles, et on considèrera toujours que n est un entier supérieur ou égal à 1.
- Un vecteur $\mathbf{x} \in \mathbb{R}^n$ sera pensé (par convention) comme un vecteur colonne. On notera $x_i \in \mathbb{R}$ sa i -ème coordonnée dans la base canonique de \mathbb{R}^n . On aura ainsi $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, que l'on notera plus succinctement $\mathbf{x} = [x_i]_{1 \leq i \leq n}$.
- Étant donné un vecteur (colonne) $\mathbf{x} \in \mathbb{R}^n$, le vecteur ligne correspondant sera noté \mathbf{x}^T . On aura donc $\mathbf{x}^T = [x_1 \ \cdots \ x_n]$ et $(\mathbf{x}^T)^T = \mathbf{x}$.
- Pour tout $n \geq 1$, les vecteurs $\mathbf{0}_n$ et $\mathbf{1}_n$ représentent les vecteurs colonnes de \mathbb{R}^n dont tous les éléments sont égaux à 0 ou 1, respectivement.

Notations matricielles

- On notera $\mathbb{R}^{m \times n}$ l'ensemble des matrices à m lignes et n colonnes à coefficients réels, où m et n seront des entiers supérieurs ou égaux à 1. Les espaces $\mathbb{R}^{m \times 1}$ et \mathbb{R}^m étant isomorphes (ce que l'on note $\mathbb{R}^{m \times 1} \simeq \mathbb{R}^m$), on pourra considérer un vecteur de \mathbb{R}^m comme une matrice de $\mathbb{R}^{m \times 1}$, et vice versa. Une matrice $\mathbf{A} \in \mathbb{R}^{n \times n}$ est dite carrée (dans le cas général, on parlera de matrice rectangulaire).
- Étant donnée une matrice $\mathbf{A} \in \mathbb{R}^n$, on notera \mathbf{A}_{ij} le coefficient en ligne i et colonne j de la matrice. La notation $[\mathbf{A}_{ij}]_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}}$ sera donc équivalente à \mathbf{A} . Sans ambiguïté sur la taille de la matrice, on notera simplement $[\mathbf{A}_{ij}]$.

- Selon les besoins, on utilisera \mathbf{a}_i^T pour la i -ème ligne de \mathbf{A} ou \mathbf{a}_j pour la j -ème colonne de \mathbf{A} .
Selon le cas, on aura donc $\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}$ ou $\mathbf{A} = [\mathbf{a}_1 \ \cdots \ \mathbf{a}_n]$.

- Pour une matrice $\mathbf{A} = [\mathbf{A}_{ij}] \in \mathbb{R}^{m \times n}$, la matrice transposée de \mathbf{A} , notée \mathbf{A}^T , est la matrice de $\mathbb{R}^{n \times m}$ telle que

$$\forall i = 1 \dots m, \forall j = 1 \dots n, \quad \mathbf{A}_{ji}^T = \mathbf{A}_{ij}.$$

Nota Bene : Cette notation généralise donc la correspondance entre vecteurs lignes et vecteurs colonnes.

- Pour tout $n \geq 1$, la matrice \mathbf{I}_n représentera la matrice identité de $\mathbb{R}^{n \times n}$ (avec des 1 sur la diagonale et des 0 partout ailleurs), tandis que les matrices $\mathbf{0}_n$ et $\mathbf{1}_n$ représenteront les matrices dont tous les éléments sont égaux à 0 ou 1, respectivement. De manière plus générale, les notations $\mathbf{0}_{m,n}$ et $\mathbf{1}_{m,n}$ seront utilisées pour les matrices de $\mathbb{R}^{m \times n}$ ne contenant respectivement que des 0 et des 1.

Partie I

Réduction de dimension

Chapitre 1

Décomposition en valeurs singulières

1.1 Rappels d'algèbre linéaire

On considèrera toujours l'espace des vecteurs \mathbb{R}^n muni de sa structure d'espace vectoriel normé de dimension n :

- Pour tous $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, la somme des vecteurs \mathbf{x} et \mathbf{y} est notée $\mathbf{x} + \mathbf{y} = [x_i + y_i]_{1 \leq i \leq n}$;
- Pour tout $\lambda \in \mathbb{R}$, on définit $\lambda \mathbf{x} := \lambda \cdot \mathbf{x} = [\lambda x_i]_{1 \leq i \leq n}$.
- La norme euclidienne $\|\cdot\|$ sur \mathbb{R}^n est définie pour tout vecteur $\mathbf{x} \in \mathbb{R}^n$ par

$$\|\mathbf{x}\| := \sqrt{\sum_{i=1}^n x_i^2}.$$

On dira que $\mathbf{x} \in \mathbb{R}^n$ est unitaire si $\|\mathbf{x}\| = 1$.

- Pour tous vecteurs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, le produit scalaire dérivé de la norme euclidienne est noté $\mathbf{x}^T \mathbf{y}$, et défini par

$$\mathbf{x}^T \mathbf{y} := \sum_{i=1}^n x_i y_i.$$

Il s'agit d'une forme bilinéaire symétrique définie positive. On a en particulier $\mathbf{y}^T \mathbf{x} = \mathbf{x}^T \mathbf{y}$.

- Il existe une famille libre et génératrice de \mathbb{R}^n de taille n . Par exemple, tout vecteur \mathbf{x} de \mathbb{R}^n s'écrit $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$, où $\mathbf{e}_i = [0 \cdots 0 \ 1 \ 0 \cdots 0]^T$ est le i -ème vecteur de la base canonique (le coefficient 1 se trouvant en i -ème position).

Définition 1.1 (Sous-espace engendré) Soient $\mathbf{x}_1, \dots, \mathbf{x}_p$ p vecteurs de \mathbb{R}^n . Le *sous-espace engendré* par les vecteurs $\mathbf{x}_1, \dots, \mathbf{x}_p$ est le sous-espace vectoriel

$$\text{vect}(\mathbf{x}_1, \dots, \mathbf{x}_p) := \left\{ \mathbf{x} = \sum_{i=1}^p \alpha_i \mathbf{x}_i \mid \alpha_i \in \mathbb{R} \ \forall i \right\}.$$

Ce sous-espace est de dimension au plus $\min(n, p)$.

Lorsque l'on travaille avec des matrices, on s'intéresse généralement aux sous-espaces définis ci-dessous.

Définition 1.2 (Sous-espaces matriciels) Soit une matrice $A \in \mathbb{R}^{m \times n}$, on définit les deux sous-espaces suivants :

- Le **noyau** (kernel/null space en anglais) de A est le sous-espace vectoriel

$$\ker(A) := \{x \in \mathbb{R}^n \mid Ax = 0_m\}$$

- L'**image** (range space) de A est le sous-espace vectoriel

$$\text{Im}(A) := \{y \in \mathbb{R}^m \mid \exists x \in \mathbb{R}^n, y = Ax\}$$

La dimension de ce sous-espace vectoriel s'appelle le **rang** de A . On la note $\text{rang}(A)$ et on a $\text{rang}(A) \leq \min\{m, n\}$.

Théorème 1.1 (Théorème du rang) Pour toute matrice $A \in \mathbb{R}^{m \times n}$, on a

$$\dim(\ker(A)) + \text{rang}(A) = n.$$

Définition 1.3 (Normes matricielles) On définit sur $\mathbb{R}^{m \times n}$ la norme d'opérateur $\|\cdot\|$ et la norme de Frobenius $\|\cdot\|_F$ par

$$\forall A \in \mathbb{R}^{m \times n}, \begin{cases} \|A\| &:= \max_{\substack{x \in \mathbb{R}^n \\ x \neq 0_n}} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\| \\ \|A\|_F &:= \sqrt{\sum_{\substack{1 \leq i \leq m \\ 1 \leq j \leq n}} A_{ij}^2}. \end{cases}$$

Nous terminons cette section par quelques définitions de sous-ensembles de matrices carrées qui nous seront utiles dans le cours.

Définition 1.4 (Matrice symétrique) Une matrice carrée $A \in \mathbb{R}^{n \times n}$ est dite **symétrique** si elle vérifie $A^T = A$.

Définition 1.5 (Matrice inversible) Une matrice carrée $A \in \mathbb{R}^{n \times n}$ est dite **inversible** s'il existe $B \in \mathbb{R}^{n \times n}$ telle que $BA = AB = I_n$ (où l'on rappelle que I_n désigne la matrice identité de $\mathbb{R}^{n \times n}$).

Si elle existe, une telle matrice B est unique : elle est appelée **l'inverse de A** et on la note A^{-1} .

