Introduction à la modélisation statistique

Présentation des principes inférentiels et de modèles supervisés usuels.

Paul MINCHELLA, Stéphane CHRÉTIEN

Janvier 2026

Licence & Master MIASHS

Table des matières

Prérequis

⇒ Calculs analytiques. Dérivées, intégrales usuels.
⇒ Calculs matriciels et théorèmes connus (théorème spectral, SVD)
⇒ Calculs probabilistes usuels.
⇒ ...

Chapitre 1

Motivation de la modélisation statistique inférentielle

La statistique inférentielle constitue l'un des piliers fondamentaux de l'analyse des données. Elle vise à tirer des conclusions générales sur une population à partir d'un nombre fini d'observations, nécessairement imparfaites et entachées d'incertitude. Avant d'introduire des outils techniques, il est essentiel de bien comprendre le cadre conceptuel dans lequel elle s'inscrit, ainsi que la nature des questions auxquelles elle cherche à répondre.

1.1 Probabilités et statistiques : deux points de vue complémentaires

Les probabilités et les statistiques reposent sur un socle mathématique commun, mais elles se distinguent par la direction du raisonnement qu'elles adoptent.

Point de vue probabiliste. Dans le cadre probabiliste classique, on suppose que le mécanisme aléatoire est entièrement connu. Plus précisément, on se donne :

— un espace probabilisé,

— une famille de lois de probabilité indexée par un paramètre $\theta \in \Theta$,

et l'on cherche à étudier le comportement d'une variable aléatoire X ou d'un vecteur aléatoire $(X_1, \ldots, X_n)$ lorsque le paramètre θ est fixé.

Autrement dit, le problème probabiliste consiste à répondre à des questions du type :

Sachant que le paramètre vaut θ, comment se répartissent les réalisations possibles de X ?

Point de vue statistique. La statistique adopte une démarche inverse. En pratique, le paramètre θ est inconnu. Ce que l'on observe, ce sont des données issues d'un phénomène réel, que l'on modélise comme des réalisations d'un processus aléatoire. La question centrale devient alors :

À partir des observations disponibles, que peut-on dire des paramètres inconnus du modèle probabiliste sous-jacent ?

C'est précisément ce renversement de perspective qui fonde la statistique inférentielle.

1.2 Variable aléatoire : du hasard à la quantité mesurable

Avant toute chose, on pose la cadre formel indispensable aux probabilités et modélisations statistiques.

Définition 1.2.1 (Variable aléatoire)**.** *Une* variable aléatoire *est une quantité numérique dont la valeur dépend du hasard. Autrement dit, c'est une application qui associe à chaque issue possible d'une expérience aléatoire un nombre réel.*

De manière formelle, on note :

$$T : \Omega \longrightarrow \mathbb{R}$$
$$\omega \longmapsto T(\omega)$$

où :

- Ω *désigne l'ensemble des issues possibles de l'expérience (appelé* univers des possibles*) ;*
- $\omega \in \Omega$ *est une issue particulière (par exemple, un individu précis ou un scénario expérimental) ;*
- $T(\omega)$ *est la valeur numérique observée pour cette issue (par exemple, le temps jusqu'à la panne de cet individu).*

Ainsi, T transforme le résultat aléatoire ω en une valeur réelle mesurable et interprétable.

Le symbole Ω représente le "monde des possibles" : toutes les situations qui pourraient se produire. La variable aléatoire T est une *fonction du hasard* : avant de tirer une issue ω, sa valeur est inconnue. Une fois l'expérience réalisée (une trajectoire observée), $T(\omega)$ devient une valeur concrète, celle que l'on observe dans les données. En analyse de survie, T désigne le temps aléatoire jusqu'à l'évènement d'intérêt (décès, rechute, défaillance, etc.), et chaque sujet de la cohorte correspond à une issue ω_i.

Exemple. Si l'on observe le temps avant la panne d'un appareil, la variable aléatoire T associe à chaque appareil la durée (en jours, mois, etc.) avant sa défaillance. Avant de réaliser l'expérience, la valeur de T est inconnue, mais elle prend certaines valeurs possibles selon un mécanisme probabiliste. L'idée clé est que T n'est pas une valeur fixe, mais une variable dont le résultat est incertain. Elle modélise le hasard sous forme numérique : durée, taille, score, nombre d'essais, etc.

Après avoir introduit le concept de variable aléatoire, voyons maintenant ce que représente *une observation* concrète de cette variable, c'est-à-dire une trajectoire.

1.3 Trajectoire ou observation : la réalisation concrète du hasard

Définition 1.3.1 (Trajectoire ou observation)**.** *Une* trajectoire *(ou* réalisation*) d'une variable aléatoire correspond à la valeur effectivement observée de cette variable pour une issue particulière du hasard.*

Formellement, si $T : \Omega \to \mathbb{R}$ est une variable aléatoire, alors pour une issue donnée $\omega \in \Omega$, la quantité

$$T(\omega)$$

désigne la trajectoire *(ou la* valeur réalisée*) de T associée à cette issue ω.*

Autrement dit, $T(\omega)$ est la valeur concrète que prend la variable aléatoire T lorsque l'expérience produit le résultat ω.

Chaque issue ω du monde des possibles Ω correspond à un individu, un appareil ou un scénario expérimental particulier. Ainsi, lorsque l'on observe un échantillon de n sujets, on observe en réalité n trajectoires :

$$T(\omega_1),\ T(\omega_2),\ \ldots,\ T(\omega_n),$$

que l'on note souvent $T_1, T_2, \ldots, T_n$.

Par exemple, si T représente le temps avant la défaillance d'un appareil, une observation $T_i = 7.3$ signifie que, pour le i-ème appareil (correspondant à l'issue ω_i), la panne est survenue au bout de 7.3 jours. Chaque sujet observé fournit donc une trajectoire particulière du phénomène aléatoire global modélisé par T.

La variable aléatoire représente le *modèle abstrait du hasard*, tandis que la trajectoire correspond à une *donnée concrète*. Autrement dit : T est un concept théorique, tandis que T_i est une observation dans un jeu de données. En statistique, on observe plusieurs réalisations $T_1, T_2, \ldots, T_n$ d'une même variable aléatoire T pour en estimer les caractéristiques (moyenne, médiane, distribution...).

Nous savons maintenant distinguer la variable aléatoire (concept théorique) de son observation (valeur empirique). Il reste à comprendre comment la *loi de probabilité* décrit mathématiquement la manière dont ces valeurs se répartissent.

1.4 Population, échantillon et modélisation

La population statistique. On appelle *population statistique* l'ensemble théorique de toutes les observations possibles associées à un phénomène donné. On la note généralement $\mathcal{P}$. Cette population est décrite par une loi de probabilité inconnue, caractérisée par certains paramètres, par exemple :

— son espérance μ,

— sa variance σ^2.

Ces quantités résument des propriétés globales de la population, mais elles ne sont pas observables directement.

L'échantillon. En pratique, on ne dispose jamais de la population entière. On observe seulement un *échantillon* de taille n, que l'on modélise par des variables aléatoires

$$(X_1, \ldots, X_n),$$

supposées issues du même mécanisme aléatoire. Lorsque l'on réalise effectivement l'expérience, on observe une trajectoire particulière, notée

$$(x_1, \ldots, x_n).$$

Il est fondamental de distinguer :

— les variables aléatoires X_i, objets théoriques,

— leurs réalisations x_i, données concrètes.

1.5 Supervisé, non-supervisé ; paramétrique et non-paramétrique

Les notions introduites dans ce chapitre – variables aléatoires, lois de probabilité, estimation, biais, variance et intervalles de confiance – constituent le socle théorique de l'apprentissage statistique. Avant d'introduire des modèles concrets de régression ou de classification, il est essentiel de clarifier deux distinctions fondamentales qui structurent l'ensemble du domaine : la distinction entre apprentissage supervisé et non supervisé, et celle entre modèles paramétriques et non paramétriques.

1.5.1 Apprentissage supervisé et non supervisé

On se place dans un cadre général où l'on observe des données décrites par des variables aléatoires. Selon la nature de l'information disponible, on distingue deux grandes familles de problèmes.

Apprentissage supervisé. On parle d'apprentissage supervisé lorsque les données observées sont constituées de couples

$$(X_1, Y_1), \ldots, (X_n, Y_n),$$

où X_i représente un vecteur de variables explicatives et Y_i une variable réponse associée. L'objectif est de construire, à partir de ces observations, une règle de prédiction permettant d'estimer la valeur de Y pour une nouvelle observation $X = x$.

Du point de vue probabiliste, l'apprentissage supervisé consiste à exploiter l'information contenue dans la loi jointe de (X, Y) afin d'approximer une quantité conditionnelle d'intérêt, telle que $\mathbb{E}[Y \mid X = x]$ en régression ou $\mathbb{P}(Y = c \mid X = x)$ en classification. Les notions d'estimation, de biais, de variance et de risque, introduites précédemment, prennent ici tout leur sens.

Les exemples typiques de l'apprentissage supervisé incluent la régression linéaire, la régression logistique, les arbres de décision, les forêts aléatoires ou encore l'algorithme des k plus proches voisins.

