Dauphine | PSL
UNIVERSITÉ PARIS

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**
  - 2024.11.05 : Ajout du chapitre 4 en version complète.
  - 2024.10.15 : 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

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 préfèrera 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).

[1] Google, 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** : $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$.
- Les lettres majuscules en gras seront utilisées pour les matrices : $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$.
- 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 $\boldsymbol{x} \in \mathbb{R}^n$ 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 $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, que l'on notera plus succinctement $\boldsymbol{x} = [x_i]_{1\leq 1\leq n}$.
- Étant donné un vecteur (colonne) $\boldsymbol{x} \in \mathbb{R}^n$, le vecteur ligne correspondant sera noté $\boldsymbol{x}^{\mathrm{T}}$. On aura donc $\boldsymbol{x}^{\mathrm{T}} = [x_1 \ \cdots \ x_n]$ et $[\boldsymbol{x}^{\mathrm{T}}]^{\mathrm{T}} = \boldsymbol{x}$.
- 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 considèrer un vecteur de $\mathbb{R}^m$ comme une matrice de $\mathbb{R}^{m\times 1}$, et vice versa. Une matrice $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ est dite carrée (dans le cas général, on parlera de matrice rectangulaire).

- Étant donnée une matrice $\boldsymbol{A} \in \mathbb{R}^n$, on notera $\boldsymbol{A}_{ij}$ le coefficient en ligne i et colonne j de la matrice. La notation $[\boldsymbol{A}_{ij}]_{\substack{1\leq i\leq m\\ 1\leq j\leq n}}$ sera donc équivalente à $\boldsymbol{A}$. Sans ambigüité sur la taille de la matrice, on notera simplement $[\boldsymbol{A}_{ij}]$.

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

- Pour une matrice $\boldsymbol{A} = [\boldsymbol{A}_{ij}] \in \mathbb{R}^{m\times n}$, la matrice transposée de $\boldsymbol{A}$, notée $\boldsymbol{A}^{\mathrm{T}}$, est la matrice de $\mathbb{R}^{n\times m}$ telle que
$$\forall i = 1\dots m,\ \forall j = 1\dots n, \quad \boldsymbol{A}^{\mathrm{T}}_{ji} = \boldsymbol{A}_{ij}.$$
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 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, la somme des vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ est notée $\boldsymbol{x} + \boldsymbol{y} = [x_i + y_i]_{1 \leq i \leq n}$;
- Pour tout $\lambda \in \mathbb{R}$, on définit $\lambda \boldsymbol{x} := \lambda \cdot \boldsymbol{x} = [\lambda x_i]_{1 \leq i \leq n}$.
- La norme euclidienne $\|\cdot\|$ sur $\mathbb{R}^n$ est définie pour tout vecteur $\boldsymbol{x} \in \mathbb{R}^n$ par

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

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

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

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

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

- Il existe une famille libre et génératrice de $\mathbb{R}^n$ de taille n. Par exemple, tout vecteur $\boldsymbol{x}$ de $\mathbb{R}^n$ s'écrit $\boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}_i$, où $\boldsymbol{e}_i = [0 \cdots 0\ 1\ 0 \cdots 0]^{\mathrm{T}}$ 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 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ p vecteurs de $\mathbb{R}^n$. Le sous-espace engendré par les vecteurs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ est le sous-espace vectoriel*

$$\mathrm{vect}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) := \left\{ \boldsymbol{x} = \sum_{i=1}^{p} \alpha_i \boldsymbol{x}_i \middle| \alpha_i \in \mathbb{R}^n \ \forall i \right\}.$$

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 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, on définit les deux sous-espaces suivants :*

- *Le* **noyau** *(kernel/null space en anglais) de $\boldsymbol{A}$ est le sous-espace vectoriel*

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

- *L'***image** *(range space) de $\boldsymbol{A}$ est le sous-espace vectoriel*

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

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

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

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

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*

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

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 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ est dite symétrique si elle vérifie $\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{A}$.*

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

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

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

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

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

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

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

Si $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ est une matrice orthogonale, alors $\boldsymbol{Q}^{\mathrm{T}}$ est également orthogonale.
On utilisera fréquemment la propriété des matrices orthogonales énoncée ci-dessous.

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

$$\|\boldsymbol{A}\| = \|\boldsymbol{U}\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{V}\| et \|\boldsymbol{A}\|_F = \|\boldsymbol{U}\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{V}\|_F,$$

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 $\boldsymbol{x} \in \mathbb{R}^m$, on a $\|\boldsymbol{U}\boldsymbol{x}\| = \|\boldsymbol{x}\|$. En utilisant la définition de la norme et celle d'une matrice orthogonale, on a :

$$\|\boldsymbol{U}\boldsymbol{x}\|^2 = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = \|\boldsymbol{x}\|^2,$$

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

$$\|\boldsymbol{U}\boldsymbol{A}\boldsymbol{x}\| = \|\boldsymbol{A}\boldsymbol{x}\|$$

pour tout vecteur $\boldsymbol{x}$. En revenant à la définition de la norme d'opérateur, on obtient ainsi

$$\|\boldsymbol{U}\boldsymbol{A}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{U}\boldsymbol{A}\boldsymbol{x}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\| = \|\boldsymbol{A}\|,$$

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

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{V}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{V}\boldsymbol{x}\| &= \max_{\|\boldsymbol{V}^{\mathrm{T}}\boldsymbol{z}\|=1} \|\boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{z}\| \\ &= \max_{\|\boldsymbol{z}\|=1} \|\boldsymbol{A}\boldsymbol{z}\| = \|\boldsymbol{A}\|,\end{aligned}$$

ce qu'il fallait démontrer.

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

$$\|\boldsymbol{A}\|_F = \sqrt{\|\boldsymbol{a}_{1\bullet}\|^2 + \cdots + \|\boldsymbol{a}_{m\bullet}\|^2} = \sqrt{\|\boldsymbol{a}_{\bullet 1}\|^2 + \cdots + \|\boldsymbol{a}_{\bullet n}\|^2}$$

où $\boldsymbol{a}_{1\bullet}, \ldots, \boldsymbol{a}_{m\bullet}$ et $\boldsymbol{a}_{\bullet 1}, \ldots, \boldsymbol{a}_{\bullet n}$ représentent les lignes et les colonnes de $\boldsymbol{A}$, respecitvement. 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 $\boldsymbol{U} \in \mathbb{R}^{m\times m}$ orthogonale, on a

$$\sqrt{\|\boldsymbol{U}\boldsymbol{a}_{\bullet 1}\|^2 + \cdots + \|\boldsymbol{U}\boldsymbol{a}_{\bullet n}\|^2} = \sqrt{\|\boldsymbol{a}_{\bullet 1}\|^2 + \cdots + \|\boldsymbol{a}_{\bullet n}\|^2},$$

d'où $\|\boldsymbol{U}\boldsymbol{A}\|_F = \|\boldsymbol{A}\|_F$. En considérant les lignes de $\boldsymbol{A}$, on montre de même que $\|\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{V}\|_F$. □

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

1.2 Valeurs propres et décomposition spectrale

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

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

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

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 $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, on a les propriétés suivantes :*

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

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

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

où $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ est une matrice orthogonale , dont les colonnes $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$ forment une base orthonormée *de vecteurs propres, et $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times n}$ est une matrice diagonale qui contient les n valeurs propres de $\boldsymbol{A}$ $\lambda_1, \ldots, \lambda_n$ 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 $\boldsymbol{A}$ par son effet sur les vecteurs $\boldsymbol{p}_i$. Ainsi, lorsque $|\lambda_i| >> 1$, on aura $\|\boldsymbol{A}\boldsymbol{p}_i\| >> \|p_i\|$, et la matrice aura donc un effet expansif dans la direction de $\boldsymbol{p}_i$ (ou sa direction opposée lorsque $\lambda_i < 0$. De même, si $|\lambda_i| << 1$, la matrice aura un effet contractant dans la direction de $\boldsymbol{p}_i$: le cas extrême est $\lambda_i = 0$, c'est-à-dire que $\boldsymbol{p}_i \in \ker(\boldsymbol{A})$ et la matrice ne conserve donc pas d'information relative à $\boldsymbol{p}_i$.

Géométriquement parlant, on voit ainsi que, pour tout vecteur $\boldsymbol{x} \in \mathbb{R}^n$ décomposé dans la base des $\boldsymbol{p}_i$ que l'on multiplie par $\boldsymbol{A}$, les composantes de ce vecteur associées aux plus grandes valeurs propres[1] 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 audios, des images, etc.

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

1.3.1 Principe de la décomposition

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

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

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

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

i) $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ *est semi-définie positive;*

ii) $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ *est symétrique.*

iii) $\ker(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \ker(\boldsymbol{A})$*;*

iv) $\mathrm{Im}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \mathrm{Im}(\boldsymbol{A}^{\mathrm{T}})$*;*

v) $\mathrm{rang}(\boldsymbol{A}^{T}\boldsymbol{A}) = \mathrm{rang}(\boldsymbol{A})$.

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 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ admet une* **décomposition en valeurs singulières** *(*SVD[2]*) de la forme*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}},$$

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

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

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*

$$\boldsymbol{A} = \underbrace{[\boldsymbol{u}_1\ \boldsymbol{u}_2\ \boldsymbol{u}_3]}_{\boldsymbol{U}} \overbrace{\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix}}^{\boldsymbol{\Sigma}} \underbrace{\begin{bmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \boldsymbol{v}_2^{\mathrm{T}} \end{bmatrix}}_{\boldsymbol{V}^{\mathrm{T}}}$$

où $\sigma_1 \geq 0, \sigma_2 \geq 0$, les $\boldsymbol{u}_i$ forment une base orthonormée de $\mathbb{R}^3$ et les $\boldsymbol{v}_i$ forment une base orthonormée de $\mathbb{R}^2$.