Définition 1.6 (Matrice (semi-)définie positive) Une matrice carrée $A \in \mathbb{R}^{n \times n}$ est dite **semi-définie positive** si elle est symétrique et que

$$\forall x \in \mathbb{R}^n, \quad x^T A x \geq 0.$$

Elle est dite **définie positive** si elle est semi-définie positive et que $x^T A x > 0$ pour tout vecteur x non nul.

Définition 1.7 (Matrice orthogonale) Une matrice carrée $P \in \mathbb{R}^{n \times n}$ est dite *orthogonale* si $P^T = P^{-1}$.

Par extension, on dira que $Q \in \mathbb{R}^{m \times n}$ avec $m \leq n$ est orthogonale si $QQ^T = I_m$ (les colonnes de Q sont donc orthonormées dans \mathbb{R}^m).

Si $Q \in \mathbb{R}^{n \times n}$ est une matrice orthogonale, alors Q^T est également orthogonale.

On utilisera fréquemment la propriété des matrices orthogonales énoncée ci-dessous.

Lemme 1.1 Soit une matrice $A \in \mathbb{R}^{m \times n}$ et $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ des matrices orthogonales (respectivement de $\mathbb{R}^{m \times m}$ et $\mathbb{R}^{n \times n}$). On a

$$\|A\| = \|UA\| = \|AV\| \text{ et } \|A\|_F = \|UA\|_F = \|AV\|_F,$$

c'est-à-dire que la multiplication par une matrice orthogonale ne modifie pas la norme d'opérateur.

Démonstration. On montre tout d'abord que pour tout vecteur $x \in \mathbb{R}^m$, on a $\|Ux\| = \|x\|$. En utilisant la définition de la norme et celle d'une matrice orthogonale, on a :

$$\|Ux\|^2 = x^T U^T U x = x^T x = \|x\|^2,$$

ce qui établit le résultat. Par conséquent, on a également

$$\|UAx\| = \|Ax\|$$

pour tout vecteur x . En revenant à la définition de la norme d'opérateur, on obtient ainsi

$$\|UA\| = \max_{\|x\|=1} \|UAx\| = \max_{\|x\|=1} \|Ax\| = \|A\|,$$

ce qui est bien le résultat recherché. Pour le résultat sur $\|AV\|$, on note que V^T est une matrice orthogonale inversible avec $VV^T = I_n$, donc que pour tout x , il existe z tel que $x = V^T z$ et $\|x\| = \|z\|$ d'après ce qui précède. On a ainsi :

$$\begin{aligned} \|AV\| &= \max_{\|x\|=1} \|AVx\| = \max_{\|V^T z\|=1} \|AVV^T z\| \\ &= \max_{\|z\|=1} \|Az\| = \|A\|, \end{aligned}$$

ce qu'il fallait démontrer.

On démontre maintenant le résultat pour la norme de Frobenius, dont on rappelle qu'elle est définie par $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$. On a ainsi

$$\|A\|_F = \sqrt{\|a_{1\bullet}\|^2 + \dots + \|a_{m\bullet}\|^2} = \sqrt{\|a_{\bullet 1}\|^2 + \dots + \|a_{\bullet n}\|^2}$$

où $a_{1\bullet}, \dots, a_{m\bullet}$ et $a_{\bullet 1}, \dots, a_{\bullet n}$ représentent les lignes et les colonnes de A , respectivement. Ainsi, la norme de Frobenius d'une matrice au carré est égale à la somme des carrés des normes de ses colonnes ou de ses lignes). Comme on a montré que la norme d'un vecteur ne change pas par transformation orthogonale, pour toute matrice $U \in \mathbb{R}^{m \times m}$ orthogonale, on a

$$\sqrt{\|Ua_{\bullet 1}\|^2 + \dots + \|Ua_{\bullet n}\|^2} = \sqrt{\|a_{\bullet 1}\|^2 + \dots + \|a_{\bullet n}\|^2},$$

d'où $\|UA\|_F = \|A\|_F$. En considérant les lignes de A , on montre de même que $\|A\|_F = \|AV\|_F$. \square

Par corollaire immédiat du lemme précédent, on note qu'une matrice $Q \in \mathbb{R}^{m \times n}$ orthogonale avec $m \leq n$ vérifie nécessairement $\|Q\| = 1$ et $\|Q\|_F = \sqrt{m}$.

1.2 Valeurs propres et décomposition spectrale

Définition 1.8 (Valeur propre) Soit une matrice $A \in \mathbb{R}^{n \times n}$. On dit que $\lambda \in \mathbb{R}$ est une **valeur propre de A** si

$$\exists v \in \mathbb{R}^n, v \neq 0_n, \quad Av = \lambda v.$$

Le vecteur v est appelé un **vecteur propre associé à la valeur propre λ** . L'ensemble des valeurs propres de A s'appelle le **spectre de A** .

Le sous-espace engendré par les vecteurs propres associés à la même valeur propre d'une matrice s'appelle un sous-espace propre. Sa dimension correspond à l'ordre de multiplicité de la valeur propre relativement à la matrice.

Proposition 1.1 Pour toute matrice $A \in \mathbb{R}^{n \times n}$, on a les propriétés suivantes :

- La matrice A possède n valeurs propres complexes mais pas nécessairement réelles.
- Si la matrice A est semi-définie positive (respectivement définie positive), alors ses valeurs propres sont réelles positives (respectivement strictement positives).
- Le noyau de A est engendré par les vecteurs propres associés à la valeur propre 0.

Théorème 1.2 (Théorème spectral) Toute matrice carrée $A \in \mathbb{R}^{n \times n}$ symétrique admet une décomposition dite **spectrale** de la forme :

$$A = P \Lambda P^{-1},$$

où $P \in \mathbb{R}^{n \times n}$ est une matrice orthogonale, dont les colonnes p_1, \dots, p_n forment une base orthonormée de vecteurs propres, et $\Lambda \in \mathbb{R}^{n \times n}$ est une matrice diagonale qui contient les n valeurs propres de A $\lambda_1, \dots, \lambda_n$ sur la diagonale.

Il est à noter que la décomposition spectrale n'est pas unique. En revanche, l'ensemble des valeurs propres est unique, que l'on prenne en compte les ordres de multiplicité ou non.

La décomposition spectrale définie dans le théorème 1.2 est particulièrement importante car elle permet de synthétiser l'information de A par son effet sur les vecteurs p_i . Ainsi, lorsque $|\lambda_i| \gg 1$, on aura $\|Ap_i\| \gg \|p_i\|$, et la matrice aura donc un effet expansif dans la direction de p_i (ou sa direction opposée lorsque $\lambda_i < 0$). De même, si $|\lambda_i| \ll 1$, la matrice aura un effet contractant dans la direction de p_i : le cas extrême est $\lambda_i = 0$, c'est-à-dire que $p_i \in \ker(A)$ et la matrice ne conserve donc pas d'information relative à p_i .

Géométriquement parlant, on voit ainsi que, pour tout vecteur $x \in \mathbb{R}^n$ décomposé dans la base des p_i que l'on multiplie par A , les composantes de ce vecteur associées aux plus grandes valeurs propres¹ seront augmentées, tandis que celles associées aux valeurs propres de petite magnitude seront réduites (voire annihilées dans le cas d'une valeur propre nulle).

1.3 Décomposition en valeurs singulières

La décomposition en valeurs singulières (ou SVD, pour *Singular Value Decomposition*) est une technique fondamentale en analyse et compression de données, particulièrement utile pour compresser des signaux audio, des images, etc.

¹On parle ici de plus grandes valeurs propres en valeur absolue, ou magnitude.

1.3.1 Principe de la décomposition

Soit une matrice rectangulaire $A \in \mathbb{R}^{m \times n}$: dans le cas général, les dimensions de la matrice diffèrent, et on ne peut donc pas parler de valeurs propres de la matrice A . On peut en revanche considérer les deux matrices

$$A^T A \in \mathbb{R}^{n \times n} \quad \text{et} \quad A A^T \in \mathbb{R}^{m \times m}.$$

Ces matrices sont symétriques réelles, et par conséquent diagonalisables. Par ailleurs, elles sont fortement liées à la matrice A . Le lemme ci-dessous illustre quelques-unes des propriétés de $A^T A$; des résultats similaires peuvent être démontrés pour $A A^T$.

Lemme 1.2 *Pour toute matrice $A \in \mathbb{R}^{m \times n}$, les propriétés suivantes sont vérifiées :*

- i) $A^T A$ est semi-définie positive;
- ii) $A^T A$ est symétrique.
- iii) $\ker(A^T A) = \ker(A)$;
- iv) $\text{Im}(A^T A) = \text{Im}(A^T)$;
- v) $\text{rang}(A^T A) = \text{rang}(A)$.

Ces résultats sont à la base de la construction de la décomposition en valeurs singulières, dont on donne l'énoncé ci-dessous.