Apprentissage non supervisé. À l'inverse, on parle d'apprentissage non supervisé lorsque seules les variables explicatives sont observées :

$$X_1, \ldots, X_n,$$

sans variable réponse associée. L'objectif n'est alors pas de prédire une quantité cible, mais de mettre en évidence une structure latente dans les données : regroupements, dépendances, directions principales de variabilité ou représentation plus compacte de l'information.

Dans ce cadre, il n'existe pas de variable Y à estimer, et la démarche inférentielle est de nature différente. Les méthodes non supervisées cherchent à résumer ou organiser les données plutôt qu'à prédire une réponse. Des exemples classiques incluent le clustering, l'analyse en composantes principales ou les méthodes de réduction de dimension.

Cette distinction est fondamentale : un algorithme est dit supervisé non pas en raison de sa complexité ou de sa forme mathématique, mais parce qu'il exploite explicitement une variable réponse observée.

1.5.2 Modèles paramétriques et non paramétriques

Indépendamment du caractère supervisé ou non d'un algorithme, on distingue également les modèles selon la manière dont la relation entre les variables est décrite.

Modèles paramétriques. Un modèle est dit paramétrique lorsqu'il suppose que la relation entre les variables est décrite par une fonction appartenant à une famille paramétrée de dimension finie. Autrement dit, il existe un vecteur de paramètres $\theta \in \mathbb{R}^p$ tel que

$$Y \approx f_\theta(X),$$

et l'apprentissage consiste à estimer le paramètre θ à partir des données.

Les modèles paramétriques reposent sur des hypothèses structurelles fortes, mais offrent en contrepartie une interprétabilité claire et une complexité contrôlée. Les notions de maximum de vraisemblance, d'estimateurs sans biais ou d'atteinte de la borne de Cramér–Rao s'inscrivent naturellement dans ce cadre. La régression linéaire et la régression logistique en sont des exemples emblématiques.

Modèles non paramétriques. Un modèle est dit non paramétrique lorsqu'aucune forme fonctionnelle globale, paramétrée par un nombre fini de paramètres, n'est imposée a priori. La complexité du modèle n'est alors pas fixée indépendamment des données, mais croît avec la taille de l'échantillon ou avec la richesse de la représentation utilisée.

Dans ce cadre, l'apprentissage vise à approximer directement des quantités fonctionnelles, souvent de manière locale. Les algorithmes non paramétriques sont généralement très flexibles et capables de modéliser des relations complexes, au prix d'une sensibilité accrue au bruit, à la dimension et au choix des hyperparamètres. L'algorithme des k plus proches voisins constitue un exemple typique de méthode supervisée non paramétrique.

1.5.3 Synthèse et rôle de l'inférence

Les deux distinctions présentées sont indépendantes mais complémentaires. Un algorithme peut être supervisé ou non supervisé, paramétrique ou non paramétrique. Ces choix conditionnent profondément la manière dont les notions inférentielles introduites dans ce chapitre s'appliquent.

Dans les modèles paramétriques supervisés, l'inférence statistique vise principalement à estimer et interpréter un nombre fini de paramètres. Dans les modèles non paramétriques, elle s'intéresse davantage aux propriétés asymptotiques, à la consistance des estimateurs et aux compromis biais–variance. Enfin, dans les méthodes non supervisées, l'inférence prend souvent la forme d'une analyse exploratoire ou géométrique des données.

Ces distinctions conceptuelles serviront de fil conducteur pour l'étude des modèles de régression, de classification et d'apprentissage automatique abordés dans la suite du cours.

Chapitre 2

Outils probabilistes fondamentaux

2.1 Loi de probabilité : description du comportement du hasard

Définition 2.1.1 (Loi de probabilité d'une variable aléatoire réelle)**.** *La* loi *(ou* distribution*) d'une variable aléatoire décrit la manière dont ses valeurs possibles se répartissent dans l'espace des nombres réels, avec leurs probabilités associées.*

Pour une variable aléatoire continue T, cette loi est décrite par :

- *une* ***fonction de répartition*** $F(t) = \mathbb{P}(T \leq t)$*, qui donne la probabilité que T prenne une valeur inférieure ou égale à t ;*
- *une* ***fonction de survie*** $\mathcal{S}(t) = 1 - F(t) = \mathbb{P}(T > t)$*, utile en analyse de survie ;*
- *une* ***densité de probabilité*** $f(t)$ *telle que $F'(t) = f(t)$ pour presque tout t.*

La loi d'une variable aléatoire est au hasard ce que la *carte de répartition* est à un territoire : elle indique quelles valeurs sont les plus ou les moins probables. Dans l'exemple de la survie, la loi de T décrit la distribution des durées de vie possibles dans la population : certaines pannes arrivent très tôt, d'autres très tard, et cette répartition est caractérisée par $\mathcal{S}$, f ou h.

Proposition 2.1.1 (La loi de probabilité comme mesure)**.** *La loi d'une variable aléatoire réelle définit une* mesure de probabilité *sur l'espace mesurable $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, où $\mathcal{B}(\mathbb{R})$ désigne la tribu borélienne de $\mathbb{R}$. Plus précisément, il existe une application*

$$\mathbb{P}_T : \mathcal{B}(\mathbb{R}) \longrightarrow [0, 1],$$

telle que, pour tout ensemble borélien $A \subset \mathbb{R}$,

$$\mathbb{P}_T(A) = \mathbb{P}(T \in A),$$

et vérifiant les propriétés suivantes :

1. ***Positivité :*** *pour tout $A \in \mathcal{B}(\mathbb{R})$, $\mathbb{P}_T(A) \geq 0$;*
2. ***Normalisation :*** *$\mathbb{P}_T(\mathbb{R}) = 1$;*
3. ***σ-additivité :*** *pour toute famille dénombrable $(A_n)_{n \geq 1}$ d'ensembles deux à deux* ***disjoints*** *de $\mathcal{B}(\mathbb{R})$,*

$$\mathbb{P}_T\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}_T(A_n).$$

2.2 Espérance : valeur moyenne et centre de gravité

L'une des grandeurs fondamentales associées à une variable aléatoire est son *espérance*. Elle représente la valeur moyenne autour de laquelle la variable aléatoire tend à se concentrer.

Définition 2.2.1 (Espérance d'une variable aléatoire continue)**.** *Soit T une variable aléatoire réelle continue de densité f. Si l'intégrale*

$$\mathbb{E}[T] = \int_{\mathbb{R}} t\, f(t)\, \mathrm{d}t$$

existe et est finie, on appelle espérance *de T cette quantité.*

L'espérance peut être interprétée comme le *centre de gravité* de la loi de probabilité. Si l'on imagine la densité $f(t)$ comme une répartition de masse sur l'axe réel, alors $\mathbb{E}[T]$ est le point d'équilibre autour duquel la masse se répartit.

Dans un cadre de survie, l'espérance correspond à la durée de vie moyenne dans la population. Il s'agit d'une caractéristique globale : deux lois différentes peuvent partager la même espérance tout en ayant des comportements très différents.

Propriétés fondamentales de l'espérance. L'espérance possède des propriétés algébriques essentielles.

Proposition 2.2.1 (Linéarité de l'espérance)**.** *Soient X et Y deux variables aléatoires admettant une espérance, et $\alpha, \beta \in \mathbb{R}$. Alors*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\, \mathbb{E}[X] + \beta\, \mathbb{E}[Y].$$

Cette propriété est valable sans hypothèse d'indépendance. Elle traduit le fait que l'espérance se comporte comme une application linéaire, ce qui en fait un outil particulièrement maniable dans les calculs probabilistes.

Théorème du transfert. Le calcul des espérances repose souvent sur un principe fondamental appelé *théorème du transfert*. Ce résultat permet d'exprimer l'espérance d'une fonction d'une variable aléatoire directement à partir de la loi de cette variable, sans avoir à déterminer explicitement la loi de l'image.

Théorème 2.2.1 (Théorème du transfert)**.** *Soit X une variable aléatoire réelle de densité f_X, et soit $\varphi : \mathbb{R} \to \mathbb{R}$ une fonction mesurable telle que l'intégrale ci-dessous soit définie. Alors*

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)\, f_X(x)\, \mathrm{d}x.$$

Preuve. Par définition, l'espérance de $\varphi(X)$ s'écrit

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} y\, f_{\varphi(X)}(y)\, \mathrm{d}y,$$

où $f_{\varphi(X)}$ désigne la densité de la variable aléatoire $\varphi(X)$. Lorsque φ est injective et régulière, on peut déterminer cette densité par changement de variable. Toutefois, cette approche devient rapidement complexe lorsque φ n'est pas bijective. Le théorème du transfert contourne cette difficulté en revenant à la définition probabiliste de l'espérance :

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)\, \mathrm{d}\mathbb{P}_X(x),$$

Il faudrait définir cette intégrale par rapport à une mesure peut-être, sans trop entrer dans les détails.

où $\mathbb{P}_X$ est la loi de X. Lorsque X admet une densité f_X, la mesure $\mathrm{d}\mathbb{P}_X(x)$ s'écrit $f_X(x)\,\mathrm{d}x$, ce qui conduit directement à la formule annoncée. □

Intuition. Le théorème du transfert affirme que pour calculer une moyenne de la quantité $\varphi(X)$, il suffit de moyenner les valeurs $\varphi(x)$ pondérées par la probabilité que X prenne la valeur x. Autrement dit, on "transfère" la transformation φ à l'intérieur de l'intégrale, sans transformer explicitement la loi de X.