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

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

i) Les carrés des valeurs singulières sont les valeurs propres de $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$.

ii) Si $\mathrm{rang}(\boldsymbol{A}) = r$, 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 $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ de rang r admet une* SVD *réduite de la forme*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{1.3.1}$$

où $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ avec $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \mathbf{I}$, $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ avec $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \mathbf{I}_r$ et $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$ 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 $(n+m)r$ réels[3] pour pouvoir reconstruire la matrice, ce qui peut être plus avantageux que de stocker les mn coefficients de la matrice $\boldsymbol{A}$.

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

$$\boldsymbol{U} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_r], \quad \boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_r], \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0\cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots 0 & \sigma_r. \end{bmatrix}.$$

On suppose que $\sigma_1 \geq \cdots \geq \sigma_r$. Alors, pour tout $k \leq r$, la décomposition $\boldsymbol{U}_k\boldsymbol{\Sigma}_{k,k}\boldsymbol{V}_k^{\mathrm{T}}$, où

$$\boldsymbol{U}_k = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_k],\ \boldsymbol{V}_k = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_k],\ \boldsymbol{\Sigma}_{k,k} = \begin{bmatrix} \sigma_1 & 0\cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots 0 & \sigma_k. \end{bmatrix}$$

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

[3] Le coût est de $(n+m+1)*r$ si on stocke les valeurs singulières séparément, mais on peut également les incorporer dans $\boldsymbol{U}$ ou $\boldsymbol{V}$, 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 $\boldsymbol{A}$ 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 $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Si on calcule la décomposition en valeurs singulières de $\boldsymbol{A}$, 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

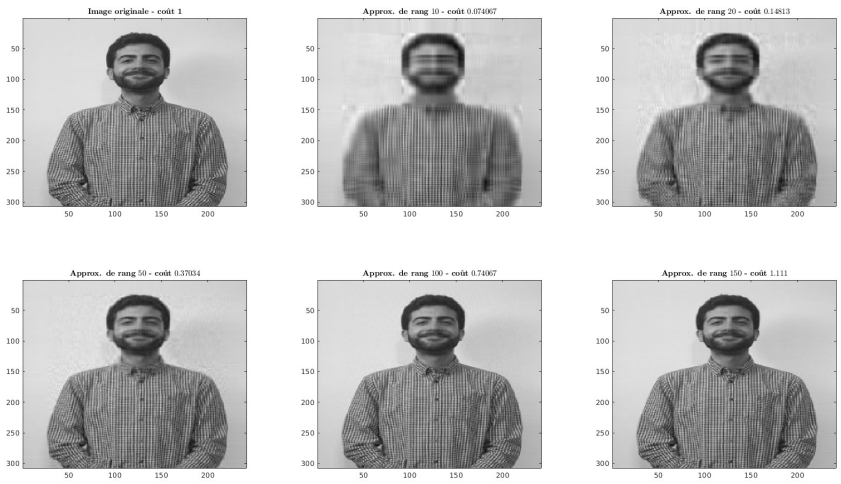

Figure 1.2: Image 307x241 et SVD tronquées avec $k \in \{10, 20, 50, 100, 150\}$; le rang de chaque matrice est indiqué, ainsi que le ratio $\frac{(m+n)*k}{mn}$.

Chapitre 2

Analyse en composantes principales

2.1 Motivation

On se place maintenant dans un contexte du modèle fonctionnel ou factoriel[1]. On supposera dans la suite que l'on dispose d'un tableau de données $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, que l'on écrira des deux manières suivantes :

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

Dans cette écriture, les lignes de la matrice $\boldsymbol{X}$ représentent des individus $\boldsymbol{x}_i \in \mathbb{R}^n$, chacun possédant n caractéristiques (ou attributs). Les colonnes de la matrice $\boldsymbol{X}$ représentent quant à elles les vecteurs des attributs : ainsi, le vecteur $\boldsymbol{v}_j \in \mathbb{R}^m$ 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 $\boldsymbol{X}$, 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 $\boldsymbol{x}_i$, qui est inconnue en général.

Dans le reste de ce chapitre, on va développer une approche géométrique d'analyse des données $\boldsymbol{X}$. L'idée est de considérer les vecteurs $\{\boldsymbol{x}_i\}_i$ ou $\{\boldsymbol{v}_j\}_j$ 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

[1] Que 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 $\boldsymbol{X}$ 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$. La moyenne arithmétique des variables, que l'on appelle également l'***individu moyen***, est définie comme le vecteur :*

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

L'individu moyen représente ainsi la **tendance centrale** des données. Sur le plan géométrique, le vecteur $\bar{\boldsymbol{x}}$ représente le **centre de gravité** du nuage de points formé par les individus $\boldsymbol{x}_i$ 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ un tableau de données et $\boldsymbol{x}_1^{\mathrm{T}}, \dots, \boldsymbol{x}_m^{\mathrm{T}}$ ses lignes. On dit que l'on centre les données lorsque l'on remplace les $\boldsymbol{x}_i$ et $\boldsymbol{X}$ par*

$$\boldsymbol{x}_i^c := \boldsymbol{x}_i - \bar{\boldsymbol{x}}, \quad \boldsymbol{X}^c := \boldsymbol{X} - \mathbf{1}_m\bar{\boldsymbol{x}}^{\mathrm{T}}. \tag{2.2.2}$$

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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ un tableau de données et $\boldsymbol{x}_1^{\mathrm{T}}, \dots, \boldsymbol{x}_m^{\mathrm{T}}$ ses lignes. La* **matrice de covariance empirique** *associée à $\boldsymbol{X}$ est définie par*

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

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

Les coeffficients de la matrice $\boldsymbol{\Sigma}$ représentent les covariances empiriques des vecteurs d'attributs $\boldsymbol{v}_i$ et $\boldsymbol{v}_k$. 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

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

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

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

$$\mathcal{I}(\boldsymbol{X}) := \operatorname{trace}(\boldsymbol{\Sigma}), \tag{2.2.4}$$

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

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

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

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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ un tableau de données et $\boldsymbol{x}_1^{\mathrm{T}}, \dots, \boldsymbol{x}_m^{\mathrm{T}}$ ses lignes. La* matrice diagonale de réduction pour *$\boldsymbol{X}$, notée $\boldsymbol{D}_{1/\boldsymbol{\sigma}}$, est définie par :*

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

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

De plus, la matrice

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

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

La réduction des données $\boldsymbol{X}$ 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ une matrice de données et $\boldsymbol{X}_{0,1}$ sa version centrée réduite. Alors, la matrice*

$$\boldsymbol{R} = \frac{1}{m}\boldsymbol{X}_{0,1}^{\mathrm{T}}\boldsymbol{X}_{0,1}. \tag{2.2.7}$$

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

Lorsque $[\boldsymbol{R}]_{ij} = 1$, cela signifie qu'il existe une relation de dépendance affine entre $\boldsymbol{v}_i$ et $\boldsymbol{v}_j$, 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 ACP[2]. 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 $\{\boldsymbol{x}_i\}_{i=1}^m \subset \mathbb{R}^n$ (que l'on considèrera comme précédemment sous la forme $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}$), 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 $\bar{\boldsymbol{x}}$. L'équation de cette droite sera donc $\{\bar{\boldsymbol{x}} + t\boldsymbol{u} | t \in \mathbb{R}\}$, où $\boldsymbol{u}$ 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 $\boldsymbol{y}_i$ la projection du vecteur $\boldsymbol{x}_i$ sur la droite, cela signifie que l'on souhaite que $\boldsymbol{y}_i$ et $\boldsymbol{x}_i$ soient les plus proches possibles. Le problème d'optimisation sous-jacent est donc

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

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

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*

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

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

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.

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

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

$$\boldsymbol{y}_i = c_i\,\boldsymbol{u} + \bar{\boldsymbol{x}}, \qquad c_i = \boldsymbol{u}^{\mathrm{T}}\,(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\,. \tag{2.3.3}$$

On dispose alors des propriétés suivantes.

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

i) On a $\bar{\boldsymbol{y}} = \bar{\boldsymbol{x}}$;

ii) $c_i\boldsymbol{u} = \boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{u} = \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, où $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ est une matrice de projection orthogonale;

iii) $\boldsymbol{Y} = (\boldsymbol{X} - \mathbf{1}_m\bar{\boldsymbol{x}}^{\mathrm{T}})\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} + \mathbf{1}_m\bar{\boldsymbol{x}}^{\mathrm{T}}$.

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

On notera que les notions de composante, axe et vecteur principaux ne dépendent que des $\boldsymbol{x}_i$.