Théorème 1.3 (Décomposition en valeurs singulières) *Toute matrice $A \in \mathbb{R}^{m \times n}$ admet une décomposition en valeurs singulières (SVD²) de la forme*

$$A = U \Sigma V^T,$$

où $U \in \mathbb{R}^{m \times m}$ est orthogonale ($U^T U = I_m$), $V \in \mathbb{R}^{n \times n}$ est orthogonale ($V^T V = I_n$) et $\Sigma \in \mathbb{R}^{m \times n}$ est telle que $\Sigma_{ij} = 0$ si $i \neq j$ et $\Sigma_{ii} \geq 0$.

L'ensemble des valeurs $\{\Sigma_{ii}\}$ pour $1 \leq i \leq \min\{m, n\}$, noté $\{\sigma_1, \dots, \sigma_{\min\{m, n\}}\}$ est appelé l'ensemble des valeurs singulières de la matrice A . Les colonnes de V (resp. de U) sont appelées les vecteurs singuliers à droite (resp. à gauche) de A .

Remarque 1.1 *Comme dans le cas de la décomposition en valeurs propres, il n'y a pas unicité de la décomposition en valeurs singulières, mais il y a unicité de l'ensemble des valeurs singulières.*

Exemple 1.1 *La décomposition en valeurs singulières d'une matrice de $\mathbb{R}^{3 \times 2}$ est de la forme*

$$A = \underbrace{[u_1 \ u_2 \ u_3]}_U \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}}_{V^T}$$

où $\sigma_1 \geq 0, \sigma_2 \geq 0$, les u_i forment une base orthonormée de \mathbb{R}^3 et les v_i forment une base orthonormée de \mathbb{R}^2 .

²Dans la suite, on utilisera fréquemment l'algorithme anglo-saxon SVD pour faire référence à la décomposition en valeurs singulières.

Proposition 1.2 Soit $A \in \mathbb{R}^{m \times n}$ et $U\Sigma V^T$ une décomposition en valeurs singulières de A . Alors :

- i) Les carrés des valeurs singulières sont les valeurs propres de $A^T A$.
- ii) Si $\text{rang}(A) = r$, alors il y a exactement r valeurs singulières non nulles.

Remarque 1.2 Une preuve constructive de la décomposition en valeurs singulières sera réalisée en TD (exercice 1.5). Si cette preuve permet de construire une décomposition en valeurs singulières, elle n'est pas forcément aisée à utiliser en pratique de par son coût en termes d'opérations algébriques.

Les implémentations modernes de la décomposition en valeurs singulières reposent sur des techniques d'algèbre linéaire (QR avec pivotage, factorisation symétrique). La version la plus utilisée, basée sur l'algorithme de Golub et Kahan [3], possède de très bonnes garanties de stabilité numérique, ce qui est fondamental pour une implémentation efficace. Cela explique en partie le succès de la décomposition en valeurs singulières et son utilisation très répandue dans de nombreuses applications.

1.3.2 Décomposition tronquée et approximation

Le principal intérêt de la décomposition en valeurs singulières est de permettre de compresser la représentation de données matricielles. Dans de nombreuses applications, il est fréquent que les matrices de données présentent peu de valeurs singulières élevées, et beaucoup de petites valeurs singulières. On peut alors se demander quelle est la perte d'information que l'on réalise en omettant ces valeurs singulières.

La première réduction d'information que l'on peut opérer consiste à éliminer les valeurs singulières nulles dans la représentation de la matrice. C'est le sens du résultat ci-dessous.

Théorème 1.4 (SVD réduite) Toute matrice $A \in \mathbb{R}^{m \times n}$ de rang r admet une SVD réduite de la forme

$$A = U\Sigma V^T, \quad (1.3.1)$$

où $U \in \mathbb{R}^{m \times r}$ avec $U^T U = I$, $V \in \mathbb{R}^{n \times r}$ avec $V^T V = I_r$ et $\Sigma \in \mathbb{R}^{r \times r}$ est diagonale à coefficients diagonaux strictement positifs.

La décomposition (1.3.1) est plus compacte que la décomposition originelle. En particulier, il suffit de stocker $(n+m)r$ réels³ pour pouvoir reconstruire la matrice, ce qui peut être plus avantageux que de stocker les mn coefficients de la matrice A .

Définition 1.9 (SVD tronquée) Soit $A \in \mathbb{R}^{m \times n}$ une matrice de rang r et $U\Sigma V^T$ sa SVD réduite, avec

$$U = [u_1 \cdots u_r], \quad V = [v_1 \cdots v_r], \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_r \end{bmatrix}.$$

On suppose que $\sigma_1 \geq \cdots \geq \sigma_r$. Alors, pour tout $k \leq r$, la décomposition $U_k \Sigma_{k,k} V_k^T$, où

$$U_k = [u_1 \cdots u_k], \quad V_k = [v_1 \cdots v_k], \quad \Sigma_{k,k} = \begin{bmatrix} \sigma_1 & 0 \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_k \end{bmatrix}$$

s'appelle la *décomposition en valeurs singulières tronquée à k valeurs*, ou *k -SVD*.

³Le coût est de $(n+m+1)*r$ si on stocke les valeurs singulières séparément, mais on peut également les incorporer dans U ou V , auquel cas le coût de stockage sera de $(n+m)*r$.

On remarque que cette factorisation est encore moins coûteuse en terme de coefficients que la décomposition en valeurs singulière réduite. Contrairement à celle-ci, la k -SVD supprime de l'information issue de la matrice \mathbf{A} dès lors que $k < r$: tout l'enjeu du processus de troncature consiste à préserver la majeure partie de l'information contenue dans la matrice, c'est-à-dire les valeurs singulières les plus importantes.

Application en compression d'images On considère une image 307x241 pixels stockée sous la forme d'une matrice $\mathbf{A} \in \mathbb{R}^{m \times n}$. Si on calcule la décomposition en valeurs singulières de \mathbf{A} , on peut voir que le rang de la matrice est $n = 241$.



Figure 1.1: Image 307x241 pixels; la matrice correspondante est de rang 241.

On peut cependant se demander si toutes les valeurs singulières sont nécessaires pour encoder l'image. On considère donc plusieurs troncatures de la *SVD*, correspondant à différentes valeurs de k inférieures au rang véritable. Les résultats de la figure 1.2 montrent qu'il n'est pas nécessaire de considérer la décomposition complète pour obtenir une image nette (voire indiscernable de l'image d'origine à l'oeil nu). Par ailleurs, on notera qu'une *SVD* tronquée peut avoir un coût mémoire supérieur à

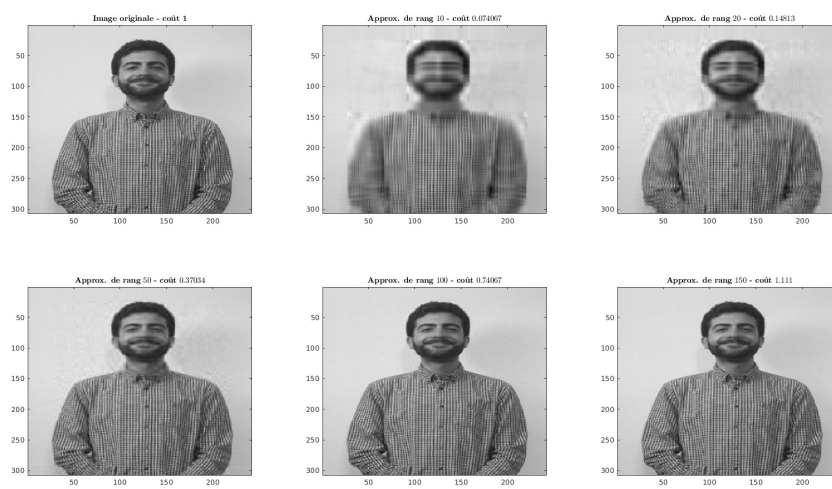


Figure 1.2: Image 307×241 et SVD tronquées avec $k \in \{10, 20, 50, 100, 150\}$; le rang de chaque matrice est indiqué, ainsi que le ratio $\frac{(m+n)*k}{mn}$.

Chapitre 2

Analyse en composantes principales

2.1 Motivation

On se place maintenant dans un contexte du modèle fonctionnel ou factoriel¹. On supposera dans la suite que l'on dispose d'un tableau de données $\mathbf{X} \in \mathbb{R}^{m \times n}$, que l'on écrira des deux manières suivantes :

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix} = [\mathbf{v}_1 \cdots \mathbf{v}_n].$$

Dans cette écriture, les lignes de la matrice \mathbf{X} représentent des individus $\mathbf{x}_i \in \mathbb{R}^n$, chacun possédant n caractéristiques (ou attributs). Les colonnes de la matrice \mathbf{X} représentent quant à elles les vecteurs des attributs : ainsi, le vecteur $\mathbf{v}_j \in \mathbb{R}^m$ contient l'ensemble des valeurs pour le j -ième attribut.