Intérêt. Ce théorème est d'un intérêt pratique majeur. Il permet notamment :

— de calculer des moments comme $\mathbb{E}[X^2]$, $\mathbb{E}[|X|]$ ou $\mathbb{E}[\exp(X)]$;

— d'établir des formules générales pour la variance et la covariance ;

— de simplifier considérablement les calculs d'espérance en statistique et en apprentissage automatique, où l'on manipule fréquemment des fonctions non linéaires de variables aléatoires.

Le théorème du transfert constitue ainsi un pont fondamental entre la loi d'une variable aléatoire et les quantités moyennes dérivées de cette loi.

2.3 Variance : mesure de la dispersion

Si l'espérance renseigne sur la position centrale de la distribution, elle ne dit rien sur la manière dont les valeurs sont dispersées autour de cette position. C'est le rôle de la variance.

Définition 2.3.1 (Variance)**.** *Soit X une variable aléatoire d'espérance $\mu = \mathbb{E}[X]$. On appelle* variance *de X la quantité*

$$\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big],$$

lorsqu'elle est finie. On dispose également de l'identité utile :

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

lorsque ces quantités sont bien définies.

La variance mesure la *dispersion quadratique* de la variable aléatoire autour de son espérance. Plus la variance est grande, plus les valeurs prises par X sont éloignées, en moyenne, du centre de gravité μ.

Le caractère quadratique de cette mesure est essentiel : les écarts positifs et négatifs ne s'annulent pas, et les grandes déviations sont davantage pénalisées.

Propriétés de la variance. Deux remarques importantes concernant la variance.

Proposition 2.3.1. *Pour toute variable aléatoire X et tous α et $\lambda \in \mathbb{R}$,*

$$\mathrm{Var}(\lambda X + \alpha) = \lambda^2 \,\mathrm{Var}(X).$$

Cette propriété montre que la variance est homogène de degré 2, ce qui explique pourquoi elle s'exprime dans le carré de l'unité de X.

2.3.1 Covariance : dépendance linéaire entre variables

Lorsque l'on considère plusieurs variables aléatoires, il devient nécessaire de mesurer la manière dont elles varient conjointement.

Définition 2.3.2 (Covariance)**.** *Soient X et Y deux variables aléatoires admettant une espérance. On appelle* covariance *de X et Y la quantité*

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big].$$

La covariance mesure la dépendance *linéaire* entre deux variables aléatoires :

— si $\mathrm{Cov}(X,Y) > 0$, les grandes valeurs de X tendent à être associées à de grandes valeurs de Y ;

— si $\mathrm{Cov}(X,Y) < 0$, les grandes valeurs de X tendent à être associées à de petites valeurs de Y ;
— si $\mathrm{Cov}(X,Y) = 0$, il n'existe pas de dépendance linéaire entre X et Y.

Propriétés algébriques. Ci-dessous les propriétés les plus importantes de l'opérateur de covariance.

Proposition 2.3.2. *La covariance est une forme bilinéaire symétrique :*
— $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$;
— $\mathrm{Cov}(\alpha X + \beta Y, Z) = \alpha\,\mathrm{Cov}(X,Z) + \beta\,\mathrm{Cov}(Y,Z)$;
— $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.

Ces propriétés rapprochent la covariance d'un produit scalaire. La différence essentielle réside dans le fait que $\mathrm{Cov}(X,X) = 0$ n'implique pas nécessairement $X = 0$ presque sûrement définir "p.s." et on peut ajouter que cela implique que X est constant p.s..

Variance d'une somme. La covariance permet d'exprimer de manière générale la variance d'une somme de variables aléatoires.

Proposition 2.3.3 (Variance d'une somme)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires admettant une variance. Alors*

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j).$$

Preuve. On écrit

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right),$$

puis on développe en utilisant la bilinéarité de la covariance. □

Proposition 2.3.4 (Corollaire induit de la variance de variables indépendantes)**.** *Si les variables $X_1, \dots, X_n$ sont indépendantes, alors*

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Ce résultat fondamental explique pourquoi l'indépendance joue un rôle central en probabilités : elle permet une propagation simple et additive de la variabilité.

2.3.2 Matrice de variance–covariance et interprétation géométrique

Les notions de variance et de covariance se généralisent naturellement lorsque l'on considère non plus une variable aléatoire réelle, mais un vecteur aléatoire

$$X = (X^1, \dots, X^p)^\top \in \mathbb{R}^p.$$

Dans ce cadre, la moyenne $\mu = \mathbb{E}[X] \in \mathbb{R}^p$ joue le même rôle que dans le cas unidimensionnel : elle représente un opérateur de position, c'est-à-dire le *centre* du nuage de points associé aux réalisations de X dans l'espace $\mathbb{R}^p$. La dispersion et les dépendances linéaires entre coordonnées sont quant à elles décrites par une matrice, appelée matrice de variance–covariance.

Définition 2.3.3 (Matrice de variance–covariance)**.** *Soit* $X = (X^1, \dots, X^p)^\top$ *un vecteur aléatoire de moyenne* $\mu = \mathbb{E}[X]$*. On appelle* matrice de variance–covariance *de* X *la matrice* $\Sigma \in \mathbb{R}^{p\times p}$ *définie par*

$$\Sigma = \mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)(X-\mu)^\top\big]\,.$$

Ses coefficients sont donnés par

$$\Sigma_{jk} = \mathrm{Cov}(X^j, X^k), \qquad 1 \le j,k \le p.$$

En particulier, les termes diagonaux vérifient $\Sigma_{jj} = \mathrm{Var}(X^j)$*.*

Par construction, la matrice Σ est symétrique et positive semi-définie, ce qui reflète le fait qu'elle encode une dispersion quadratique.

Exemple 2.3.1 (Illustration en dimension 2.)**.** *Considérons* $X = (X^1, X^2)^\top$ *et notons*

$$\mu = \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} = \begin{pmatrix}\mathbb{E}[X^1]\\ \mathbb{E}[X^2]\end{pmatrix}.$$

Alors,

$$\Sigma = \mathbb{E}\left[\begin{pmatrix}X^1-\mu_1\\ X^2-\mu_2\end{pmatrix}\begin{pmatrix}X^1-\mu_1 & X^2-\mu_2\end{pmatrix}\right]$$

$$= \begin{pmatrix}\mathbb{E}\big[(X^1-\mu_1)^2\big] & \mathbb{E}\big[(X^1-\mu_1)(X^2-\mu_2)\big]\\ \mathbb{E}\big[(X^2-\mu_2)(X^1-\mu_1)\big] & \mathbb{E}\big[(X^2-\mu_2)^2\big]\end{pmatrix}.$$

On retrouve donc explicitement

$$\Sigma = \begin{pmatrix}\mathrm{Var}(X^1) & \mathrm{Cov}(X^1,X^2)\\ \mathrm{Cov}(X^1,X^2) & \mathrm{Var}(X^2)\end{pmatrix}.$$

Comme étayéen section précédente, le terme $\mathrm{Cov}(X^1, X^2)$ mesure la dépendance linéaire entre les deux coordonnées : son signe indique si les variations de X^1 et X^2 tendent à aller dans le même sens (covariance positive) ou en sens opposé (covariance négative), et sa valeur absolue quantifie l'intensité de cette relation en unités quadratiques.

Interprétation géométrique : ellipses de dispersion. Prenons le cas de la dimension 2, comme illustré en figure 2.1. La matrice de variance–covariance admet une interprétation géométrique particulièrement instructive. Intuitivement, la moyenne μ indique la position centrale du nuage de points, tandis que Σ – dont le rôle est d'illustrer la dispersion autour de la moyenne μ – décrit sa forme et son orientation. Considérons l'ensemble défini par

$$\mathcal{E}_c = \big\{x \in \mathbb{R}^2 \;:\; (x-\mu)^\top \Sigma^{-1}(x-\mu) \le c^2\big\}\,,$$

où $c > 0$ est une constante et où l'on suppose Σ inversible. Cet ensemble est une ellipse centrée en μ. Les axes principaux de l'ellipse sont orientés selon les vecteurs propres de Σ, et leurs longueurs sont proportionnelles aux racines carrées des valeurs propres de Σ. Autrement dit :

— la moyenne μ joue le rôle de *centre* (opérateur de position) ;

— les variances déterminent la dispersion selon les directions principales ;

— la covariance oriente l'ellipse : si $\mathrm{Cov}(X^1, X^2) \neq 0$, l'ellipse est inclinée par rapport aux axes.

Cette représentation est particulièrement utile en statistique multivariée et en apprentissage automatique : elle permet de visualiser en un coup d'œil la dispersion d'un groupe d'observations, la présence de corrélations entre variables, et la structure géométrique induite par la matrice de covariance. Elle constitue également le point de départ naturel de méthodes de réduction de dimension (telles que l'ACP), qui cherchent précisément à identifier les directions de plus grande variance.