Calcul explicite de la composante principale

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

$$\max_{\boldsymbol{u}\in\mathbb{R}^n, \|\boldsymbol{u}\|=1} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{u}. \tag{2.3.4}$$

Le problème (2.3.4) est un problème très classique de mathématiques appliquées : il est lié au calcul des vecteurs propres de la matrice $\boldsymbol{\Sigma}$, 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 $\boldsymbol{\Sigma}$.*

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 considèrant 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

$$\bar{\boldsymbol{x}} + \mathcal{S}_k, \qquad \mathcal{S}_k := \mathrm{vect}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k\}, \tag{2.3.5}$$

où $\mathcal{S}$ est un sous-espace vectoriel de dimension k, et $\boldsymbol{U}_k = [\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_k] \in \mathbb{R}^{n\times k}$ est une base orthonormale du sous-espace $\mathcal{S}$ (c'est-à-dire $\boldsymbol{U}_k^{\mathrm{T}}\boldsymbol{U}_k = \boldsymbol{I}_k$, où $\boldsymbol{I}_k$ 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

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

où $\boldsymbol{y}_i$ représente la projection de $\boldsymbol{x}_i$ sur l'espace $\bar{\boldsymbol{x}} + \mathcal{S}_k$. 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*

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

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

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 $\boldsymbol{U}_k$ et passant par le vecteur $\bar{\boldsymbol{x}}$. Alors, pour tout $i = 1, \ldots, m$, la projection de $\boldsymbol{x}_i$ sur ce sous-espace est donné par :*

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

Cette propriété conduit à la définition des composantes principales.

Définition 2.10 (Composantes principales) *Les vecteurs $\boldsymbol{c}_i$, qui représentent les coordonnées des $\boldsymbol{y}_i$ dans le repère $(\bar{\boldsymbol{x}}; \boldsymbol{u}_1, \ldots, \boldsymbol{u}_k)$, sont appelées les ($k$ premières) composantes principales. Les vecteurs $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ 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 $\boldsymbol{U}_k$ la solution du problème* (2.3.6) *en $k < n$ composantes principales. Alors, le problème*

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

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

$$\max_{\substack{\boldsymbol{u}\in\mathbb{R}^n, \|\boldsymbol{u}\|=1 \\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}_1=0 \\ \cdots \\ \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}_k=0}} \mathcal{I}(\boldsymbol{Z}), \tag{2.3.9}$$

où $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{z}_m^{\mathrm{T}} \end{bmatrix}$ *désigne l'ensemble des projections des points* $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$.

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* $\boldsymbol{\Sigma}$ *associés aux* k *plus grandes valeurs propres de* $\boldsymbol{\Sigma}$*, que l'on note* $\lambda_1 \geq \cdots \geq \lambda_k$.

On notera qu'en cas de valeurs propres multiples, l'ensemble des solutions peut être ambigü (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 $\boldsymbol{y}_i$ en tant qu'approximations des $\boldsymbol{x}_i$.

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

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

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

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

iii) Si $k = n$*, alors* $\boldsymbol{y}_i = \boldsymbol{x}_i$ *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 $\{\boldsymbol{0}; \boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$, où $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ sont les vecteurs de la base canonique de $\mathbb{R}^n$, au repère $\{\bar{\boldsymbol{x}}; \boldsymbol{u}_1, \dots, \boldsymbol{u}_k, \dots, \boldsymbol{u}_n\}$ (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 $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ à une matrice de composantes principales $\boldsymbol{C} = \left[\boldsymbol{c}_i^{\mathrm{T}}\right]_{i=1,\dots,m} \in \mathbb{R}^{m \times k}$, qui est donc moins coûteuse à stocker en mémoire;

iii) Un centrage des données, car les composantes forment un jeu de données centrées :

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

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 :

$$\boldsymbol{\Sigma}' = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k),$$

où $\lambda_1 \geq \cdots \geq \lambda_k$ sont les k plus grandes valeurs propres de $\boldsymbol{\Sigma}$.

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 commes 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 caractéristiques[1] sous la forme d'un vecteur de $\mathbb{R}^n$. Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ ces vecteurs : on les regroupe alors sous la forme d'une matrice de données

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m \times n}. \tag{3.1.1}$$

Exemple 3.1

- *Chaque ligne de $\boldsymbol{X}$ représente un individu, et les n composantes de $\boldsymbol{x}_i$ sont des données médicales (âge, poids, taux de cholestérol, etc).*
- *Chaque ligne de $\boldsymbol{X}$ est une "vectorisation" d'une image 2D, et les valeurs de $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i$ 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 $\boldsymbol{y} \in \mathbb{R}^m$.

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

[1] Ou *features* en anglais.

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

$$h(\boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \boldsymbol{y}_i \qquad \forall i = 1, \dots, m,$$

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

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

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 $\beta_1, \dots, \beta_n$ est donné par*

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

ou, sous forme compacte,

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

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

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

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

où $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, $\boldsymbol{y} \in \mathbb{R}^n$, 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 $\boldsymbol{X}$ et du vecteur $\boldsymbol{y}$, comme le montrent les exemples ci-dessous.

Exemple 3.2 *a) Le système*

$$\left\{ \begin{array}{rcl} \beta_1 + \beta_2 & = & 0, \\ 3\beta_1 + 2\beta_2 & = & 1. \end{array} \right.$$

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

b) Le système

$$\left\{ \begin{array}{rcl} \beta_1 & = & 0, \\ \beta_1 & = & -1. \end{array} \right.$$

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 $\boldsymbol{X}$ inversible : dans cette situation, il existe une caractérisation de la solution du système.

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

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

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

3.2.2 Cas d'un système rectangulaire

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

$$\boldsymbol{X\beta} = \boldsymbol{y}, \qquad \boldsymbol{X} \in \mathbb{R}^{m\times n},\ m \neq n. \tag{3.2.2}$$

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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ est de* **rang plein** *si* $\mathrm{rang}(\boldsymbol{X}) = \min\{m, n\}$.

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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ de rang plein et $\boldsymbol{y} \in \mathbb{R}^m$ avec $m \neq n$. On a :*

a) Si $\operatorname{rang}(\boldsymbol{X}) = n$, alors $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ est inversible et le système (3.2.2) possède une unique solution donnée par $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ lorsque $\boldsymbol{y} \in \operatorname{Im}(\boldsymbol{X})$;

b) Si $\operatorname{rang}(\boldsymbol{X}) = m$, alors $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ est inversible et le vecteur $\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{y}$ est solution de (3.2.2)

Il est donc toujours possible de déterminer une solution au problème (3.2.2) lorsque $\boldsymbol{X}$ est de rang plein. Lorsque $\boldsymbol{X}$ 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-inverse[2] que l'on décrit ci-dessous.

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

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

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

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 $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times n}$ 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 $\boldsymbol{\Sigma}$ est la matrice $\boldsymbol{\Sigma}^{\dagger} \in \mathbb{R}^{n\times m}$ définie par

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

[2] On 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$ une décomposition en valeurs singulières de cette matrice. Alors, la pseudo-inverse de $\boldsymbol{X}$ est donnée par*

$$\boldsymbol{X}^{\dagger} = \boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}\boldsymbol{U}^{\mathrm{T}}. \tag{3.2.3}$$

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 montrer le corollaire suivant.

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

i) Si $\mathrm{rang}(\boldsymbol{X}) = m$, alors $\boldsymbol{X}^{\dagger} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}$;

ii) Si $\mathrm{rang}(\boldsymbol{X}) = n$, alors $\boldsymbol{X}^{\dagger} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$;

iii) Si $\mathrm{rang}(\boldsymbol{X}) = n$ et que $n = m$, alors $\boldsymbol{X}$ est carrée inversible, et les deux formules ci-dessus correspondent à $\boldsymbol{X}^{\dagger} = \boldsymbol{X}^{-1}$.

La pseudo-inverse est ainsi apparue dans les sections 3.2.1 et 3.2.2 lorsque l'on supposait que la matrice $\boldsymbol{X}$ 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^{m}$, on ne cherche plus à résoudre $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ de manière exacte, mais plutôt à minimiser l'écart entre les vecteurs $\boldsymbol{X}\boldsymbol{\beta}$ et $\boldsymbol{y}$. 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 :

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

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

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

Cet objectif que l'on cherche à réduire représente donc bien l'attache 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^m$. L'ensemble des vecteurs $\boldsymbol{\beta} \in \mathbb{R}^n$ tels que la valeur de $\frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$ soit minimale, que l'on note :*

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2, \tag{3.3.2}$$

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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^m$. L'ensemble des vecteurs $\boldsymbol{\beta} \in \mathbb{R}^n$ tels que $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$ 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^m$, les propriétés suivantes sont vérifiées :*

1. *l'ensemble défini par (3.3.2) est toujours non vide;*
2. *le vecteur $\boldsymbol{X}^\dagger\boldsymbol{y}$ est une solution au sens des moindres carrés;*
3. *parmi toutes les solutions au sens des moindres carrés, le vecteur $\boldsymbol{X}^\dagger\boldsymbol{y}$ est la solution de norme minimale :*
$$\forall\, \boldsymbol{\gamma} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2, \quad \|\boldsymbol{\gamma}\| \geq \|\boldsymbol{X}^\dagger\boldsymbol{y}\|.$$

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.