Le but de ce chapitre est de présenter une technique pour extraire l'information importante de \mathbf{X} , afin d'obtenir une représentation plus compacte des données. Nous avons déjà vu un exemple d'un tel outil dans le chapitre 1, avec la décomposition en valeurs singulières tronquée. Cette dernière permet en effet de compresser les données, mais l'interprétation des composantes obtenues n'est pas toujours évidente. Par ailleurs, la SVD ne fournit aucune garantie statistique sur les données, car elle ne fait pas intervenir la distribution des \mathbf{x}_i , qui est inconnue en général.

Dans le reste de ce chapitre, on va développer une approche géométrique d'analyse des données \mathbf{X} . L'idée est de considérer les vecteurs $\{\mathbf{x}_i\}_i$ ou $\{\mathbf{v}_j\}_j$ comme des nuages de points, respectivement dans \mathbb{R}^n et \mathbb{R}^m . On peut alors exprimer une distribution empirique sur ces données, et chercher une représentation de celles-ci qui soit de dimension réduite tout en étant la plus conforme possible aux données d'origine. L'**analyse en composantes principales** repose précisément sur ce principe.

2.2 Statistique empirique et pré-traitement des données

Dans cette section, nous décrivons les statistiques qui peuvent être calculées à partir d'un échantillon de données. Comme dit plus haut, la distribution sous-jacente des données n'est généralement pas connue (on suppose cependant qu'il en existe une). On peut toutefois estimer des statistiques de notre distribution, qui permettent une première transformation des données. Cette étape, appelée **pré-traitement**, permet de se placer dans un contexte plus favorable à l'application de l'analyse en

¹Que l'on peut également considérer comme une problématique d'apprentissage non supervisé.

composantes principales. Mathématiquement parlant, cela correspond à appliquer une transformation linéaire sur les données.

2.2.1 Individu moyen et données centrées

Une première étape dans l'analyse statistique de la matrice de données \mathbf{X} consiste à centrer les données : cette procédure est extrêmement classique en analyse de données, et possède également une justification mathématique forte. On en donne ici

Définition 2.1 (Individu moyen) Soit une matrice $\mathbf{X} \in \mathbb{R}^{m \times n}$. La moyenne arithmétique des variables, que l'on appelle également l'**individu moyen**, est définie comme le vecteur :

$$\bar{x} := \frac{1}{m} \sum_{i=1}^m x_i = \frac{1}{m} \mathbf{X}^T \mathbf{1}_m, \quad \text{avec } \mathbf{1}_m = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^m. \quad (2.2.1)$$

L'individu moyen représente ainsi la **tendance centrale** des données. Sur le plan géométrique, le vecteur \bar{x} représente le **centre de gravité** du nuage de points formé par les individus x_i dans \mathbb{R}^n .

Une opération courante en analyse de données consiste à opérer un **centrage des données**. Cela consiste à translater les individus par rapport à leur moyenne empirique, de sorte à obtenir un nuage de points qui soit centré en l'origine dans \mathbb{R}^n .

Définition 2.2 (Données centrées) Soit $\mathbf{X} \in \mathbb{R}^{m \times n}$ un tableau de données et x_1^T, \dots, x_m^T ses lignes. On dit que l'on **centre les données** lorsque l'on remplace les x_i et \mathbf{X} par

$$x_i^c := x_i - \bar{x}, \quad \mathbf{X}^c := \mathbf{X} - \mathbf{1}_m \bar{x}^T. \quad (2.2.2)$$

2.2.2 Dispersion et dépendance

La première statistique que nous avons établie concernait la moyenne des données; il est naturel de chercher à estimer également la covariance des données (on parlera ici de covariance car on considère des vecteurs aléatoires). Pour ce faire, on introduit la notion suivante.

Définition 2.3 (Matrice de covariance empirique) Soit $\mathbf{X} \in \mathbb{R}^{m \times n}$ un tableau de données et x_1^T, \dots, x_m^T ses lignes. La **matrice de covariance empirique** associée à \mathbf{X} est définie par

$$\Sigma := \frac{1}{m} \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{m} (\mathbf{X}^c)^T \mathbf{X}^c \in \mathbb{R}^{n \times n}, \quad (2.2.3)$$

où \bar{x} désigne l'individu moyen et \mathbf{X}^c la matrice des données centrées.

Les coefficients de la matrice Σ représentent les covariances empiriques des vecteurs d'attributs v_i et v_k . Les composantes diagonales de la matrice de covariance sont particulièrement intéressantes du point de vue géométrique : elles représentent les variances empiriques des n variables (et non des m individus !). On note ainsi

$$\sigma_j^2 := [\Sigma]_{jj} = \frac{1}{m} \sum_{i=1}^m ([x_i]_j - [\bar{x}]_j)^2 = \frac{1}{m} \sum_{i=1}^n \left([v_j]_i - \frac{1}{m} \sum_{k=1}^m [v_j]_k \right)^2.$$

On regroupe généralement ces quantités dans la notion d'inertie.

Définition 2.4 (Inertie) L'**inertie** du nuage de points $\mathbf{X} \in \mathbb{R}^{m \times n}$ est définie par

$$\mathcal{I}(\mathbf{X}) := \text{trace}(\boldsymbol{\Sigma}), \quad (2.2.4)$$

où $\boldsymbol{\Sigma}$ est la matrice de covariance empirique de \mathbf{X} .

L'inertie d'un nuage de points \mathbf{X} décrit ainsi la dispersion des individus \mathbf{x}_i autour de leur centre de gravité. On notera que l'on a

$$\mathcal{I}(\mathbf{X}) = \sum_{j=1}^n \sigma_j^2 = \frac{1}{m} \sum_{i=1}^m \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2.$$

De la même manière que nous avons centré les données à l'aide de l'individu moyen, nous pouvons opérer une transformation des données en utilisant la diagonale de la matrice de covariance empirique.

Définition 2.5 (Données centrées réduites) Soit $\mathbf{X} \in \mathbb{R}^{m \times n}$ un tableau de données et $\mathbf{x}_1^T, \dots, \mathbf{x}_m^T$ ses lignes. La matrice diagonale de réduction pour \mathbf{X} , notée $\mathbf{D}_{1/\sigma}$, est définie par :

$$\mathbf{D}_{1/\sigma} := \begin{bmatrix} \frac{1}{\sigma_1} & 0 \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_n} \end{bmatrix} \in \mathbb{R}^{n \times n} \quad (2.2.5)$$

où $\sigma_1, \dots, \sigma_n$ sont les coefficients diagonaux de la matrice de covariance empirique de \mathbf{X} .

De plus, la matrice

$$\mathbf{X}_{0,1} := \mathbf{X}^c \mathbf{D}_{1/\sigma} = \left[\frac{[\mathbf{x}_i]_j - [\bar{\mathbf{x}}]_j}{\sigma_j} \right]_{\substack{i=1, \dots, m \\ j=1, \dots, n}} \in \mathbb{R}^{m \times n} \quad (2.2.6)$$

s'appelle la matrice de données **centrées réduites**.

La réduction des données \mathbf{X} correspond donc à une transformation diagonale, où l'on normalise chaque attribut selon sa variance empirique. On peut alors étudier la corrélation de cette nouvelle matrice.

Définition 2.6 (Matrice de corrélation empirique) Soit $\mathbf{X} \in \mathbb{R}^{m \times n}$ une matrice de données et $\mathbf{X}_{0,1}$ sa version centrée réduite. Alors, la matrice

$$\mathbf{R} = \frac{1}{m} \mathbf{X}_{0,1}^T \mathbf{X}_{0,1}. \quad (2.2.7)$$

s'appelle la matrice de **corrélation empirique** associée à \mathbf{X} .

Lorsque $[\mathbf{R}]_{ij} = 1$, cela signifie qu'il existe une relation de dépendance affine entre \mathbf{v}_i et \mathbf{v}_j , qui suggère que cette relation affine est valide pour les deux attributs sur l'ensemble de la distribution. On peut dans ce cas réduire les deux attributs à un seul.

La matrice de corrélation empirique permet donc d'identifier des corrélations claires (affines). Dans la pratique, les corrélations peuvent se présenter sous des formes plus complexes, et il n'est pas trivial de réduire les n attributs à un nombre inférieur. L'analyse en composantes principales, que nous détaillons dans la partie suivante, permet précisément de passer de n à $k \leq n$ attributs en conservant le maximum d'information.