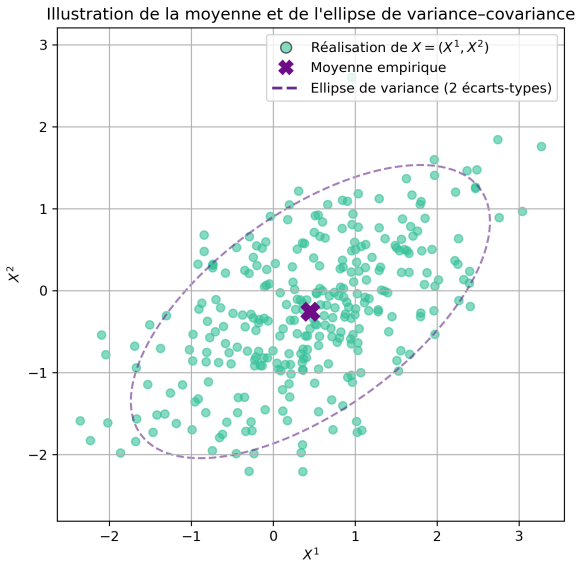


FIGURE 2.1 – Illustration géométrique de la moyenne et de la matrice de variance–covariance pour un vecteur aléatoire bidimensionnel $X = (X^1, X^2)$. Les points représentent des réalisations de X, la croix indique la moyenne empirique, qui joue le rôle d'opérateur de position du nuage de points dans le plan. L'ellipse en pointillé correspond à une courbe d'iso-dispersion associée à la matrice de variance–covariance : sa taille traduit l'intensité de la dispersion, tandis que son orientation reflète la présence d'une covariance non nulle entre X^1 et X^2.

2.4 Corrélation

La covariance introduite précédemment permet de mesurer la dépendance linéaire entre deux variables aléatoires. Toutefois, sa valeur dépend des unités de mesure de X et Y, ce qui rend les comparaisons délicates. La *corrélation* vise précisément à pallier cette difficulté en proposant une version *normalisée* de la covariance.

Définition 2.4.1 (Corrélation linéaire)**.** *Soient X et Y deux variables aléatoires admettant une espérance et une variance strictement positive. On appelle* corrélation *(ou coefficient de corrélation linéaire de Pearson) entre X et Y la quantité*

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}}.$$

Par construction, la corrélation est une quantité *sans dimension* : elle ne dépend pas des unités dans lesquelles X et Y sont mesurées.

Intuition et interprétation. La corrélation peut être interprétée comme une mesure du *lien linéaire* entre X et Y, évalué de manière symétrique. Le numérateur, la covariance, mesure la manière dont X et Y varient conjointement. Le dénominateur, produit des écarts types de X et Y, joue un rôle de normalisation : il ramène cette mesure à une échelle comparable, comprise entre -1 et 1.

Ainsi :

— $\rho_{X,Y} > 0$ indique que les grandes valeurs de X sont en moyenne associées à de grandes valeurs de Y ;

— $\rho_{X,Y} < 0$ indique une association inverse ;

— $\rho_{X,Y} \approx 0$ suggère l'absence de dépendance linéaire significative.

D'un point de vue géométrique, la corrélation peut être vue comme le *cosinus de l'angle* entre les variables centrées $X - \mathbb{E}[X]$ et $Y - \mathbb{E}[Y]$ dans un espace de Hilbert probabiliste. Cette lecture explique

naturellement pourquoi la corrélation est bornée par -1 et 1, et pourquoi elle quantifie un alignement linéaire entre les deux variables.

Propriétés fondamentales. Quelques propriétés remarquables associées :

Proposition 2.4.1. *Soient X et Y deux variables aléatoires de variance non nulle. Alors :*

- $-1 \leq \rho_{X,Y} \leq 1$;
- $\rho_{X,Y} = \rho_{Y,X}$ *(symétrie) ;*
- *pour tout* $\alpha, \beta \in \mathbb{R} \setminus \{0\}$,
$$\rho_{\alpha X, \beta Y} = \operatorname{sign}(\alpha\beta)\, \rho_{X,Y};$$
- $\rho_{X,Y} = \pm 1$ *si et seulement si il existe* $a \in \mathbb{R}$ *et* $b \in \mathbb{R}$ *tels que*
$$Y = aX + b \quad \textit{presque sûrement.}$$

La dernière propriété montre que la corrélation égale à ± 1 correspond à une dépendance linéaire parfaite, sans bruit.

Lien avec la régression linéaire. La corrélation joue un rôle central dans l'étude de la régression linéaire simple. En effet, lorsque l'on cherche à expliquer Y comme une fonction linéaire de X à un terme d'erreur près,
$$Y = \beta_0 + \beta_1 X + \varepsilon,$$
le coefficient de corrélation mesure l'intensité de la relation linéaire entre X et Y, indépendamment de l'échelle des variables. On montrera ultérieurement que le coefficient de régression β_1 est directement lié à la covariance entre X et Y, et donc à leur corrélation, ce qui confère à cette dernière une interprétation statistique et géométrique fondamentale.

Limites. Il est important de souligner qu'une corrélation nulle n'implique pas nécessairement l'indépendance entre X et Y, sauf dans des cas particuliers (notamment lorsque les variables sont gaussiennes). La corrélation ne capture que les dépendances *linéaires* et peut être aveugle à des relations non linéaires pourtant marquées.

2.5 La mesure de Dirac

Avant d'introduire les lois de probabilité générales, il est utile de s'arrêter sur un objet élémentaire : la *mesure de Dirac*. Celle-ci permet de formaliser mathématiquement l'idée de masse concentrée en un point, notion omniprésente aussi bien en probabilités qu'en statistique.

Motivation et intuition. Considérons une situation extrême dans laquelle une variable aléatoire prend une valeur déterministe $a \in \mathbb{R}$ avec certitude. Intuitivement, toute la masse de probabilité est alors concentrée en ce point unique. La mesure de Dirac formalise précisément cette situation. D'un point de vue statistique, la mesure de Dirac permet également de représenter une observation individuelle : une donnée observée X_i peut être vue comme une masse de probabilité concentrée en X_i.

Définition 2.5.1 (Mesure de Dirac)**.** *Soit $a \in \mathbb{R}$. La* mesure de Dirac en a, *notée δ_a, est la mesure définie sur la tribu borélienne $\mathcal{B}(\mathbb{R})$, pour tout $A \in \mathcal{B}(\mathbb{R})$, par*
$$\delta_a(A) = \begin{cases} 1 & \textit{si } a \in A, \\ 0 & \textit{sinon.} \end{cases}$$

Ainsi, δ_a attribue toute la masse à l'unique point a et aucune masse ailleurs.

Interprétation intégrale. L'intérêt principal de la mesure de Dirac apparaît lorsqu'on l'intègre contre une fonction test.

Proposition 2.5.1. *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction mesurable. Alors*

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}\delta_a(x) = f(a). \tag{2.1}$$

Cette propriété justifie l'écriture heuristique souvent rencontrée

$$\int f(x)\,\delta_a(x)\,\mathrm{d}x = f(a),$$

qu'il convient d'interpréter rigoureusement comme une intégration par rapport à une mesure, et non comme une fonction ordinaire.

Exemple 2.5.1. *Soit X une variable aléatoire telle que $X = a$ presque sûrement. La loi de X est alors la mesure de Dirac δ_a, et son espérance vaut*

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\,\mathrm{d}\delta_a(x) = a.$$

Cet exemple illustre que la mesure de Dirac correspond au cas limite d'une variable aléatoire sans incertitude.

Intérêts en probabilités et en statistique. La mesure de Dirac constitue un outil fondamental pour comprendre la transition entre données observées et modélisation probabiliste et joue donc un rôle central à plusieurs niveaux :

- **En probabilités**, elle permet de représenter des lois dégénérées et sert de brique élémentaire pour construire des lois discrètes comme combinaisons linéaires de mesures de Dirac.
- **En statistique**, elle est au cœur de la définition de la *mesure empirique*, qui approxime une loi inconnue à partir d'un échantillon :
$$\widehat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}. \tag{2.2}$$
- Elle fournit ainsi un lien naturel entre observations individuelles, lois de probabilité et estimateurs *plug-in*.

2.6 Fonction caractéristique : définition, exemples et intérêt

La *fonction caractéristique* est l'outil fondamental de la théorie de la convergence en loi, notamment parce qu'elle existe toujours (sans hypothèse d'intégrabilité).

2.6.1 Définition

Définition 2.6.1 (Fonction caractéristique)**.** *Soit X une variable aléatoire réelle. Sa* fonction caractéristique *est la fonction*

$$\varphi_X(t) := \mathbb{E}\big[e^{itX}\big], \qquad t \in \mathbb{R}, \tag{2.3}$$

où $i^2 = -1$ désigne l'imaginaire pur.

Remarque 1. Comme $|e^{itX}| = 1$, on a toujours $|\varphi_X(t)| \leq 1$ et (2.3) est toujours bien défini.