[3] Pour 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 $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, $\boldsymbol{y} \in \mathbb{R}^m$ et $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\dagger \boldsymbol{y}$.*

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

- *a) Si $\operatorname{rang}(\boldsymbol{X}) = m = n$, on a $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{-1}\boldsymbol{y}$, et il s'agit de l'unique solution au sens classique et au sens des moindres carrés;*
- *b) Si $\operatorname{rang}(\boldsymbol{X}) < m$ et $\boldsymbol{y} \in \operatorname{Im}(\boldsymbol{X})$, alors $\hat{\boldsymbol{\beta}}$ 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 $\operatorname{rang}(\boldsymbol{X}) < m$ et $\boldsymbol{y} \notin \operatorname{Im}(\boldsymbol{X})$, alors il n'existe pas de solution au sens classique; en revanche, $\hat{\boldsymbol{\beta}}$ est la solution de norme minimale au sens des moindres carrés.*

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

- *b) Si $\operatorname{rang}(\boldsymbol{X}) = m$, alors $\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{y}$ 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 $\operatorname{rang}(\boldsymbol{X}) < m$ et $\boldsymbol{y} \in \operatorname{Im}(\boldsymbol{X})$, alors $\hat{\boldsymbol{\beta}}$ 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 $\operatorname{rang}(\boldsymbol{X}) < m$ et $\boldsymbol{y} \notin \operatorname{Im}(\boldsymbol{X})$, alors il n'existe pas de solution au sens classique; en revanche, $\hat{\boldsymbol{\beta}}$ est la solution de norme minimale au sens des moindres carrés.*

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

- *a) Si $\operatorname{rang}(\boldsymbol{X}) = n$, alors $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ et il s'agit de l'unique solution au sens des moindres carrés. C'est une solution au sens classique lorsque $\boldsymbol{y} \in \operatorname{Im}(\boldsymbol{X})$.*
- *b) Si $\operatorname{rang}(\boldsymbol{X}) < n$ et $\boldsymbol{y} \in \operatorname{Im}(\boldsymbol{X})$, alors $\hat{\boldsymbol{\beta}}$ 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 $\operatorname{rang}(\boldsymbol{X}) < n$ et $\boldsymbol{y} \notin \operatorname{Im}(\boldsymbol{X})$, alors il n'existe pas de solution au sens classique; en revanche, $\hat{\boldsymbol{\beta}}$ 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

Nous abordons maintenant le cadre principal de construction du modèle linéaire, celui de la régression linéaire. Dans ce contexte, nous considèrerons tout d'abord la régression linéaire dite **simple**, où le modèle est défini par un seul paramètre réel, puis celui de la régression linéaire **multiple**, où on cherchera à calculer un vecteur de paramètres pour obtenir le modèle.

4.1 Introduction (d'aléatoire)

Dans le chapitre précédent, nous avons construit des modèles linéaires à partir de données dont nous avons supposé qu'elles étaient déterministes. Dans la pratique, on dispose généralement de données bruitées, soit aléatoirement, soit numériquement (par l'utilisation de simulateurs, notamment), ce qui peut engendrer des soucis de cohérence de modèle. Par ailleurs, si les données proviennent d'une distribution, il peut être intéressant de considérer le modèle obtenu comme aléatoire, et d'en étudier la distribution. C'est ce que l'on se propose de réaliser dans la suite.

À cet effet, le présent chapitre fournit quelques rappels de probabilités, puis pose les bases de l'estimation statistique : la construction de modèles linéaires pourra être considérée dans ce cadre, qui correspondra alors à une situation d'apprentissage statistique.

4.2 Éléments de statistiques

La notion de probabilité découle de la théorie de la mesure. Nos résultats reposeront ainsi sur la notion d'*espace probabilisé*, c'est-à-dire de triplets $(\Omega, \mathcal{A}, \mathbb{P})$, où :

- Ω est un ensemble de valeurs appelé univers;
- $\mathcal{A}$ est une famille de parties de Ω appelée ensemble des évènements[1];
- $\mathbb{P} : \mathcal{A} \to [0, 1]$ est une mesure de probabilité, qui vérifie notamment $\mathbb{P}(\emptyset) = 0$ et $\mathbb{P}(\Omega) = 1$.

Une variable aléatoire est alors une fonction d'un espace probabilisé vers un autre espace qui y induit une nouvelle mesure de probabilité. La notion de variable aléatoire est généralement scalaire : on parlera plutôt de vecteur aléatoire dans le cas d'un espace vectoriel.

Pour les besoins de ce cours, on considèrera principalement deux types de quantités aléatoires :

[1] On notera que $\mathcal{A}$ doit satisfaire certaines propriétés qui en font une tribu, ou σ-algèbre.

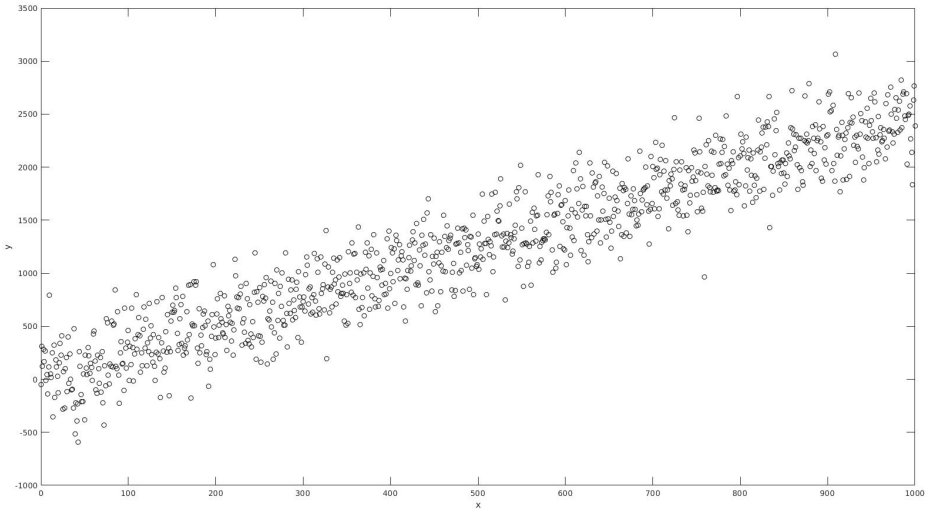

Figure 4.1: Données bruitées générées à partir d'une fonction linéaire auquel on a ajouté un bruit gaussien.

- les variables aléatoires z sur un espace probabilisé $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$ par

$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \mathbb{P}(z \in B) = \mathbb{P}(B);$$

- les vecteurs aléatoires $\boldsymbol{z} = \begin{bmatrix} z_1 \\ \cdots \\ z_m \end{bmatrix}$ de taille m, définis sur l'espace probabilisé $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), \mathbb{P})$.

Dans les deux cas, l'ensemble des évènements considéré sera la tribu de Borel $\mathcal{B}(\mathbb{R}^n)$.

4.2.1 Variables aléatoires

L'étude des variables aléatoires peut s'effectuer en considérant qu'elles peuvent prendre un continuum de valeurs. Nous ferons cependant la distinction entre variables dites discrètes et variables dites continues.

Définition 4.1 (Variable aléatoire discrète) *Une variable aléatoire discrète est définie par :*

- *L'ensemble discret des valeurs possibles $\mathcal{Z} = \{z_i\}$;*
- *L'ensemble des probabilités associées $p = \{p_i\}$, où $\sum_i p_i = 1$ et*

$$\forall \mathcal{S} \subset \mathcal{Z}, \quad \mathbb{P}(z \in \mathcal{S}) = \sum_{z_i \in \mathcal{S}} p_i.$$

Définition 4.2 (Variable aléatoire continue) *Une variable aléatoire continue est définie par :*

- *L'ensemble continu des valeurs possibles $\mathcal{Z} \subset \mathbb{R}$;*

- *La densité de probabilité associée* $p : \mathcal{Z} \to \mathbb{R}$, *où* $\int_{\mathbb{R}} p(z)\,dz = 1$ *et*

$$\forall \mathcal{S} \subset \mathcal{Z}, \quad \mathbb{P}(z \in \mathcal{S}) = \int_{z \in \mathcal{S}} p(z)\,dz.$$

Définition 4.3 (Espérance/Moyenne) *Soit* z *une variable aléatoire réelle distribuée selon* p, *que l'on note* $z \sim p$. *L'***espérance de** z *est définie par*

$$\mathbb{E}[z] = \mathbb{E}_z[z] = \begin{cases} \sum_{z_i \in \mathcal{Z}} z_i\, p(z = z_i) & \textit{(cas discret)} \\ \int_{\mathcal{Z}} z\, p(z)\,dz & \textit{(cas continu).} \end{cases}$$

L'espérance possède plusieurs propriétés qui en facilitent l'utilisation, au premier des rangs desquelles celle ci-dessous.

Proposition 4.1 *L'espérance est linéaire, c'est-à-dire que pour toute variable aléatoire* z *et tous réels* α, β, *on a :*

$$\mathbb{E}[\alpha z + \beta] = \alpha\,\mathbb{E}[z] + \beta;$$

Définition 4.4 (Variance et écart-type) *Soit* z *une variable aléatoire réelle.*

- *La variance de* z *est définie par*

$$\mathrm{Var}[z] = \mathbb{E}\left[z^2\right] - \mathbb{E}[z]^2.$$

- *L'écart-type de* z *est la racine carrée de la variance, qui est toujours positive.*

Dans certains cas, la variance possède une expression simplifiée.

Lemme 4.1

- *Si* z *suit une distribution discrète, alors* $\mathrm{Var}[z] = \sum_i p_i z_i^2 - [\sum_i p_i z_i]^2$;
- *Si* $\mathbb{E}[z] = 0$, *alors* $\mathrm{Var}[z] = \mathbb{E}\left[z^2\right]$.

4.2.2 Couple de variables aléatoires

Lorsque deux variables aléatoires suivent la même loi sur le même espace probabilisé, on parle de variables **identiquement distribuées**. Dans le cas général, on peut s'intéresser à la distribution du couple formé par deux variables aléatoires.