2.3 Principe de l'analyse en composantes principales

Dans cette partie, nous fournissons une analyse détaillée de l'analyse en composantes principales, aussi appelée ACP². Nous séparons le cas d'une et de plusieurs composantes principale(s).

2.3.1 Analyse en une composante principale

Nous étudions tout d'abord le cas d'une seule composante principale. Du point de vue géométrique, on considère donc le problème suivant : étant donné un nuage de points $\{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^n$ (que

l'on considèrera comme précédemment sous la forme $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix}$), on souhaite projeter ces

points sur une droite dans \mathbb{R}^n de sorte à conserver le maximum d'information. Cela revient à chercher à remplacer les n variables définissant un individu (ses attributs) par une unique variable appelée **composante principale**, obtenue par combinaison linéaire des variables originelles, qui soit de variance ou d'inertie maximale pour le nuage de points considéré.

Formulation mathématique Afin de préserver le maximum d'information, la droite sur laquelle nous allons projeter devra nécessairement passer par l'individu moyen $\bar{\mathbf{x}}$. L'équation de cette droite sera donc $\{\bar{\mathbf{x}} + t\mathbf{u} | t \in \mathbb{R}\}$, où \mathbf{u} est un vecteur de \mathbb{R}^n que l'on considèrera unitaire (de norme 1) sans perte de généralité.

On souhaite une droite qui représente au mieux chacun des points. Si on note \mathbf{y}_i la projection du vecteur \mathbf{x}_i sur la droite, cela signifie que l'on souhaite que \mathbf{y}_i et \mathbf{x}_i soient les plus proches possibles. Le problème d'optimisation sous-jacent est donc

$$\min_{\mathbf{u} \in \mathbb{R}^n, \|\mathbf{u}\|=1} \frac{1}{m} \sum_{i=1}^m \|\mathbf{y}_i - \mathbf{x}_i\|^2. \quad (2.3.1)$$

La proposition suivante montre que le problème peut s'exprimer relativement au nuage de points formé par les projections \mathbf{y}_i .

Proposition 2.1 *Minimiser l'erreur entre points originaux et points projetés revient à maximiser l'inertie, au sens où le problème (2.3.1) possède le même ensemble de solutions que le problème*

$$\max_{\mathbf{u} \in \mathbb{R}^n, \|\mathbf{u}\|=1} \mathcal{I}(\mathbf{Y}) = \frac{1}{m} \sum_{i=1}^m \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2, \quad (2.3.2)$$

où $\bar{\mathbf{y}}$ désigne le centre de gravité de $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_m^T \end{bmatrix}$.

En résolvant le problème de projection (2.3.1), on cherche donc à maximiser la dispersion des projections sur la droite. La **projection orthogonale** réalise précisément cet objectif.

²Ou PCA, de l'anglais *Principal Component Analysis*.

Définition 2.7 (Projection orthogonale sur une droite) La projection orthogonale d'un ensemble de points $\{x_i\}_i \subset \mathbb{R}^n$ sur une droite de vecteur directeur $u \in \mathbb{R}^n$ (avec $\|u\| = 1$) et passant par \bar{x} est donnée par :

$$y_i = c_i u + \bar{x}, \quad c_i = u^T (x_i - \bar{x}). \quad (2.3.3)$$

On dispose alors des propriétés suivantes.

Lemme 2.1 Soit $\{x_i\}_{i=1}^m$ un nuage de points de \mathbb{R}^n et $\{y_i\}_{i=1}^m$ les projections orthogonales des x_i sur une droite de vecteur directeur u , où $\|u\| = 1$. Alors, les propriétés suivantes sont vérifiées :

- i) On a $\bar{y} = \bar{x}$;
- ii) $c_i u = u^T (x_i - \bar{x}) u = u u^T (x_i - \bar{x})$, où $u u^T$ est une matrice de projection orthogonale;
- iii) $Y = (X - \mathbf{1}_m \bar{x}^T) u u^T + \mathbf{1}_m \bar{x}^T$.

Définition 2.8 On appelle **composante principale** la coordonnée c_i de y_i dans le repère $(\bar{x}; u)$. La droite de vecteur directeur u passant par \bar{x} s'appelle l'**axe principal**, et u est appelé le **vecteur principal**.

On notera que les notions de composante, axe et vecteur principaux ne dépendent que des x_i .

Calcul explicite de la composante principale

Théorème 2.1 Le problème $\max_{u \in \mathbb{R}^n, \|u\|=1} \mathcal{I}(Y)$ possède le même ensemble de solutions que le problème

$$\max_{u \in \mathbb{R}^n, \|u\|=1} u^T \Sigma u. \quad (2.3.4)$$

Le problème (2.3.4) est un problème très classique de mathématiques appliquées : il est lié au calcul des vecteurs propres de la matrice Σ , qui est symétrique et semi-définie positive par construction.

Corollaire 2.1 L'ensemble des solutions du problème (2.3.2) est donné par l'ensemble des vecteurs propres unitaires associés à la plus grande valeur propre de Σ .

Remarque 2.1 Dans la pratique, il sera fréquent que la plus grande valeur propre soit de multiplicité 1, donc qu'il n'existe que deux solutions au problème (2.3.4) qui seront égales au signe près. En cas d'égalité, cependant, on aurait plusieurs directions possibles pour la composante principale, ce qui indiquerait que plusieurs composantes possèderaient la même importance. Ces considérations justifient en partie le recours à l'analyse en plusieurs composantes principales décrite dans la section ci-dessous.

2.3.2 Analyse en plusieurs composantes principales

On étend maintenant l'analyse de la partie précédente en considérant le calcul de plusieurs composantes principales. Du point de vue mathématique, cela revient à projeter sur un sous-espace de dimension k , où $k \geq 1$. On considèrera comme précédemment un sous-espace affine de la forme

$$\bar{x} + \mathcal{S}_k, \quad \mathcal{S}_k := \text{vect}\{u_1, \dots, u_k\}, \quad (2.3.5)$$

où \mathcal{S} est un sous-espace vectoriel de dimension k , et $\mathbf{U}_k = [\mathbf{u}_1 \cdots \mathbf{u}_k] \in \mathbb{R}^{n \times k}$ est une base orthonormale du sous-espace \mathcal{S} (c'est-à-dire $\mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k$, où \mathbf{I}_k est la matrice identité de $\mathbb{R}^{k \times k}$). Comme dans le cas d'une composante principale, on souhaite calculer la projection la plus proche des données originelles. On considère donc le problème

$$\min_{\mathbf{U}_k \in \mathbb{R}^{n \times k}, \mathbf{U}_k^T \mathbf{U}_k = \mathbf{I}_k} \frac{1}{m} \sum_{i=1}^m \|\mathbf{y}_i - \mathbf{x}_i\|^2. \quad (2.3.6)$$

où \mathbf{y}_i représente la projection de \mathbf{x}_i sur l'espace $\bar{\mathbf{x}} + \mathcal{S}_k$. Comme dans le cas d'une composante principale, on va reformuler le problème en tant que maximisation de l'inertie.

Proposition 2.2 *Minimiser l'erreur entre points originaux et points projetés revient à maximiser l'inertie, au sens où le problème (2.3.1) possède le même ensemble de solutions que le problème*

$$\max_{\mathbf{U} \in \mathbb{R}^{n \times k}, \mathbf{U}^T \mathbf{U} = \mathbf{I}_k} \mathcal{I}(\mathbf{Y}) = \frac{1}{m} \sum_{i=1}^m \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2, \quad (2.3.7)$$

où $\bar{\mathbf{y}}$ désigne le centre de gravité de l'ensemble des points projetés $\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_m^T \end{bmatrix}$.

Comme dans le cas unidimensionnel, on dispose d'une formule explicite sur les projections orthogonales.

Définition 2.9 (Projection orthogonale sur un sous-espace affine) *On considère un sous-espace affine de dimension k engendré par une base orthonormale \mathbf{U}_k et passant par le vecteur $\bar{\mathbf{x}}$. Alors, pour tout $i = 1, \dots, m$, la projection de \mathbf{x}_i sur ce sous-espace est donné par :*

$$\mathbf{y}_i = \mathbf{U}_k \mathbf{c}_i + \bar{\mathbf{x}}, \quad \text{où } \mathbf{c}_i = \mathbf{U}_k^T (\mathbf{x}_i - \bar{\mathbf{x}}).$$

Cette propriété conduit à la définition des composantes principales.