Exemple 2.6.1 (Exemples usuels). *Si $X = a$ p.s., alors*

$$\varphi_X(t) = e^{ita}.$$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors

$$\varphi_X(t) = \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right).$$

Si $X \sim \mathcal{P}(\lambda)$, alors

$$\varphi_X(t) = \exp\left(\lambda(e^{it} - 1)\right).$$

Si $X \sim \mathrm{Exp}(\lambda)$, alors

$$\varphi_X(t) = \frac{\lambda}{\lambda - it}.$$

2.6.2 Intérêts majeurs

1) Unicité (caractérisation de la loi). La fonction caractéristique *détermine* la loi : si X et Y sont deux variables aléatoires telles que

$$\varphi_X(t) = \varphi_Y(t) \quad \forall t \in \mathbb{R}, \tag{2.4}$$

alors X et Y ont la même loi.

2) Sommes de variables indépendantes. Si X et Y sont indépendantes, alors

$$\varphi_{X+Y}(t) = \varphi_X(t)\,\varphi_Y(t). \tag{2.5}$$

Cette propriété rend les fonctions caractéristiques particulièrement efficaces pour étudier des sommes et établir des théorèmes limites.

3) Convergence en loi (théorème de continuité de Lévy). La convergence des fonctions caractéristiques est *équivalente* à la convergence en distribution (sous une condition de continuité en 0 automatiquement satisfaite). C'est un pilier des preuves du théorème central limite.

Chapitre 3

Statistiques descriptives et illustration des lois connues

Ce chapitre constitue une transition naturelle entre les fondements probabilistes et inférentiels introduits au chapitre précédent et les méthodes de modélisation statistique étudiées par la suite. Avant toute démarche de modélisation, il est essentiel de comprendre, résumer et interpréter les données disponibles. Cette étape repose sur deux piliers complémentaires : l'analyse descriptive, qui vise à synthétiser l'information observée, et l'inférence statistique, qui permet de généraliser ces observations à une population sous-jacente en tenant compte de l'aléa.

Cadre statistique. On considère un échantillon de données réelles $(x_1, \dots, x_n)$ issu de l'observation d'une variable aléatoire réelle X. Les statistiques descriptives ont pour objectif de fournir une représentation synthétique de cet échantillon, à la fois sur le plan numérique et graphique.

3.1 Indicateurs numériques

Grandeurs de position. Les indicateurs de position permettent de localiser le "centre" d'un jeu de données.

Définition 3.1.1 (La moyenne dite "empirique")**.** *La* moyenne empirique *est définie par*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{3.1}$$

Elle peut être interprétée comme le centre de gravité du nuage de points associé aux observations, chaque valeur contribuant proportionnellement à son poids.

On verra plus tard que la moyenne empirique (3.1) est l'estimateur *plug-in* **??** de l'espérance 2.2.1 avec la mesure empirique (**??**).

Définition 3.1.2 (La médiane)**.** *La* médiane *est définie comme une valeur* $m_{0.5}$ *telle que*

$$\mathbb{P}_n(X \leq m_{0.5}) \geq \tfrac{1}{2} \quad \textit{et} \quad \mathbb{P}_n(X \geq m_{0.5}) \geq \tfrac{1}{2},$$

où $\mathbb{P}_n$ *désigne la mesure empirique associée à l'échantillon.*

La médiane correspond donc au quantile d'ordre 0.5 de la fonction de répartition empirique et partage les données en deux parties de taille égale.

De manière plus générale, pour $\alpha \in (0, 1)$, le *quantile empirique d'ordre* α, souvent noté q_α, est construit de sorte que $\mathbb{P}_n(X \leq q_\alpha) \geq \alpha$.

Définition 3.1.3 (Quantile empirique d'ordre α)**.** *Soit $\alpha \in (0,1)$. Le* quantile empirique d'ordre α *est défini comme le* pseudo-inverse généralisé à gauche *de la fonction de répartition empirique :*

$$q_\alpha = F_n^{-1}(\alpha) := \inf\big\{x \in \mathbb{R} \,:\, F_n(x) \geq \alpha\big\}.$$

On présentera pleinement la fonction de répartition empirique en définition **??**. Cette définition garantit l'existence du quantile même lorsque la fonction F_n présente des sauts, ce qui est toujours le cas pour une fonction de répartition empirique associée à un échantillon fini.

Enfin, une autre grandeur de position dont il faut au moins connaître le nom : le mode.

Définition 3.1.4 (Le mode)**.** *Le* mode *est la valeur la plus fréquemment observée dans l'échantillon. Il est particulièrement pertinent pour des données discrètes ou catégorielles, mais peut être moins informatif pour des données continues.*

Grandeurs de dispersion. Les indicateurs de dispersion quantifient la variabilité des données autour d'un indicateur de position.

Définition 3.1.5 (La variance et l'écart-type)**.** *La* variance *qui décrit une population* [a] *est définie par*

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

*et mesure la dispersion quadratique des observations autour de la moyenne. Sa racine carrée, l'*écart-type empirique $\sigma = \sqrt{V_n}$*, est exprimée dans la même unité que la variable observée.*

a. On rappelle que cette grandeur est à nuancer avec la variance empirique estimée, comme détaillé en section **??**.

D'autres indicateurs de dispersion sont parfois utilisés : l'*étendue*, définie comme $\max_i x_i - \min_i x_i$; l'*écart moyen*, qui correspond à la moyenne des distances absolues à la moyenne ; ou encore le *coefficient de variation*, donné par $\sigma/\bar{x}$, utile pour comparer la dispersion relative de variables de natures différentes.

3.2 Représentations graphiques

Les représentations graphiques complètent l'analyse numérique en offrant une visualisation directe et souvent irremplaçable de la structure des données. Elles permettent de détecter des phénomènes que les indicateurs numériques seuls ne suffisent pas toujours à révéler, tels que des relations non linéaires, des asymétries marquées ou la présence de valeurs aberrantes.

Le *nuage de points* (*scatter plot*) est un outil fondamental dès lors que l'on étudie conjointement deux variables quantitatives. Il permet d'observer visuellement l'existence d'une relation entre une variable explicative X et une variable réponse Y : relation linéaire ou non, croissante ou décroissante, présence de groupes ou d'observations atypiques. Il est également central dans l'étude de séries temporelles, où l'on représente une variable X_t en fonction du temps t, afin de mettre en évidence des tendances, des ruptures ou des comportements cycliques. En pratique, le nuage de points constitue presque toujours le point de départ de toute analyse exploratoire bivariée.

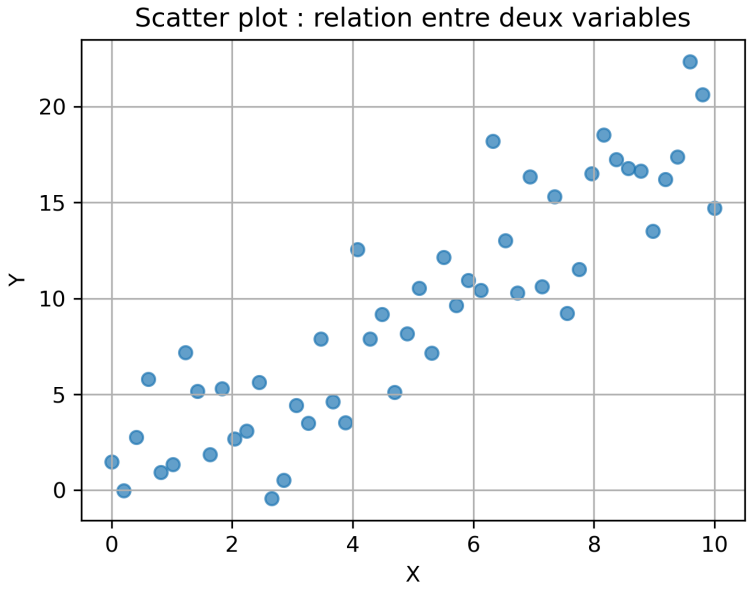


FIGURE 3.1 – Nuage de points illustrant la relation entre deux variables quantitatives.

L'*histogramme* est tout aussi essentiel pour l'analyse univariée d'une variable continue. En regroupant les observations par classes, il fournit une approximation visuelle de la distribution empirique de la variable. À lui seul, un histogramme permet souvent de capter la quasi-totalité de l'information descriptive pertinente : on peut identifier le mode, situer approximativement la moyenne et la médiane, apprécier la dispersion, détecter une asymétrie ou une multimodalité, et se faire une première idée de la variance. C'est un outil central pour formuler des hypothèses sur la loi sous-jacente suivie par la variable observée.

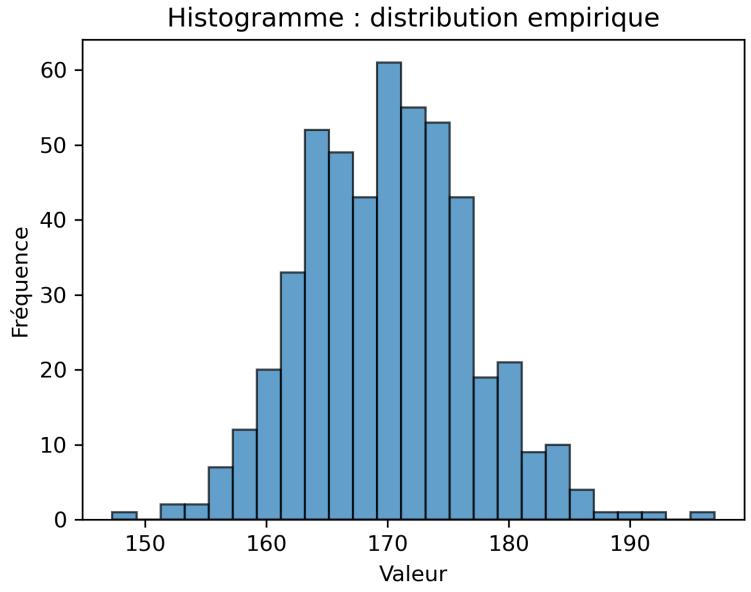


FIGURE 3.2 – Histogramme représentant la distribution empirique d'une variable continue.