Définition 4.5 (Loi jointe (cas discret)) *Soient deux variables aléatoires discrètes* z *et* w, *respectivement à valeurs dans* $\mathcal{Z} = \{z_i\}$ *et* $\mathcal{W} = \{w_j\}$. *La variable aléatoire* (z, w) *est définie par :*

- *L'ensemble de définition* $\mathcal{Z} \times \mathcal{W} = \{(z_i, w_j)\}$;
- *La loi de probabilité discrète* $\{p_{i,j}\}$, *avec :*

$$p_{i,j} = \mathbb{P}(z = z_i, w = w_j).$$

Définition 4.6 (Loi jointe (cas continu)) *Soient deux variables aléatoires continues z et w respectivement à valeurs dans $\mathcal{Z}$ et $\mathcal{W}$. La variable aléatoire (z, w) est définie par :*

- *L'ensemble de définition $\mathcal{Z} \times \mathcal{W}$;*
- *la loi de probabilité $p : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ telle que*

$$\int_z \int_w p(z, w)\, dz\, dw \;=\; 1.$$

Dans les définitions précédentes, on est partis de deux variables aléatoires pour définir la loi jointe du couple de variables aléatoires. On peut aussi effectuer le chemin inverse.

Définition 4.7 (Lois marginales (cas discret)) *Soient z et w deux variables aléatoires discrètes à valeurs respectives dans $\mathcal{Z} = \{z_i\}$ et $\mathcal{W} = \{w_j\}$. Soit $\{p_{i,j}\}$ la loi jointe de (z, w).*

- *La loi marginale de z est donnée par*

$$\mathbb{P}(z = z_i) = \sum_{j | w_j \in \mathcal{W}} \mathbb{P}(z = z_i,\ w = w_j) = \sum_j p_{i,j} := p_{i\bullet}.$$

- *On définit de la même manière la loi marginale $\{p_{\bullet j}\}_j$ de w.*

Définition 4.8 (Lois marginales (cas continu)) *Soient z et w deux variables aléatoires continues à valeurs respectives dans $\mathcal{Z}$ et $\mathcal{W}$. Soit $p : (z, w) \mapsto p(z, w)$ la loi jointe de (z, w).*

- *La loi marginale de z, notée p_z ou $p(z, \bullet)$, est la fonction $p_z : \mathcal{Z} \to \mathbb{R}^+$ définie par*

$$\forall z \in \mathcal{Z}, \quad p_z(z) = \int_{\mathcal{W}} p(z, w)\, dw.$$

- *On définit de même $p_w : \mathcal{W} \to \mathbb{R}^+$.*

Définition 4.9 (Covariance) *Soient z et w deux variables aléatoires réelles. La covariance de z et w est définie par:*

$$\mathrm{Cov}[z, w] = \mathbb{E}_{z,w}\left[(z - \mathbb{E}[z])\,(w - \mathbb{E}[w])\right].$$

Définition 4.10 (Corrélation) *Soient z et w deux variables aléatoires réelles. La corrélation de z et w est définie par*

$$\mathrm{Corr}[z, w] = \frac{\mathrm{Cov}[z, w]}{\sqrt{\mathrm{Var}_z[z]}\sqrt{\mathrm{Var}_w[w]}}.$$

Variables indépendantes La notion d'indépendance joue un rôle majeur dans l'obtention de résultats de statistiques. Elle est souvent combinée avec la notion de variables identiquement distribuées: on parlera alors de variables **i.i.d.**, pour indépendantes et identiquement distribuées.

Définition 4.11 (Variables indépendantes) *Soient deux variables aléatoires z et w d'ensembles de valeurs $\mathcal{Z}$ et $\mathcal{W}$, et de densités de probabilité p_z et p_w. On dit que z et w sont* **indépendantes** *si la variable aléatoire produit (z, w) vérifie :*

$$\forall \mathcal{S} \times \mathcal{T} \subset \mathcal{Z} \times \mathcal{W}, \quad \mathbb{P}(z \in \mathcal{S}, w \in \mathcal{T}) \;=\; \mathbb{P}(z \in \mathcal{S})\, \mathbb{P}(w \in \mathcal{T}).$$

Proposition 4.2 *Soient deux variables aléatoires z et w indépendantes. Alors, leur loi jointe est le produit des lois marginales : on a donc*

$$\begin{cases} p_{ij} = p_{i\bullet} \times p_{\bullet j} & \text{(cas discret)} \\ p(z,w) = p_z(z) \times p_w(w) & \text{(cas continu).} \end{cases}$$

Proposition 4.3 *Soient deux variables aléatoires z et w indépendantes. Alors, ces deux variables sont décorrélées, c'est-à-dire que* $\mathrm{Cov}\,[z,w] = \mathrm{Corr}\,[z,w] = 0$.

4.2.3 Statistique multidimensionnelle

La plupart des résultats précédents sur les variables aléatoires peuvent être étendus au cas des **vecteurs aléatoires**, c'est-à-dire des quantités aléatoires multidimensionnelles. Sans être exhaustifs, nous donnons ci-dessous les concepts de base.

Définition 4.12 (Loi d'un vecteur aléatoire) *Soit $\boldsymbol{z} = [z_i]_i$ un vecteur aléatoire de $\mathbb{R}^n$: la loi de $\boldsymbol{z}$ est définie par la loi jointe de ses composantes. En particulier, on définit les moments de ces distributions suivants :*

- *l'espérance de $\boldsymbol{z}$ est définie comme le vecteur des espérances :*

$$\mathbb{E}\,[\boldsymbol{z}] = \{\mathbb{E}\,[z_i]\}_i \in \mathbb{R}^n;$$

- *la* **matrice de covariance** *de $\boldsymbol{z}$, notée* $\mathrm{Var}\,[\boldsymbol{z}]$ *ou* $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ *:*

$$\forall 1 \leq i,j \leq n, \quad [\boldsymbol{\Sigma}_{\boldsymbol{z}}]_{i,j} := \mathbb{E}\,[(z_i - \mathbb{E}\,[z_i])(z_j - \mathbb{E}\,[z_j])]\,.$$

On remarque donc que la variance du vecteur $\boldsymbol{z}$ est la matrice des covariances de ses composantes. On note également que

$$\boldsymbol{\Sigma}_{\boldsymbol{z}} = \mathbb{E}\left[(\boldsymbol{z} - \mathbb{E}\,[\boldsymbol{z}])(\boldsymbol{z} - \mathbb{E}\,[\boldsymbol{z}])^{\mathrm{T}}\right] \in \mathbb{R}^{n\times n}.$$

Définition 4.13 (Vraisemblance) *Soient $z_1, z_2, \ldots, z_m$ des variables aléatoires dont on suppose que leur loi dépend d'un paramètre θ. La* **vraisemblance**[2] *est la densité correspondant à la loi jointe de $z_1, \ldots, z_m$, que l'on note $L(z_1, \ldots, z_m; \theta)$.*

La vraisemblance décrit donc, étant donné une valeur de paramètre θ (qui n'est pas forcément la vraie valeur du paramètre de la loi), quelle est la loi la plus vraisemblable pour les $z_1, \ldots, z_m$. Elle peut servir à caractériser une loi de probabilité dépendant d'un paramètre inconnu.

Avant cela, on donne une propriété très utile de la vraisemblance, qui est une simple conséquence de sa définition ainsi que des résultats sur l'indépendance de la section 4.2.2.

Lemme 4.2 *Soient $z_1, \ldots, z_m$ des variables aléatoires i.i.d. dont on suppose que leur loi dépend d'un paramètre θ. On notera cette loi $p(\cdot, \theta)$; alors,*

$$L(z_1, \ldots, z_m; \theta) = \prod_{i=1}^{m} p_i(z_i; \theta) = \prod_{i=1}^{m} p(z_i; \theta).$$

[2] Ou *likelihood* en anglais.

4.3 Régression linéaire simple

Dans le cadre de la régression linéaire simple, nous supposons que nos données de travail sont fournies sous la forme

$$\boldsymbol{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \in \mathbb{R}^{m\times 1}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ v_m \end{bmatrix} \in \mathbb{R}^{m},$$

où $m \geq 1$. On va donc chercher à construire un modèle linéaire $h : \mathbb{R} \to \mathbb{R}$ qui explique les données. Celui-ci sera de la forme

$$h(x) = x\beta \quad \text{avec} \quad \beta \in \mathbb{R},$$

de sorte que le système d'équations $h(x_i) = y_i,\ i = 1, \ldots, m$ corresponde au système linéaire $\boldsymbol{X}\beta = \boldsymbol{y}$.

Pour introduire des éléments statistiques dans la construction du modèle, on considèrera que les éléments de $\boldsymbol{X}$ sont fixés, mais que les composantes de $\boldsymbol{y}$ correspondent à des observations d'un modèle linéaire entaché d'aléatoire. Plus formellement, on considère que les données $\boldsymbol{X}$ et $\boldsymbol{y}$ vérifient

$$\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{\epsilon} \iff y_i = h(x_i) + \epsilon_i\ \forall i, \tag{4.3.1}$$

où $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^m \in \mathbb{R}^m$ est un vecteur aléatoire représentant l'erreur de modèle.

L'approche de la régression linéaire consiste donc à modéliser l'erreur $y_i - x_i\beta^*$ par une variable aléatoire. On cherche idéalement à calculer β^*, ou une approximation de cette quantité. Comme on va le voir, il existe deux approches possibles pour traiter ce problème, qui se basent sur ce que nous avons vu aux chapitres précédents.