Définition 2.10 (Composantes principales) *Les vecteurs \mathbf{c}_i , qui représentent les coordonnées des \mathbf{y}_i dans le repère $(\bar{\mathbf{x}}; \mathbf{u}_1, \dots, \mathbf{u}_k)$, sont appelées les (k premières) **composantes principales**. Les vecteurs $\mathbf{u}_1, \dots, \mathbf{u}_k$ sont appelés les axes principaux.*

Calcul explicite de k composantes principales L'analyse en composantes principales repose sur la création d'une base orthonormale, ce qui permet d'établir le lien suivant entre composantes principales

Proposition 2.3 *Soit \mathbf{U}_k la solution du problème (2.3.6) en $k < n$ composantes principales. Alors, le problème*

$$\max_{\mathbf{U} \in \mathbb{R}^{n \times (k+1)}, \mathbf{U}^T \mathbf{U} = \mathbf{I}_{k+1}} \mathcal{I}(\mathbf{Y}) \quad (2.3.8)$$

admet comme solution $\mathbf{U}_{k+1} = [\mathbf{U}_k \ \mathbf{u}_{k+1}]$, où \mathbf{u}_{k+1} est la composante principale du nuage dans l'espace orthogonal à $\text{vect}\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$, c'est-à-dire une solution du problème

$$\begin{aligned} \max_{\substack{\mathbf{u} \in \mathbb{R}^n, \|\mathbf{u}\|=1 \\ \mathbf{u}^T \mathbf{u}_1 = 0 \\ \vdots \\ \mathbf{u}^T \mathbf{u}_k = 0}} \mathcal{I}(\mathbf{Z}), \end{aligned} \quad (2.3.9)$$

où $Z = \begin{bmatrix} z_1^T \\ \vdots \\ z_m^T \end{bmatrix}$ désigne l'ensemble des projections des points x_1, \dots, x_n .

On voit ainsi que le calcul de k composantes principales se ramène à k problèmes en **une** composante principale résolus de manière successive sur des espaces orthogonaux. Cette idée est en lien direct avec la décomposition en valeurs propres (qui utilise des bases orthonormales donc orthogonales), et conduit au résultat suivant.

Théorème 2.2 *Les solutions du problème (2.3.6) sont donné par les ensembles de k vecteurs propres unitaires de Σ associés aux k plus grandes valeurs propres de Σ , que l'on note $\lambda_1 \geq \dots \geq \lambda_k$.*

On notera qu'en cas de valeurs propres multiples, l'ensemble des solutions peut être ambigu (ce qui peut indiquer qu'une analyse en $k' > k$ composantes principales apportera une information intéressante).

Le résultat ci-dessous permet d'établir des garanties sur les projections y_i en tant qu'approximations des x_i .

Corollaire 2.2 *Sous les hypothèses du théorème 2.2, on a les propriétés suivantes :*

i) *L'écart d'inertie entre les nuages de points projeté et originel est donné par*

$$\frac{\mathcal{I}(Y)}{\mathcal{I}(X)} = \frac{\sum_{j=1}^k \lambda_j}{\mathcal{I}(X)} = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^n \lambda_j}.$$

ii) *Dans le cas général, le tableau Y est solution du problème d'optimisation en variables matricielles suivant :*

$$Y = \underset{M \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \|M - X\|_F^2 \quad \text{sous la contrainte} \quad \operatorname{rang}(M) = k.$$

iii) *Si $k = n$, alors $y_i = x_i$ pour tout i , mais la base utilisée pour décrire les vecteurs est différente.*

En bref : Comment penser l'ACP ? L'analyse en composantes principales peut être vue comme réalisant les opérations suivantes :

- i) Un changement de repère trouvant les axes d'inertie maximale pour les données : on passe ainsi du repère canonique $\{0; e_1, \dots, e_n\}$, où e_1, \dots, e_n sont les vecteurs de la base canonique de \mathbb{R}^n , au repère $\{\bar{x}; u_1, \dots, u_k, \dots, u_n\}$ (et on ne considère généralement qu'une partie des coordonnées dans ce nouveau repère).
- ii) Une réduction de la dimension pour $k < n$: on passe en effet d'une matrice $X \in \mathbb{R}^{m \times n}$ à une matrice de composantes principales $C = [c_i^T]_{i=1, \dots, m} \in \mathbb{R}^{m \times k}$, qui est donc moins coûteuse à stocker en mémoire;
- iii) Un centrage des données, car les composantes forment un jeu de données centrées :

$$\bar{c} = \frac{1}{m} \sum_{i=1}^m c_i = 0;$$

- iv) Une décorrélation des composantes, car la matrice de covariance des composantes centrées (qui sont identiques aux composantes d'origine d'après le point précédent) est diagonale :

$$\Sigma' = C^T C = \text{diag}(\lambda_1, \dots, \lambda_k),$$

où $\lambda_1 \geq \dots \geq \lambda_k$ sont les k plus grandes valeurs propres de Σ .

2.3.3 Application : Reconnaissance de visage

L'une des applications les plus connues de l'analyse en composantes principales est la reconnaissance de visages. L'idée est de partir d'une base d'images de visages, vues comme des vecteurs de dimension égale au nombre de pixels. Dans ce contexte, on peut calculer les axes principaux, qui sont appelées les **eigenfaces**, ou "visages propres" (par analogie avec les vecteurs propres). On peut ainsi se servir de ces eigenfaces pour approcher une image, mais aussi pour reconnaître un visage sur une image absente du jeu de données originel.

Partie II

Régression linéaire

Chapitre 3

Premiers pas avec le modèle linéaire

Dans le chapitre 1, nous avons introduit la décomposition en valeurs singulières, une technique visant à extraire de l'information d'un jeu de données : ce paradigme est celui de l'apprentissage *non supervisé*. Ce chapitre aborde un autre paradigme, celui de l'apprentissage *supervisé*, dans lequel il s'agira de déterminer un modèle (une fonction linéaire pour les besoins de ce cours) décrivant une relation entre différents éléments d'un jeu de données, pour potentiellement prédire (on parle également d'inférer) le comportement de données futures.

3.1 Introduction

On considère un jeu de données ayant m éléments ou individus, et on associe à chaque individu n caractéristiques¹ sous la forme d'un vecteur de \mathbb{R}^n . Soient x_1, \dots, x_n ces vecteurs : on les regroupe alors sous la forme d'une matrice de données

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times n}. \quad (3.1.1)$$

Exemple 3.1 • Chaque ligne de X représente un individu, et les n composantes de x_i sont des données médicales (âge, poids, taux de cholestérol, etc).

- Chaque ligne de X est une "vectorisation" d'une image 2D, et les valeurs de x_i sont celles des pixels, en niveau de gris. Ainsi, une image de taille 480*640 serait transformée (en mettant les lignes bout à bout, par exemple) en un vecteur de taille $n = 480 * 640 = 307200$.

Sans autre information que la matrice elle-même, on peut appliquer des techniques d'algèbre linéaire pour en extraire de l'information : c'est ce que réalisait la SVD dans le chapitre précédent. Dans un contexte d'apprentissage supervisé, on associe chaque vecteur de caractéristiques x_i à un **label** $y_i \in \mathbb{R}$, qui peut représenter une classe à laquelle l'individu appartient (malade/non malade, image de chien ou de chat, etc). Ces labels sont concaténés pour former un vecteur de labels $y \in \mathbb{R}^m$.

Par conséquent, notre but n'est plus seulement d'analyser l'information de la matrice X , mais bien de trouver une relation entre les caractéristiques X et les labels y . Pour ce cours, on postulera que

¹Ou *features* en anglais.

cette relation est linéaire : on va donc chercher une fonction $h : \mathbb{R}^n \rightarrow \mathbb{R}$ de la forme $h(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$. On souhaite que h permette d'obtenir les y_i à partir des x_i , c'est-à-dire que l'on voudrait avoir

$$h(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta} = y_i \quad \forall i = 1, \dots, m,$$

que l'on peut ré-écrire sous forme matricielle comme

$$\mathbf{X} \boldsymbol{\beta} = \mathbf{y}.$$

On se trouve donc en présence d'un système linéaire que l'on va vouloir résoudre. Rien ne garantit a priori que ce système possède une solution, ou que cette solution (si elle existe) est unique. Une étude plus approfondie des systèmes d'équations linéaires semble donc nécessaire.

3.2 Résolution de systèmes non linéaires

Définition 3.1 Un système linéaire de m équations à n inconnues β_1, \dots, β_n est donné par

$$\begin{array}{cccccc} x_{11}\beta_1 & + & x_{12}\beta_2 & + & \dots & + & x_{1n}\beta_n & = & y_1 \\ x_{21}\beta_1 & + & x_{22}\beta_2 & + & \dots & + & x_{2n}\beta_n & = & y_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ x_{m1}\beta_1 & + & x_{m2}\beta_2 & + & \dots & + & x_{mn}\beta_n & = & y_m \end{array}$$

ou, sous forme compacte,

$$\mathbf{X} \boldsymbol{\beta} = \mathbf{y},$$

avec $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\boldsymbol{\beta} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$.