Le *diagramme en boîte* (*boxplot*) propose une synthèse compacte de la distribution à partir des quantiles. Il met en évidence la médiane, l'étendue interquartile, la dispersion globale ainsi que les valeurs aberrantes éventuelles. Le boxplot est particulièrement utile pour comparer plusieurs distributions ou pour détecter rapidement des observations atypiques qui pourraient influencer de manière

excessive les indicateurs numériques classiques.

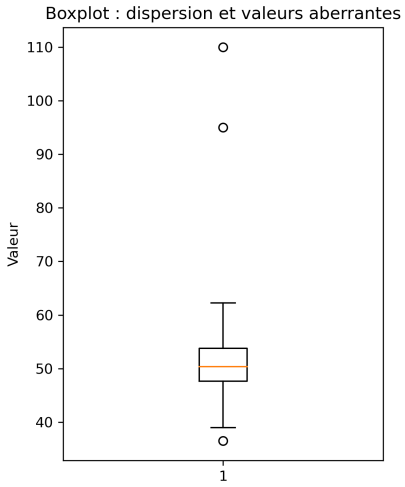


FIGURE 3.3 – Diagramme en boîte mettant en évidence la dispersion et les valeurs aberrantes.

Enfin, le *diagramme circulaire* (*camembert*) est utilisé pour représenter graphiquement des proportions associées à des modalités catégorielles. Il permet de visualiser rapidement la répartition relative des catégories au sein d'un ensemble de données. Toutefois, son usage doit rester limité à des situations où le nombre de modalités est faible, sous peine de nuire à la lisibilité et à l'interprétation.

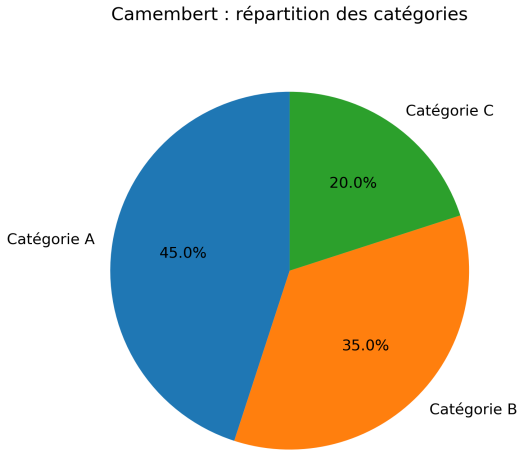


FIGURE 3.4 – Diagramme circulaire illustrant la répartition de modalités catégorielles.

3.3 Lois discrètes fondamentales

Avant d'aborder les lois continues, il est essentiel de rappeler plusieurs lois discrètes fondamentales, qui jouent un rôle central tant en modélisation probabiliste qu'en statistique inférentielle. Ces lois interviennent naturellement lorsqu'on modélise des phénomènes comptables, des succès/échecs ou des occurrences d'événements.

Loi de Bernoulli. Une variable aléatoire X suit une loi de Bernoulli de paramètre $p \in [0,1]$, notée $\mathcal{B}(p)$, si

$$\mathbb{P}(X=1) = p, \qquad \mathbb{P}(X=0) = 1-p.$$

Cette loi modélise une expérience élémentaire à deux issues, souvent interprétées comme *succès* et *échec*. Elle constitue la brique de base de nombreuses constructions probabilistes. On a

$$\mathbb{E}[X] = p, \qquad \mathrm{Var}(X) = p(1-p).$$

La loi de Bernoulli est fondamentale en inférence, car l'estimation d'une proportion repose directement sur ce modèle.

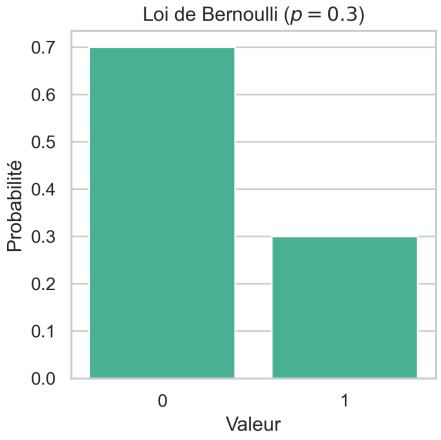


FIGURE 3.5 – Illustration de la loi de Bernoulli de paramètre p. La variable aléatoire ne prend que deux valeurs possibles, 0 et 1, associées respectivement à un échec et à un succès, avec des probabilités $1-p$ et p.

Loi binomiale. Une variable aléatoire X suit une loi binomiale de paramètres (n, p), notée $\mathcal{B}(n, p)$, si elle peut s'écrire comme la somme de n variables de Bernoulli indépendantes et identiquement distribuées :

$$X = \sum_{i=1}^{n} X_i, \qquad X_i \sim \mathcal{B}(p).$$

La loi binomiale modélise le nombre de succès obtenus lors de n répétitions indépendantes d'une même expérience de Bernoulli. Sa fonction de masse est donnée par

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, \dots, n.$$

Elle intervient naturellement dans les problèmes de comptage, d'échantillonnage et d'estimation de proportions. Lorsque n est grand, elle est bien approximée par une loi normale, ce qui justifie de nombreuses approximations asymptotiques.

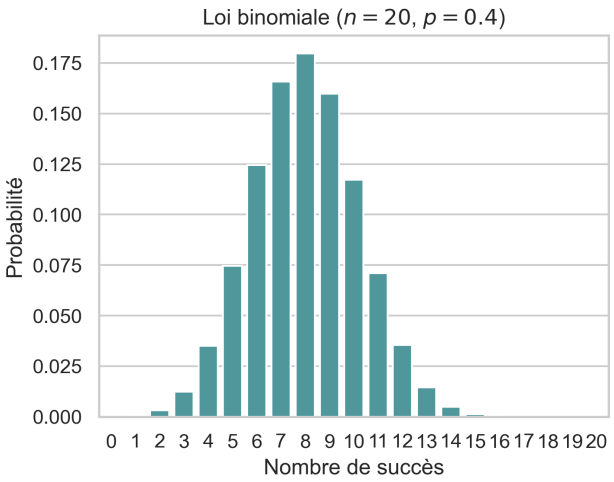


FIGURE 3.6 – Illustration de la loi binomiale de paramètres (n, p). La variable aléatoire représente le nombre de succès obtenus lors de n répétitions indépendantes d'une expérience de Bernoulli de probabilité de succès p.

Loi uniforme discrète. Une variable aléatoire X suit une loi uniforme discrète sur un ensemble fini $\{a_1, \ldots, a_m\}$ si

$$\mathbb{P}(X = a_k) = \frac{1}{m}, \qquad k = 1, \ldots, m.$$

Cette loi modélise un hasard équilibré, où toutes les issues possibles sont équiprobables. Elle est souvent utilisée comme modèle de référence ou comme hypothèse de neutralité, notamment dans les tests statistiques ou les simulations aléatoires.

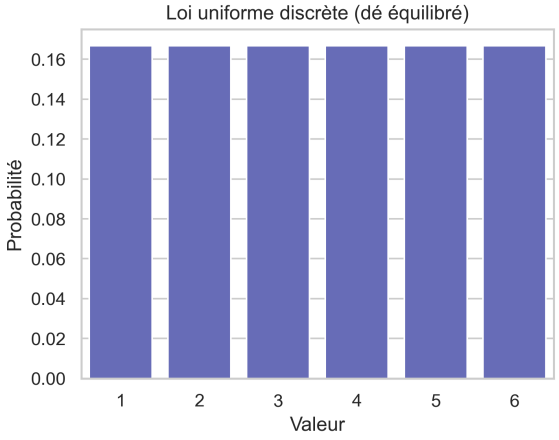


FIGURE 3.7 – Illustration d'une loi uniforme discrète sur un ensemble fini. Toutes les valeurs possibles sont équiprobables, ce qui modélise un hasard parfaitement équilibré.

Loi de Poisson. Une variable aléatoire X suit une loi de Poisson de paramètre $\lambda > 0$, notée $\mathcal{P}(\lambda)$, si

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k \in \mathbb{N}.$$

La loi de Poisson modélise le nombre d'occurrences d'un événement rare sur un intervalle donné, lorsque ces occurrences sont indépendantes et apparaissent à un taux moyen constant. Elle est largement utilisée pour modéliser des comptages (arrivées de clients, défauts, événements biologiques ou physiques).

Une propriété remarquable est que

$$\mathbb{E}[X] = \mathrm{Var}(X) = \lambda.$$

Elle peut être obtenue comme limite d'une loi binomiale lorsque $n \to \infty$ et $p \to 0$ avec $np \to \lambda$.