4.3.1 Approche par moindres carrés

L'approche **par moindres carrés** du modèle linéaire consiste à ne pas exploiter le caractère aléatoire de l'erreur, mais plutôt à minimiser l'erreur entre le modèle $h(x_i)$ et l'observation y_i pour tout $i = 1, \ldots, m$. Pour ce faire, on construit la fonctionnelle aux moindres carrés

$$f(\beta) = \frac{1}{2}\|\boldsymbol{X}\beta - \boldsymbol{y}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(x_i\beta - y_i)^2.$$

Le problème aux moindres carrés correspondant est ainsi

$$\min_{\beta\in\mathbb{R}} f(\beta). \tag{4.3.2}$$

On cherche ainsi à obtenir le modèle h ou, de manière équivalente, le réel β, tel que l'erreur est minimale.

Remarque 4.1 *En utilisant* (4.3.1)*, on remarque que*

$$f(\beta^*) = \frac{1}{2}\sum_{i=1}^{n}\epsilon_i^2.$$

On peut donc caractériser l'erreur que produit l'utilisation du véritable modèle (parfois appelé vérité terrain, de l'anglais ground truth*). Il n'est cependant pas garanti qu'il s'agisse de la solution du problème aux moindres carrés. En effet, s'il se trouve qu'il existe β tel que $\boldsymbol{X}\beta = \boldsymbol{y}$, la résolution du problème aux moindres carrés renverra cette valeur.*

Avec ces définitions, on considère une solution du problème aux moindres carrés, auquel on donne le nom d'estimateur des moindres carrés linéaires.

Définition 4.14 (Estimateur des moindres carrés ordinaires) *Pour un problème de régression linéaire simple posé sur des données $\boldsymbol{X} \in \mathbb{R}^{m\times 1}$ et $\boldsymbol{y} \in \mathbb{R}^m$, où $\boldsymbol{y}$ est obtenu via (4.3.1), l'***estimateur des moindres carrés ordinaires***, noté $\hat{\beta}^{OLS}$, est défini comme la solution de norme minimale au problème*

$$\min_{\beta\in\mathbb{R}} \frac{1}{2}\|\boldsymbol{X}\beta - \boldsymbol{y}\|^2. \tag{4.3.3}$$

D'après les résultats du chapitre 3, on sait qu'il existe deux possibilités de résolution du problème. Celles-ci sont rappelées dans le théorème ci-dessous.

Théorème 4.1 (Calcul de l'estimateur des moindres carrés ordinaires) *On considère le problème de régression linéaire simple posé sur des données $\boldsymbol{X} \in \mathbb{R}^{m\times 1}$ et $\boldsymbol{y} \in \mathbb{R}^m$, où $\boldsymbol{y}$ est obtenu via (4.3.1). Alors,*

a) Si $\boldsymbol{X} = \mathbf{0}_m$, alors $\hat{\beta}^{OLS} = 0$;

b) Si $\operatorname{rang}(\boldsymbol{X}) = \min\{m, 1\} = 1$, alors

$$\hat{\beta}^{OLS} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \frac{\sum_{i=1}^m x_i y_i}{\sum_{i=1}^m x_i^2}. \tag{4.3.4}$$

Comme on le voit, l'estimateur des moindres carrés ordinaires peut être calculé directement en faisant appel aux résultats du chapitre 3. Cet estimateur ignore la présence d'aléatoire que mettait en avant la formulation (4.3.1). Pour cette raison, et même si cet estimateur possède de bonnes propriétés algébriques (il est optimal au sens du problème (4.3.2), il est de norme minimale parmi l'ensemble des solutions), on ne peut pas le munir a priori de garanties **statistiques**. En revanche, le calcul de $\hat{\beta}^{OLS}$ ne requiert pas d'hypothèses sur la distribution de ces erreurs, ce qui en fait un outil intéressant dans le cas d'une régression linéaire générique.

4.3.2 Approche par maximum de vraisemblance

On se place toujours dans le contexte de données $\boldsymbol{X} \in \mathbb{R}^{m\times 1}$ et $\boldsymbol{y} \in \mathbb{R}^m$, que l'on souhaite expliquer au moyen d'un modèle linéaire $x \mapsto x\beta$. La seconde approche au problème part de la formule de définition des données y_i, à savoir

$$y_i \;=\; x_i\,\beta^* + \epsilon_i, \quad \forall i = 1, \ldots, m, \tag{4.3.5}$$

et suppose que la loi des erreurs ϵ_i est connue. Sous cette hypothèse, il va s'agir de déterminer β tel que la loi $x_i\beta + \epsilon_i$ soit la plus vraisemblable pour y_i. On ramène donc ce problème à de l'estimation de paramètres, ici du paramètre β^*.

Pour un modèle β fixé, on va donc supposer que $y_i - x_i\beta$ suit la même loi que ϵ_i. Cela nous permet ensuite de caractériser la loi des y_i en fonction de β, et par la suite de former la **vraisemblance** des y_i selon β, c'est-à-dire la loi jointe des y_i en fonction de β. On peut alors définir un estimateur statistique de β comme suit.

Définition 4.15 (Estimateur du maximum de vraisemblance) *Soient des données $\boldsymbol{X} \in \mathbb{R}^{m\times 1}$ et $\boldsymbol{y} \in \mathbb{R}^m$ vérifiant (4.3.5) pour un certain β^* et des erreurs ϵ_i dont on suppose la loi connue.*

*Pour tout $\beta \in \mathbb{R}$, soit $L(y_1, \dots, y_m; \beta)$ la vraisemblance des y_i à β donné. L'**estimateur du maximum de vraisemblance**, noté $\hat{\beta}^{MV}$, est défini comme la solution du problème d'optimisation*[3]

$$\max_{\beta\in\mathbb{R}} \ln\left[L(y_1, \dots, y_m; \beta)\right]. \tag{4.3.6}$$

Nous détaillons dans la section suivante le calcul de cet estimateur dans le cas fondamental d'erreurs i.i.d. gaussiennes.

4.3.3 Calcul explicite dans le cas gaussien

On se place dans les hypothèses suivantes.

Hypothèse 4.1 *On considère des données sous la forme d'une matrice $\boldsymbol{X} \in \mathbb{R}^{m\times 1}$ non nulle fixée et d'un vecteur $\boldsymbol{y} \in \mathbb{R}^m$ généré par*

$$\boldsymbol{y} = \boldsymbol{X}\beta^* + \boldsymbol{\epsilon}, \tag{4.3.7}$$

avec $\beta^* \in \mathbb{R}$ et $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^m$, où les ϵ_i sont des variables aléatoires i.i.d. suivant une loi normale centrée réduite $\mathcal{N}(0, \sigma^2)$.

Sous cette hypothèse, on peut alors caractériser la distribution des variables y_i.

Lemme 4.3 *Sous l'hypothèse 4.1, les variables y_i sont indépendantes, et pour tout $i = 1, \dots, m$, y_i suit une loi normale $\mathcal{N}(x_i\beta^*, \sigma^2)$.*

Notre but est de construire un estimateur pour β^*. Pour ce faire, on suppose que β est tel que $y_i - x_i\beta$ suit une loi normale centrée réduite, et on en tire la formulation de la vraisemblance. Ayant caractérisé la loi de chaque y_i, on peut alors exprimer leur vraisemblance.

Proposition 4.4 *Sous l'hypothèse 4.1, soit $L(y_1, \dots, y_m; \beta)$ la vraisemblance, c'est-à-dire la loi jointe des y_i pour β fixé, avec les hypothèses précédentes. On a :*

$$\begin{aligned} L(y_1, \dots, y_m; \beta) &= \left[\frac{1}{\sqrt{2\pi}}\right]^m \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m (x_i\beta - y_i)^2\right) \\ &= \left[\frac{1}{\sqrt{2\pi}}\right]^m \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{X}\beta - \boldsymbol{y}\|^2\right). \end{aligned}$$

On cherche maintenant à calculer l'estimateur du maximum de vraisemblance. Celui-ci est obtenu comme une solution du problème On cherche ainsi la valeur de β la plus probable au vu des observations.

$$\max_{\beta\in\mathbb{R}} \ln(L(y_1, \dots, y_m; \beta)) = \max_{\beta\in\mathbb{R}} -\frac{1}{2\sigma^2}\|\boldsymbol{X}\beta - \boldsymbol{y}\|_2^2.$$

On procède comme décrit en section 4.3.2, en résolvant

$$\frac{\partial}{\partial\beta} \ln(L(y_1, \dots, y_m; \beta)) = 0 \;\Leftrightarrow\; -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\beta - \boldsymbol{y}) = 0.$$

[3] L'estimateur est parfois noté $\hat{\beta}^{MLE}$, de l'anglais *maximum likelihood estimator*.

Comme $\boldsymbol{X}$ est non nulle par hypothèse, ce système linéaire possède une unique solution, qui est $\boldsymbol{X}^\dagger \boldsymbol{y} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$.

Par ailleurs, on a :

$$\frac{\partial^2}{\partial \beta^2} \ln(L(y_1, \dots, y_m; \boldsymbol{X}^\dagger \boldsymbol{y})) = -\frac{1}{\sigma^2}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = -\frac{1}{\sigma^2}\sum x_i^2 < 0,$$

ce qui garantit que la valeur obtenue est bien un maximum local.

Théorème 4.2 *Sous l'hypothèse 4.1, l'estimateur du maximum de vraisemblance est donné par*

$$\hat{\beta}^{MV} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} = \frac{\sum_{i=1}^m x_i y_i}{\sum_{i=1}^m x_i^2}. \tag{4.3.8}$$

Remarque 4.2 *On retrouve la solution au sens des moindres carrés :* **sous l'hypothèse 4.1 sur la loi des ϵ_i**, *les problèmes aux moindres carrés et le problème de maximisation de la vraisemblance sont donc équivalents, et $\hat{\beta}^{OLS} = \hat{\beta}^{MV}$.*

On peut maintenant étudier les propriétés statistiques de cet estimateur.