Dans la suite, nous allons déterminer les conditions d'existence de solutions à ce système linéaire.

3.2.1 Cas d'un système carré

On s'intéresse tout d'abord au cas où $n = m$: on a donc un système dit "carré" de n équations à n inconnues de la forme

$$\mathbf{X} \boldsymbol{\beta} = \mathbf{y}, \tag{3.2.1}$$

où $\mathbf{X} \in \mathbb{R}^{n \times n}$, $\mathbf{y} \in \mathbb{R}^n$, et $\boldsymbol{\beta} \in \mathbb{R}^n$ représente les paramètres inconnus de notre modèle.

L'existence et l'unicité de solutions au système (3.2.1) dépendent des propriétés de la matrice \mathbf{X} et du vecteur \mathbf{y} , comme le montrent les exemples ci-dessous.

Exemple 3.2 a) Le système

$$\begin{cases} \beta_1 + \beta_2 & = & 0, \\ 3\beta_1 + 2\beta_2 & = & 1. \end{cases}$$

possède une unique solution $\beta_1 = 1$, $\beta_2 = -1$.

b) Le système

$$\begin{cases} \beta_1 & = & 0, \\ \beta_1 & = & -1. \end{cases}$$

ne possède pas de solution.

c) Le système

$$\begin{cases} \beta_1 + 2\beta_2 = 2, \\ 2\beta_1 + 4\beta_2 = 4. \end{cases}$$

possède une infinité de solutions.

On peut donc se trouver dans trois cas différents : pour les deux derniers, on ne sait pas ce qui peut être fait. En revanche, le premier cas correspond à une matrice \mathbf{X} inversible : dans cette situation, il existe une caractérisation de la solution du système.

Théorème 3.1 (Résolution d'un système carré inversible) Soient $\mathbf{X} \in \mathbb{R}^{n \times n}$ une matrice inversible et $\mathbf{y} \in \mathbb{R}^n$. Le système carré inversible $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ possède une unique solution $\boldsymbol{\beta}^*$ donnée par

$$\boldsymbol{\beta}^* = \mathbf{X}^{-1}\mathbf{y}.$$

Lorsque la matrice \mathbf{X} n'est pas inversible en revanche, c'est-à-dire que $\text{rang}(\mathbf{X}) < n$, il existera une infinité de solutions si $\mathbf{y} \in \text{Im}(\mathbf{X})$, et aucune si $\mathbf{y} \notin \text{Im}(\mathbf{X})$.

3.2.2 Cas d'un système rectangulaire

On considère maintenant le cas d'un système linéaire rectangulaire non carré, soit

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y}, \quad \mathbf{X} \in \mathbb{R}^{m \times n}, \quad m \neq n. \quad (3.2.2)$$

Comme le montrent les exemples ci-dessous, on retrouve les mêmes cas que pour un système carré.

Exemple 3.3 a) Le système

$$\begin{aligned} \beta_1 + \beta_2 &= 0, \\ 3\beta_1 + 2\beta_2 &= 1, \\ 6\beta_1 + 5\beta_2 &= 1. \end{aligned}$$

possède une unique solution ($\beta_1 = 1, \beta_2 = -1$).

b) Le système

$$\beta_1 + 2\beta_2 = 2.$$

possède une infinité de solutions.

c) Le système

$$\begin{aligned} \beta_1 &= 2, \\ \beta_2 &= 3, \\ \beta_1 + \beta_2 &= 0. \end{aligned}$$

ne possède pas de solution.

Sans information sur le système linéaire, il ne semble donc pas possible de déterminer s'il possède ou non des solutions. On peut cependant être plus spécifique pour certains types de systèmes satisfaisant la propriété ci-dessous.

Définition 3.2 Une matrice $\mathbf{X} \in \mathbb{R}^{m \times n}$ est de **rang plein** si $\text{rang}(\mathbf{X}) = \min\{m, n\}$.

Toute matrice carrée inversible est de rang plein, mais cette notion est plus générique. On a ainsi les cas particuliers suivants.

Théorème 3.2 Soient $\mathbf{X} \in \mathbb{R}^{m \times n}$ de rang plein et $\mathbf{y} \in \mathbb{R}^m$ avec $m \neq n$. On a :

- a) Si $\text{rang}(\mathbf{X}) = n$, alors $\mathbf{X}^T \mathbf{X}$ est inversible et le système (3.2.2) possède une unique solution donnée par $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ lorsque $\mathbf{y} \in \text{Im}(\mathbf{X})$;
- b) Si $\text{rang}(\mathbf{X}) = m$, alors $\mathbf{X} \mathbf{X}^T$ est inversible et le vecteur $\mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{y}$ est solution de (3.2.2)

Il est donc toujours possible de déterminer une solution au problème (3.2.2) lorsque \mathbf{X} est de rang plein. Lorsque \mathbf{X} n'est pas de rang plein, on retombe en revanche dans les mêmes difficultés que pour un système carré.

Comme on le verra dans la section suivante, les différentes expressions établies dans les théorèmes 3.1 et 3.2 correspondent en fait à une même formule.

3.2.3 Pseudo-inverse et SVD

Le concept d'inverse d'une matrice carrée peut être généralisé au cas d'une matrice rectangulaire : c'est le principe de la pseudo-inverse² que l'on décrit ci-dessous.

Définition 3.3 (Pseudo-inverse d'une matrice) Soit une matrice $\mathbf{X} \in \mathbb{R}^{m \times n}$. Il existe une unique matrice $\mathbf{M} \in \mathbb{R}^{n \times m}$ vérifiant les équations de Penrose :

$$\begin{aligned} \mathbf{X} \mathbf{M} \mathbf{X} &= \mathbf{X} \\ \mathbf{M} \mathbf{X} \mathbf{M} &= \mathbf{M} \\ (\mathbf{X} \mathbf{M})^T &= \mathbf{X} \mathbf{M} \\ (\mathbf{M} \mathbf{X})^T &= \mathbf{M} \mathbf{X} \end{aligned}$$

Cette matrice s'appelle la **pseudo-inverse** de \mathbf{X} et on la note \mathbf{X}^\dagger .

Sous sa forme générale, le calcul de la pseudo-inverse ne semble pas évident. Il existe heureusement une formule explicite de la pseudo-inverse basée sur la décomposition en valeurs singulières.

Proposition 3.1 Soit une matrice diagonale par blocs $\Sigma \in \mathbb{R}^{m \times n}$ de la forme

$$\left[\begin{array}{ccc|c} \sigma_1 & 0 \cdots & 0 & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & \cdots 0 & \sigma_r & 0 \\ \hline 0 & \cdots & \cdots & 0 \end{array} \right]$$

avec $\sigma_1 \geq \cdots \geq \sigma_r > 0$. La pseudo-inverse de la matrice Σ est la matrice $\Sigma^\dagger \in \mathbb{R}^{n \times m}$ définie par

$$\Sigma^\dagger = \left[\begin{array}{ccc|c} \frac{1}{\sigma_1} & 0 \cdots & 0 & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & \cdots 0 & \frac{1}{\sigma_r} & 0 \\ \hline 0 & \cdots & \cdots & 0 \end{array} \right].$$

²On parle aussi d'inverse généralisée ou d'inverse de Moore-Penrose.

Le résultat de la proposition 3.1 illustre bien le concept de pseudo-inverse : on a ainsi “inversé” uniquement le bloc diagonal contenant des valeurs non nulles, le reste étant simplement transposé pour inverser les espaces de départ et d’arrivée.

On a alors le résultat générique suivant.

Théorème 3.3 (Formule de pseudo-inverse) *Soit une matrice $X \in \mathbb{R}^{m \times n}$ et $U\Sigma V^T$ une décomposition en valeurs singulières de cette matrice. Alors, la pseudo-inverse de X est donnée par*

$$X^\dagger = V\Sigma^\dagger U^T. \quad (3.2.3)$$

On peut vérifier que la formule (3.2.3) satisfait bien aux équations de Penrose. Ces dernières suggèrent qu’il est possible d’utiliser la pseudo-inverse d’une manière similaire à celle de l’inverse pour une matrice carrée inversible : le lien entre inverse et pseudo-inverse est encore plus ténu, comme le montre le corollaire suivant.