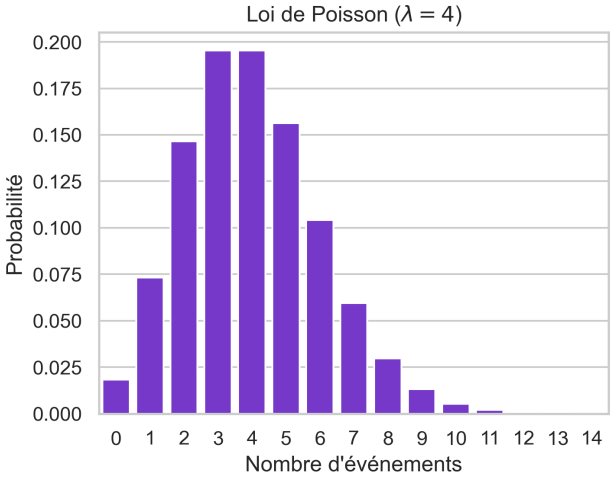


FIGURE 3.8 – Illustration de la loi de Poisson de paramètre λ. Cette loi discrète modélise le nombre d'occurrences d'un événement rare observé sur un intervalle donné, lorsque ces occurrences apparaissent de manière indépendante à un taux moyen constant.

Loi géométrique. Une variable aléatoire X suit une loi géométrique de paramètre $p \in (0, 1)$ si

$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \qquad k \in \mathbb{N}^*.$$

Elle modélise le nombre d'essais nécessaires avant l'obtention du premier succès dans une suite d'expériences de Bernoulli indépendantes. La loi géométrique possède une propriété de *sans mémoire*, analogue discret de celle de la loi exponentielle :

$$\mathbb{P}(X > k + \ell \mid X > k) = \mathbb{P}(X > \ell).$$

Cette loi est utilisée pour modéliser des temps d'attente discrets ou des durées exprimées en nombre d'essais.

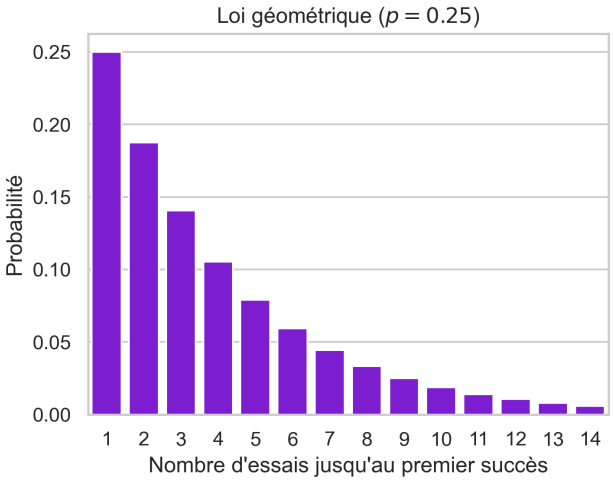


FIGURE 3.9 – Illustration de la loi géométrique sur $\mathbb{N}^*$ de paramètre p. Cette loi modélise le nombre d'essais nécessaires avant l'obtention du premier succès dans une suite d'expériences de Bernoulli indépendantes.

Ces lois discrètes constituent un socle essentiel pour l'analyse probabiliste et statistique. Elles apparaissent fréquemment comme modèles directs de phénomènes observés, ou comme lois limites ou intermédiaires dans des raisonnements asymptotiques. Elles préparent naturellement l'introduction des lois continues fondamentales et des méthodes inférentielles associées.

3.4 Lois continues fondamentales

Certaines lois continues jouent un rôle central en probabilités et en statistique, soit par leur fréquence d'apparition dans les phénomènes naturels, soit par leurs propriétés mathématiques remarquables, soit encore par leur rôle fondamental dans les méthodes d'inférence. Cette section présente les principales lois continues utilisées en modélisation statistique.

3.4.1 La loi normale (ou gaussienne)

La *loi normale* occupe une place absolument centrale en statistique, tant par sa fréquence d'apparition dans les phénomènes naturels que par ses propriétés mathématiques remarquables.

Définition 3.4.1 (Loi normale)**.** *Une variable aléatoire réelle X suit une loi normale de moyenne $m \in \mathbb{R}$ et de variance $\sigma^2 > 0$, notée*

$$X \sim \mathcal{N}(m, \sigma^2),$$

si elle admet pour densité la fonction

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{3.2}$$

Interprétation des paramètres. Le paramètre m correspond au centre de la distribution (et à son espérance), tandis que σ^2 mesure la dispersion autour de ce centre. La densité est symétrique par rapport à m et décroît rapidement lorsque l'on s'en éloigne, ce qui reflète la concentration de la masse de probabilité autour de la moyenne.

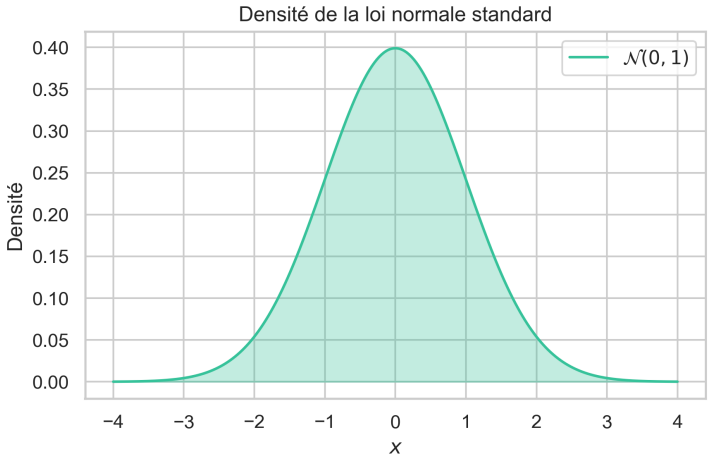


FIGURE 3.10 – Densité de la loi normale standard $\mathcal{N}(0, 1)$. Cette loi symétrique est caractérisée par sa moyenne nulle et sa variance unitaire, et joue un rôle central en statistique et en inférence.

Définition 3.4.2 (Z-score (centrage–réduction))**.** *Soit X une variable aléatoire réelle de moyenne*

m et d'écart-type $\sigma > 0$. On appelle Z-score *(ou variable centrée réduite) la variable aléatoire*

$$Z = \frac{X - m}{\sigma}. \tag{3.3}$$

Par construction, la variable Z est sans dimension, de moyenne nulle et de variance égale à 1. Lorsque X suit une loi normale $\mathcal{N}(m, \sigma^2)$, on a

$$Z \sim \mathcal{N}(0, 1).$$

Intérêt du Z-score. Le Z-score fournit une mesure normalisée de l'écart d'une observation à la moyenne, exprimée en nombre d'écarts-types. Cette normalisation présente plusieurs intérêts fondamentaux :

- **Comparabilité :** elle permet de comparer des observations issues de distributions différentes, dès lors que celles-ci sont approximativement symétriques et centrées autour de leur moyenne.
- **Détection d'observations atypiques :** dans une distribution à peu près symétrique, des valeurs de $|Z|$ élevées indiquent des observations éloignées du comportement central. Par exemple, dans le cas gaussien,
$$\mathbb{P}(|Z| \leq 1.96) \simeq 0.95,$$
ce qui suggère qu'une observation telle que $|Z| > 2$ est peu probable et peut être considérée comme atypique.
- **Intuition des tests statistiques :** de nombreux tests reposent sur la comparaison d'une statistique normalisée à des seuils issus de la loi normale standard. Le Z-score constitue ainsi une première intuition des mécanismes de décision statistique, fondés sur l'éloignement relatif à la moyenne sous une hypothèse de référence.

Lorsque la distribution sous-jacente est seulement *approximativement* symétrique, le Z-score reste un outil heuristique pertinent, à condition d'interpréter les seuils avec prudence.

Stabilité par addition. Une propriété structurelle majeure de la loi normale est sa *stabilité par addition* : si $X \sim \mathcal{N}(m_1, \sigma_1^2)$ et $Y \sim \mathcal{N}(m_2, \sigma_2^2)$ sont indépendantes, alors

$$X + Y \sim \mathcal{N}(m_1 + m_2,\ \sigma_1^2 + \sigma_2^2).$$

Cette propriété explique pourquoi la loi normale apparaît naturellement comme loi limite pour des sommes de variables aléatoires indépendantes.

Lien avec le théorème central limite. Le *théorème central limite* (voir théorème **??**) établit que, sous des hypothèses très générales, la moyenne d'un grand nombre de variables aléatoires i.i.d. est approximativement normale, indépendamment de la loi initiale. Ce résultat confère à la loi normale un rôle universel en statistique.

Propriété d'entropie maximale. Un autre argument fondamental justifiant le rôle central de la loi normale en modélisation statistique repose sur une notion issue de la théorie de l'information : l'entropie différentielle, qu'on définira plus tard en **??**. Cette quantité mesure le degré de dispersion ou d'incertitude associé à une loi de probabilité continue : plus l'entropie est élevée, plus la distribution est étalée et moins elle est informative.