Proposition 4.5 *Sous l'hypothèse 4.1, soit $\hat{\beta}^{MV}$ l'estimateur du maximum de vraisemblance. Alors,*

i) $\mathbb{E}_{y_1,\dots,y_m}\left[\hat{\beta}^{MV}\right] = \beta^*$ *(estimateur sans biais);*

ii) $\mathrm{Var}\left[\hat{\beta}^{MV}\right] = \frac{\sigma^2}{\sum_{i=1}^m x_i^2}$ *(estimateur convergent);*

iii) Il s'agit de l'estimateur efficace.

Remarque 4.3 *En règle générale, l'estimateur du maximum de vraisemblance n'est pas forcément unique ou sans biais (mais il sera asymptotiquement sans biais, voire asymptotiquement efficace si cette définition a un sens). De même, dans un cadre non gaussien, les estimateurs des moindres carrés ordinaires et du maximum de vraisemblance ne seront pas nécessairement égaux.*

4.4 Régression linéaire multiple

Dans cette section, on considère maintenant l'apprentissage d'un modèle linéaire multi-dimensionnel, c'est-à-dire paramétré par un vecteur de réels et non plus par un réel. Plus concrètement, on considère une matrice de caractéristiques $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m\times n}$ fixée et un vecteur de labels $\boldsymbol{y} \in \mathbb{R}^m$. Notre but est alors de construire un modèle linéaire $h : \boldsymbol{x} \mapsto \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$ tel que $h(\boldsymbol{x}_i) \approx y_i$ pour tout $i = 1, \dots, m$.

Afin d'adapter les résultats de la section 4.3, nous allons mobiliser des résultats de statistique multi-dimensionnelle, ainsi que d'optimisation en plusieurs variables. Du point de vue modélisation, on considère en effet que le modèle vérifie l'équation :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

où $\boldsymbol{\beta}^* \in \mathbb{R}^n$ et $\boldsymbol{\epsilon} \in \mathbb{R}^m$ est un vecteur aléatoire représentant l'erreur du modèle.

Comme dans le cas de la régression linéaire simple, il existe deux manières d'obtenir un modèle linéaire : soit en formulant un problème aux moindres carrés, soit en calculant l'estimateur du maximum de vraisemblance.

4.4.1 Approche par moindres carrés

Dans l'approche par moindres carrés, on cherche un vecteur $\boldsymbol{\beta} \in \mathbb{R}^n$ solution du problème :

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2.$$

Comme dans le cas de la section 4.3.1, on cherche ainsi la valeur de β qui minimise l'erreur entre y_i et $h(\boldsymbol{x}_i)$ pour chaque i; on minimise ainsi la norme du vecteur des erreurs sans prise en compte d'aléatoire.

On a vu au chapitre 3 qu'il est possible de construire une solution à ce problème pour tous $\boldsymbol{X}$ et $\boldsymbol{y}$. Cette solution sera définie comme la solution du problème aux moindres carrés.

Définition 4.16 (Estimateur des moindres carrés ordinaires) *L'estimateur des moindres carrés ordinaires, noté $\hat{\boldsymbol{\beta}}^{OLS}$, est défini comme la solution de norme minimale du problème*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2. \tag{4.4.1}$$

Il s'agit donc de $\hat{\boldsymbol{\beta}}^{OLS} = \boldsymbol{X}^{\dagger}\boldsymbol{y}$.

Lorsque $\mathrm{rang}(\boldsymbol{X}) = n \leq m$, on a $\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ et il s'agit de l'unique solution du problème aux moindres carrés (cf chapitre 3).

4.4.2 Méthode du maximum de vraisemblance

Dans l'approche du maximum de vraisemblance, on part de

$$y_i \;=\; \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_i, \quad \forall i = 1, \ldots, m, \tag{4.4.2}$$

et on suppose que la loi des erreurs ϵ_i est connue. On cherche ensuite $\boldsymbol{\beta}$ tel que la loi $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$ soit la plus vraisemblable pour y_i. On définit alors l'estimateur suivant.

Définition 4.17 (Estimateur du maximum de vraisemblance) *Soient des données $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ et $\boldsymbol{y} \in \mathbb{R}^m$ vérifiant (4.4.2) pour un certain $\boldsymbol{\beta}^* \in \mathbb{R}^n$ et des erreurs ϵ_i dont on suppose la loi connue. Pour tout $\boldsymbol{\beta} \in \mathbb{R}^n$, soit $L(y_1, \ldots, y_m; \boldsymbol{\beta})$ la vraisemblance des y_i à $\boldsymbol{\beta}$ donné. L'***estimateur du maximum de vraisemblance**, noté $\hat{\boldsymbol{\beta}}^{MV}$, est défini comme la solution du problème d'optimisation*

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \ln\left[L(y_1, \ldots, y_m; \boldsymbol{\beta})\right]. \tag{4.4.3}$$

Remarque 4.4 *En cas de solutions multiples, on choisira une solution avec la plus grande valeur de L, ou éventuellement une solution avec de bonnes propriétés (en termes de biais, variance, norme, etc).*

4.4.3 Calcul explicite dans le cas gaussien

Afin d'illustrer la section précédente, on se place dans un cadre particulier, qui constitue l'exemple de base d'estimateur de la régression linéaire.

Hypothèse 4.2 *On considère des données sous la forme d'une matrice $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ de rang $n \leq m$ et d'un vecteur $\boldsymbol{y} \in \mathbb{R}^m$ généré par*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{4.4.4}$$

avec $\boldsymbol{\beta}^* \in \mathbb{R}^n$ et $\boldsymbol{\epsilon} = [\epsilon_i]_{i=1}^m$, où les ϵ_i sont des variables aléatoires i.i.d. suivant une loi normale centrée réduite $\mathcal{N}(0, \sigma^2)$.

On dira que le vecteur $\boldsymbol{\epsilon}$ est gaussien, c'est-à-dire qu'il suit lui-même une loi normale de moyenne $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ et de matrice de covariance $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \sigma^2\boldsymbol{I}_m$.

Remarque 4.5 *L'hypothèse $n \leq m$ est typique du cas de la régression linéaire, où l'on cherche à construire un modèle linéaire avec plus de points qu'il n'en faudrait. Dans d'autres modèles d'apprentissage (par exemple, les réseaux de neurones), on pourrait avoir plus de paramètres que d'éléments dans le jeu de données.*

Afin de calculer l'estimateur du maximum de vraisemblance, on commence par former la vraisemblance.

Proposition 4.6 *Sous l'hypothèse 4.2, soit $L(y_1, \ldots, y_m; \boldsymbol{\beta})$ la vraisemblance, c'est-à-dire la loi jointe des y_i pour $\boldsymbol{\beta}$ fixé, avec les hypothèses précédentes.On a :*

$$\begin{aligned} L(y_1, \ldots, y_m; \boldsymbol{\beta}) &= \left[\frac{1}{\sqrt{2\pi}}\right]^m \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^m (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2\right) \\ &= \left[\frac{1}{\sqrt{2\pi}}\right]^m \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2\right). \end{aligned}$$

Théorème 4.3 *Sous l'hypothèse 4.2, l'estimateur du maximum de vraisemblance est donné par*

$$\hat{\boldsymbol{\beta}}^{MV} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}. \tag{4.4.5}$$

Remarque 4.6 *On retrouve la solution au sens des moindres carrés : sous les hypothèses 4.2 sur la loi des ϵ_i et le rang de X, les problèmes aux moindres carrés et le problème de maximisation de la vraisemblance sont donc équivalents, et $\hat{\boldsymbol{\beta}}^{OLS} = \hat{\boldsymbol{\beta}}^{MV}$.*

On peut enfin établir les propriétés statistiques de cet estimateur.

Proposition 4.7 *Sous l'hypothèse 4.2, soit $\hat{\beta}^{MV}$ l'estimateur du maximum de vraisemblance. Alors,*

i) $\mathbb{E}_{y_1,\ldots,y_m}\left[\hat{\boldsymbol{\beta}}^{MV}\right] = \boldsymbol{\beta}^*$;

ii) la matrice de covariance de $\hat{\boldsymbol{\beta}}^{MV}$ est donnée par

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^{MV}} = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}.$$

Remarque 4.7 *Nous avons défini les notions d'estimateurs sans biais, convergents et efficaces dans le cadre de paramètres uni-dimensionnels, mais celles-ci s'étendent au cas de paramètres multi-dimensionnels. Une extension triviale de la notion de biais permet de caractériser* $\hat{\boldsymbol{\beta}}^{MV}$ *comme étant sans biais, tandis que la notion d'estimateur convergent correspondante est vérifiée ici car on peut montrer que* $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^{MV}} \to \mathbf{0}$ *quand* m *tend vers* ∞.

4.5 Régression linéaire régularisée

On considère dans cette partie le cadre de la régression linéaire multiple. On suppose donc que l'on dispose d'un jeu de données sous la forme d'une matrice $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{m \times n}$ et d'un vecteur $\boldsymbol{y} = [y_i] \in \mathbb{R}^m$. On travaillera sous l'hypothèse 4.2, en supposant donc qu'il existe $\boldsymbol{\beta}^* \in \mathbb{R}^n$ tel que

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

avec $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ et $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathrm{T}}\right] = \sigma^2 \boldsymbol{I}_n$, où $\sigma > 0$.