Corollaire 3.1 *Soit $X \in \mathbb{R}^{m \times n}$ une matrice de rang plein. Alors,*

- i) *Si $\text{rang}(X) = m$, alors $X^\dagger = X^T(XX^T)^{-1}$;*
- ii) *Si $\text{rang}(X) = n$, alors $X^\dagger = (X^T X)^{-1}X^T$;*
- iii) *Si $\text{rang}(X) = n$ et que $n = m$, alors X est carrée inversible, et les deux formules ci-dessus correspondent à $X^\dagger = X^{-1}$.*

La pseudo-inverse est ainsi apparue dans les sections 3.2.1 et 3.2.2 lorsque l’on supposait que la matrice X était de rang plein. En ce sens, il semble que la technique de pseudo-inverse soit adaptée aux problèmes bien posés. L’approche par moindres carrés, développée dans la section suivante, va permettre de formaliser cette propriété, et de mettre en lumière le rôle plus large joué par la pseudo-inverse.

3.3 Moindres carrés linéaires

Comme expliqué dans la section précédente, la notion de solution d’un système linéaire perd de son sens lorsque le système ne possède pas de solution, ou une infinité. Pour cette raison, on définit un autre concept de solution, dite au sens des moindres carrés.

3.3.1 Solution au sens des moindres carrés

Un problème aux moindres carrés linéaires est un problème d’optimisation, et plus précisément de minimisation : étant donnés $X \in \mathbb{R}^{m \times n}$ et $y \in \mathbb{R}^m$, on ne cherche plus à résoudre $X\beta = y$ de manière exacte, mais plutôt à minimiser l’écart entre les vecteurs $X\beta$ et y . Du point de vue mathématique, on évaluera cet écart via la norme euclidienne, et on cherchera donc à résoudre le problème (dit aux moindres carrés linéaires) suivant :

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \|X\beta - y\|^2. \quad (3.3.1)$$

Il s'agit d'un problème de minimisation de la fonction $\beta \mapsto \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}\|^2$ selon β , sous sa forme standard en optimisation.³ Notons que si l'on décompose la norme, on obtient :

$$\|\mathbf{X}\beta - \mathbf{y}\|^2 = \sum_{i=1}^m (\mathbf{x}_i^T \beta - y_i)^2.$$

Cet objectif que l'on cherche à réduire représente donc bien l'attachement aux données, c'est-à-dire la correspondance entre notre modèle linéaire et les labels de chaque individu.

Définition 3.4 (Solution au sens des moindres carrés) Soient $\mathbf{X} \in \mathbb{R}^{m \times n}$ et $\mathbf{y} \in \mathbb{R}^m$. L'ensemble des vecteurs $\beta \in \mathbb{R}^n$ tels que la valeur de $\frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}\|^2$ soit minimale, que l'on note :

$$\arg \min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}\|^2, \quad (3.3.2)$$

s'appelle l'ensemble des **solutions au sens des moindres carrés**.

La notion de solution au sens des moindres carrés est à distinguer de celle d'une solution du système linéaire, pour laquelle nous introduisons la terminologie ci-dessous.

Définition 3.5 (Solution au sens classique) Soient $\mathbf{X} \in \mathbb{R}^{m \times n}$ et $\mathbf{y} \in \mathbb{R}^m$. L'ensemble des vecteurs $\beta \in \mathbb{R}^n$ tels que $\mathbf{X}\beta = \mathbf{y}$ s'appelle l'ensemble des **solutions au sens classique** du système linéaire.

Cet ensemble peut être vide, et il est nécessairement inclus dans l'ensemble des solutions au sens des moindres carrés.

Le concept de solution au sens des moindres carrés nous permet d'introduire différents concepts, liés notamment à la pseudo-inverse.

Théorème 3.4 Pour tous $\mathbf{X} \in \mathbb{R}^{m \times n}$ et $\mathbf{y} \in \mathbb{R}^m$, les propriétés suivantes sont vérifiées :

1. l'ensemble défini par (3.3.2) est toujours non vide;
2. le vecteur $\mathbf{X}^\dagger \mathbf{y}$ est une solution au sens des moindres carrés;
3. parmi toutes les solutions au sens des moindres carrés, le vecteur $\mathbf{X}^\dagger \mathbf{y}$ est la solution de norme minimale :

$$\forall \gamma \in \arg \min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|\mathbf{X}\beta - \mathbf{y}\|^2, \quad \|\gamma\| \geq \|\mathbf{X}^\dagger \mathbf{y}\|.$$

Ce théorème ne sera pas démontré dans ce cours (voir à cet égard le cours de *Méthodes numériques : Optimisation* au semestre 2, ou le cours *Mathématiques pour les sciences des données*). En revanche, nous exploiterons fortement ce résultat pour calculer une solution au sens des moindres carrés.

³Pour des raisons de normalisation, on introduit notamment un facteur 1/2.

3.3.2 Résolution du problème aux moindres carrés

Le théorème ci-dessous récapitule l'ensemble des cas à considérer dans le calcul d'un modèle linéaire.

Théorème 3.5 Soient $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{y} \in \mathbb{R}^m$ et $\beta^* = \mathbf{X}^\dagger \mathbf{y}$.

1) Si $m = n$, on distingue trois cas :

- a) Si $\text{rang}(\mathbf{X}) = m = n$, on a $\beta^* = \mathbf{X}^{-1} \mathbf{y}$, et il s'agit de l'unique solution au sens classique et au sens des moindres carrés;
- b) Si $\text{rang}(\mathbf{X}) < m$ et $\mathbf{y} \in \text{Im}(\mathbf{X})$, alors β^* est une solution au sens classique et de norme minimale au sens des moindres carrés. Les problèmes (3.2.1) et (3.3.1) admettent chacun une infinité de solutions.
- c) Si $\text{rang}(\mathbf{X}) < m$ et $\mathbf{y} \notin \text{Im}(\mathbf{X})$, alors il n'existe pas de solution au sens classique; en revanche, β^* est la solution de norme minimale au sens des moindres carrés.

2) Si $m < n$ (système sous-déterminé), on distingue trois cas :

- b) Si $\text{rang}(\mathbf{X}) = m$, alors $\beta^* = \mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{y}$ et il s'agit à la fois d'une solution classique et de la solution de norme minimale au sens des moindres carrés. Les problèmes (3.2.2) et (3.3.1) admettent chacun une infinité de solutions.
- b) Si $\text{rang}(\mathbf{X}) < m$ et $\mathbf{y} \in \text{Im}(\mathbf{X})$, alors β^* est une solution au sens classique et de norme minimale au sens des moindres carrés. Les problèmes (3.2.1) et (3.3.1) admettent chacun une infinité de solutions.
- b) Si $\text{rang}(\mathbf{X}) < m$ et $\mathbf{y} \notin \text{Im}(\mathbf{X})$, alors il n'existe pas de solution au sens classique; en revanche, β^* est la solution de norme minimale au sens des moindres carrés.

3) Si $m > n$ (système sur-déterminé), on distingue trois cas :

- a) Si $\text{rang}(\mathbf{X}) = n$, alors $\beta^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ et il s'agit de l'unique solution au sens des moindres carrés. C'est une solution au sens classique lorsque $\mathbf{y} \in \text{Im}(\mathbf{X})$.
- b) Si $\text{rang}(\mathbf{X}) < n$ et $\mathbf{y} \in \text{Im}(\mathbf{X})$, alors β^* est une solution au sens classique et de norme minimale au sens des moindres carrés. Les problèmes (3.2.1) et (3.3.1) admettent chacun une infinité de solutions.
- c) Si $\text{rang}(\mathbf{X}) < n$ et $\mathbf{y} \notin \text{Im}(\mathbf{X})$, alors il n'existe pas de solution au sens classique; en revanche, β^* est la solution de norme minimale au sens des moindres carrés.

3.4 Conclusion

Dans le contexte de l'apprentissage supervisé, on peut être amené à vouloir expliquer nos données par un modèle linéaire. Ce choix de modélisation fait naturellement apparaître des systèmes linéaires, dont l'étude repose sur des propriétés issues de l'algèbre linéaire. Il apparaît alors que le problème peut être bien ou mal posé, selon que le système possède une, des ou même aucune solution(s).

On a ainsi introduit le concept de *problème aux moindres carrés associé à un système linéaire*, qui a permis de formuler la procédure d'apprentissage du modèle linéaire en prenant en compte les cas où le modèle ne peut pas expliquer les données de manière unique. Grâce à la pseudo-inverse, nous avons pu caractériser une solution du problème qui fournit la meilleure erreur d'approximation au sens des moindres carrés, tout en permettant d'avoir un modèle plus simple au sens de la norme.

Chapitre 4

Régression linéaire

Bibliographie

- [1] S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra - Vectors, Matrices and Least Squares*. Cambridge University Press, Cambridge, United Kingdom, 2018.
- [2] S. L. Brunton and J. N. Kutz. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, Cambridge, United Kingdom, 2019.
- [3] G. H. Golub and C. F. Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, fourth edition, 2013.