Supposons que la matrice $\boldsymbol{X}$ ne soit pas de rang plein, c'est-à-dire $\mathrm{rang}(\boldsymbol{X}) < \min\{m, n\}$. Dans ce cas, on sait qu'il existe une infinité de solutions possibles au sens des moindres carrés, parmi lesquelles $\hat{\boldsymbol{\beta}}^{OLS} = \boldsymbol{X}^{\dagger}\boldsymbol{y}$ est une solution de norme minimale. Cette propriété supplémentaire peut se révéler utile pour interpréter le modèle sous-jacent, voir obtenir un modèle simplifié. On a également vu que le modèle possède de bonnes propriétés statistiques.

On se pose maintenant la question de la meilleure manière de garantir des propriétés, ou une certaine structure, sur le modèle linéaire que l'on cherche à calculer. Comme on va le voir, cela est possible dans le cadre des moindres carrés en considérant un problème régularisé; cela correspond à introduire un a priori sur le maximum de vraisemblance.

4.5.1 Maximum a posteriori

Dans les parties précédentes, nous avons vu que la formulation aux moindres carrés linéaires trouvait une correspondance dans le cadre statistique, via le problème de maximisation de vraisemblance. On va ici décrire l'équivalent de la régularisation, qui correspond à imposer un a priori sur le modèle linéaire.

On rappelle que l'estimateur du maximum de vraisemblance est obtenu en résolvant le problème d'optimisation suivant:

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^n} \ln\left[L(y_1, \ldots, y_m; \boldsymbol{\beta})\right],$$

où L représente la vraisemblance des $y_1, \ldots, y_m$ à $\boldsymbol{\beta}$ donné. L'utilisation de la loi des "y_i sachant $\boldsymbol{\beta}$" nous permet d'exploiter la connaissance de la forme de la loi des y_i; cependant, pour déterminer $\boldsymbol{\beta}$ et sachant que l'on observe les y_i, il semblerait plus judicieux d'optimiser la loi de $\boldsymbol{\beta}$ sachant les y_i, c'est-à-dire la vraisemblance $L(\boldsymbol{\beta}; y_1, \ldots, y_m)$. Or, il est possible de montrer que ces deux lois sont liées par une relation de proportionalité[4] : on a ainsi

$$L(\boldsymbol{\beta}; y_1, \ldots, y_m) \;\propto\; L(y_1, \ldots, y_m; \boldsymbol{\beta}) L(\boldsymbol{\beta}), \tag{4.5.1}$$

[4] On écrira $f(\boldsymbol{\beta}) \propto g(\boldsymbol{\beta})$ lorsqu'il existe une constante c indépendante de $\boldsymbol{\beta}$ telle que $f(\boldsymbol{\beta}) = c g(\boldsymbol{\beta})$.

où $L(\boldsymbol{\beta})$ représente la loi du vecteur $\boldsymbol{\beta}$. Notre souci est précisément que nous ne connaissons pas la loi du vecteur $\boldsymbol{\beta}$ (que l'on voudrait ponctuelle égale à la vraie valeur). On va donc pré-supposer une loi, que l'on appelle un *prior*, ou loi a priori. Partant de cette loi et de la vraisemblance, on va obtenir une formule a posteriori pour l'estimateur.

Définition 4.18 (Estimateur du maximum a posteriori) *On suppose a priori que $\boldsymbol{\beta}$ suit une distribution de loi $L(\boldsymbol{\beta}$.Alors, une solution du problème*

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \ln\left[L(y_1,\dots,y_m;\boldsymbol{\beta})\, L(\boldsymbol{\beta})\right] \tag{4.5.2}$$

est appelée **un estimateur du maximum a posteriori** *avec a priori $L(\boldsymbol{\beta})$, et on la note $\hat{\boldsymbol{\beta}}^{MAP}$.*

4.5.2 Calcul explicite du maximum a posteriori dans le cas gaussien

Dans cette partie, on va supposer un a priori gaussien sur $\boldsymbol{\beta}$, c'est-à-dire que l'on postule que $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\boldsymbol{I}_n)$, avec $\lambda > 0$. La loi a priori de $\boldsymbol{\beta}$ s'écrit donc

$$L(\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi/\lambda}} \exp\left[-\sum_{i=1}^{n} \frac{\beta_i^2}{2/\lambda}\right]. \tag{4.5.3}$$

Pour une valeur de λ suffisamment forte, on peut alors montrer que l'estimateur du maximum a posteriori est unique; notons que si $\lambda = \infty$, alors $\boldsymbol{\beta}$ suit une loi de Dirac en $\mathbf{0}$, et donc l'estimateur du maximum a posteriori est nécessairement nul.

On travaille à nouveau sous l'hypothèse 4.2. Dans ce cadre, la vraisemblance $L(y_1,\dots,y_m;\boldsymbol{\beta})$ s'écrit

$$L(y_1,\dots,y_m;\boldsymbol{\beta}) = \frac{1}{(2\pi)^{m/2}} \exp\left[-\sum_{i=1}^{n} \frac{(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2}{2\sigma^2}\right].$$

Par conséquent, on a :

$$\ln\left[L(y_1,\dots,y_m;\boldsymbol{\beta})L(\boldsymbol{\beta})\right] = \ln\left[\frac{1}{(2\pi)^{(m+1)/2}/\lambda}\right] - \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2}{2\sigma^2} - \frac{\lambda}{2}\sum_{i=1}^{n}\beta_i^2.$$

On cherche ainsi une solution du problème

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^n} \left\{\ln\left[\frac{1}{(2\pi)^{(m+1)/2}/\lambda}\right] - \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2 - \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2\right\}.$$

La première structure que l'on peut vouloir imposer sur un estimateur est que sa valeur soit peu sensible à des variations du jeu de données. En termes statistiques, cela se traduit par une matrice de covariance dont la norme sera plus faible (potentiellement au prix d'un biais non nul); du point de vue algébrique, cela correspondra à restreindre la norme du vecteur représentant le modèle via une modification du problème d'optimisation. On parle ainsi de **régression linéaire avec régularisation écrêtée** (*ridge* en anglais), ou de **régression linéaire avec régularisation ℓ_2**.

Définition 4.19 *On se donne une valeur $\lambda \geq 0$ et on considère le problème aux moindres carrés linéaires avec régularisation ℓ_2 défini par*

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{m}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - y_i)^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\beta_i^2 \;=\; \min_{\boldsymbol{\beta}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2 \tag{4.5.4}$$

Lorsque $\lambda = 0$, on retrouve la définition du problème aux moindres carrés donnée en partie 4.4.1.

Remarque 4.8 *Le problème* (4.5.4) *peut se reformuler de manière équivalente comme un problème aux moindres carrés linéaires avec contraintes sur* $\|\boldsymbol{w}\|$ *: en effet, pour tout* $\lambda \geq 0$*, il existe* $s(\lambda) \in \mathbb{R}^+$ *tel que le problème s'écrive*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 \quad \text{s.c.} \quad \|\boldsymbol{\beta}\|_2^2 \leq s(\lambda).$$

Théorème 4.4 *On considère un problème aux moindres carrés linéaires avec régularisation* ℓ_2 *tel que décrit par la définition 4.19, où on suppose que* $\lambda > 0$*. Alors, la solution du problème est unique : on l'appelle l'estimateur des moindres carrés avec régularisation* ℓ_2*, et elle est donnée par :*

$$\hat{\beta}_2 := (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y}. \tag{4.5.5}$$

A noter que $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}$ est nécessairement inversible car $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \succeq \boldsymbol{0}$ et $\lambda > 0$. Par conséquent, le problème régularisé possède toujours une unique solution, contrairement au problème non régularisé.

Théorème 4.5 *Sous l'hypothèse 4.2 et en postulant un a priori gaussien* $\mathcal{N}(\boldsymbol{0}, \frac{1}{\lambda}\boldsymbol{I})$*, l'estimateur du maximum a posteriori est donné par*

$$\hat{\boldsymbol{\beta}}^{MAP} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{y}. \tag{4.5.6}$$

Remarque 4.9 *On notera que lorsque* $\sigma = 1$*, on retrouve exactement la formule obtenue dans le cadre de la régression* ℓ_2*; de manière plus générale, il y a équivalence entre les deux formulations si l'on postule que* $\boldsymbol{\beta}$ *suit une loi gaussienne* $\mathcal{N}(\boldsymbol{0}, \frac{\sigma^2}{\lambda}\boldsymbol{I})$*, où* $\lambda > 0$ *est le paramètre utilisé dans la régularisation* ℓ_2 *des moindres carrés, et* σ^2 *est la variance des erreurs.*

On peut enfin établir les propriétés statistiques de cet estimateur.

Proposition 4.8 *Sous l'hypothèse 4.2, soit* $\hat{\beta}^{MAP}$ *l'estimateur du maximum a posteriori avec un a priori gaussien* $\mathcal{N}(\boldsymbol{0}, \frac{1}{\lambda}\boldsymbol{I})$*. Alors,*

i) $\mathbb{E}_{y_1,\ldots,y_m}\left[\hat{\boldsymbol{\beta}}^{MAP}\right] = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}^*$*;*

ii) la matrice de covariance de $\hat{\boldsymbol{\beta}}^{MAP}$ *est donnée par*

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^{MAP}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\sigma^2\boldsymbol{I})^{-1}.$$

et on a $\|\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^{MAP}}\| \leq \|\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^{MV}}\|$.

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.