Un théorème remarquable **??** établit que, parmi toutes les lois continues de moyenne m et de variance σ^2 fixées, la loi normale $\mathcal{N}(m, \sigma^2)$ est celle qui *maximise* l'entropie différentielle. Autrement dit, sous ces seules contraintes de premier et second ordre, la loi normale est la distribution la moins informative possible, ce qui renforce son statut de modèle de référence en l'absence d'information supplémentaire.

Interprétation statistique. Ce résultat signifie que la loi normale est la distribution la moins informative possible lorsque seules la moyenne et la variance sont spécifiées. Elle constitue ainsi un choix naturel de modélisation en l'absence d'information supplémentaire, ce qui explique sa présence omniprésente en statistique et en inférence.

3.4.2 La loi exponentielle

La *loi exponentielle* est une loi continue définie sur $\mathbb{R}_+$, largement utilisée pour modéliser des durées de vie ou des temps d'attente.

Définition 3.4.3 (Loi exponentielle)**.** *Une variable aléatoire X suit une loi exponentielle de paramètre $\lambda > 0$, notée*

$$X \sim \text{Exp}(\lambda),$$

si sa densité est donnée par

$$f(x) = \lambda e^{-\lambda x}\, \mathbb{1}_{x \geq 0}. \tag{3.4}$$

Comme on a pu le montrer dans l'exemple **??** pour $X \sim \text{Exp}(\lambda)$, on a $\mathbb{E}[X] = 1/\lambda$ et $\text{Var}(X) = 1/\lambda^2$.

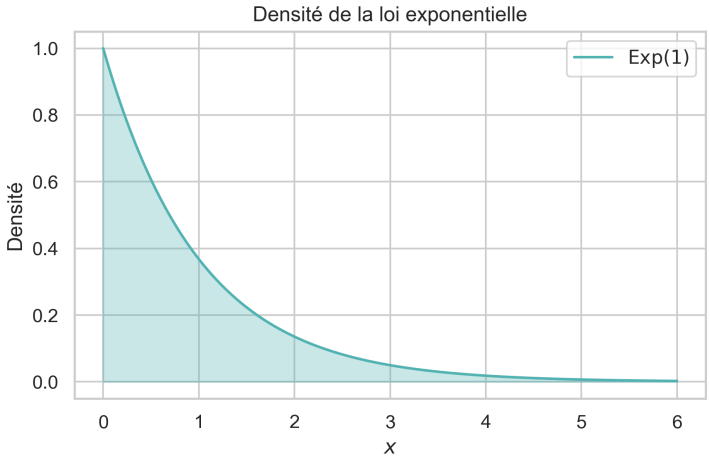


FIGURE 3.11 – Densité de la loi exponentielle de paramètre $\lambda = 1$. Cette loi est définie sur $\mathbb{R}_+$ et est utilisée pour modéliser des temps d'attente ou des durées de vie sans mémoire.

Propriété de non-mémoire. La loi exponentielle est caractérisée par la propriété de *non-mémoire* :

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t), \tag{3.5}$$

pour tous $s, t \geq 0$. Cette propriété signifie que le temps d'attente restant ne dépend pas du temps déjà écoulé, ce qui la rend particulièrement adaptée à la modélisation de pannes ou d'événements aléatoires sans vieillissement.

3.4.3 La loi de Student

La *loi de Student* joue un rôle fondamental en inférence statistique, notamment lorsque la variance d'une population normale est inconnue.

Une variable aléatoire T suit une loi de Student à $\nu > 0$ degrés de liberté, notée

$$T \sim t_\nu,$$

si elle peut s'écrire comme

$$T = \frac{Z}{\sqrt{U/\nu}}, \tag{3.6}$$

où $Z \sim \mathcal{N}(0,1)$ et $U \sim \chi^2(\nu)$ sont indépendantes.

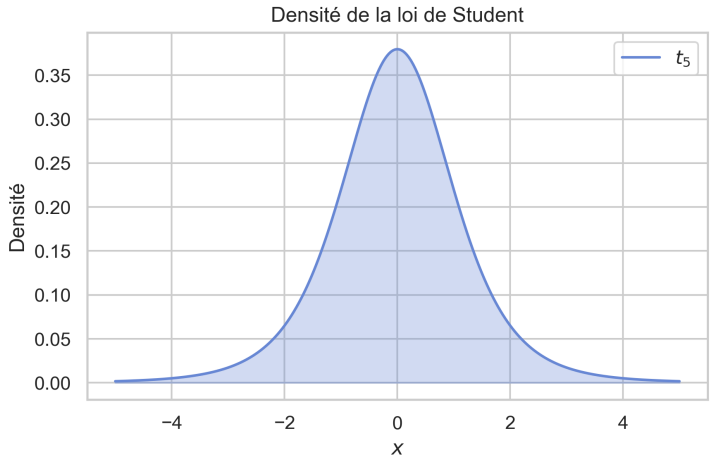


FIGURE 3.12 – Densité de la loi de Student à $\nu = 5$ degrés de liberté. Par rapport à la loi normale, elle présente des queues plus épaisses, traduisant l'incertitude liée à l'estimation de la variance.

Interprétation. La loi de Student ressemble à une loi normale centrée, mais possède des *queues plus épaisses*, reflétant l'incertitude supplémentaire liée à l'estimation de la variance. Lorsque $\nu \to \infty$, la loi de Student converge vers la loi normale standard.

Rôle en inférence. La loi de Student apparaît naturellement dans les tests de moyenne (test de Student) et les intervalles de confiance lorsque la variance est inconnue et estimée à partir des données.

3.4.4 Les lois du χ^2

La loi du χ^2 est étroitement liée à la loi normale et joue un rôle central dans l'estimation de la variance et les tests d'hypothèses.

Une variable aléatoire U suit une loi du χ^2 à ν degrés de liberté, notée

$$U \sim \chi^2(\nu),$$

si

$$U = \sum_{i=1}^{\nu} Z_i^2, \tag{3.7}$$

où $Z_1, \ldots, Z_\nu$ sont des variables i.i.d. de loi $\mathcal{N}(0,1)$.

On a

$$\mathbb{E}[U] = \nu, \qquad \mathrm{Var}(U) = 2\nu.$$

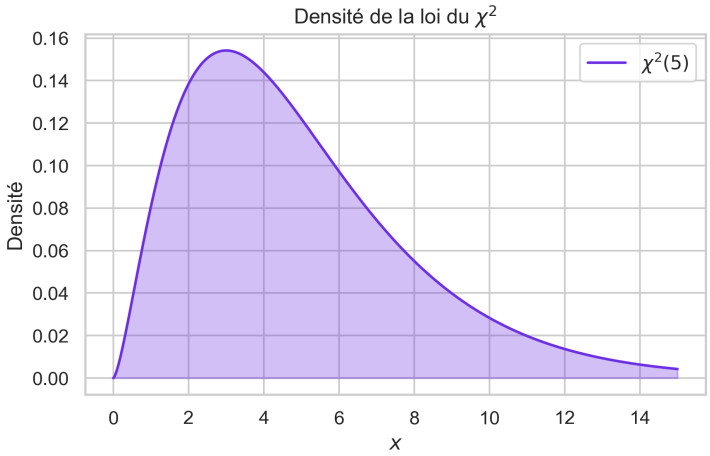


FIGURE 3.13 – Densité de la loi du χ^2 à $\nu = 5$ degrés de liberté. Cette loi asymétrique intervient naturellement dans l'estimation de la variance et la construction de nombreuses statistiques de test.

Rôle statistique. La loi du χ^2 intervient dans :

— l'estimation et les tests sur la variance d'une population normale,

— les tests d'adéquation,

— la construction de la loi de Student et de nombreuses statistiques de test.

3.4.5 Autres lois continues usuelles

D'autres lois continues apparaissent fréquemment en modélisation statistique :

— la *loi uniforme*, qui modélise un hasard parfaitement équilibré sur un intervalle ;

— la *loi log-normale*, utilisée lorsque le logarithme de la variable est normal (revenus, tailles, concentrations) ;

— la *loi Gamma*, généralisation de la loi exponentielle, adaptée à la modélisation de durées ou de quantités positives ;

— les lois issues de transformations de lois normales, très présentes en inférence statistique.

Ces lois constituent une boîte à outils essentielle pour la modélisation probabiliste et l'inférence, chacune étant associée à des hypothèses et des contextes d'application spécifiques.

Chapitre 4

Outils topologiques pour la modélisation

De nombreux algorithmes de machine learning reposent, explicitement ou implicitement, sur une notion de *proximité* entre observations : classification par plus proches voisins, méthodes à noyaux, clustering, réduction de dimension, arbres et forêts (via des critères de séparation), etc. Avant d'introduire ces modèles, il est donc utile de formaliser ce que signifie "être proche" dans un espace de données $\mathcal{X}$. Cette formalisation passe naturellement par la *topologie*, généralement induite par une *distance* (métrique) ou par une *similarité*.

4.1 Topologie sur des données quantitatives

4.1.1 Distance, métrique et topologie induite

Soit $\mathcal{X}$ un ensemble représentant l'espace des observations (par exemple $\mathbb{R}^p$ lorsque l'on manipule p variables quantitatives). Une manière standard de structurer $\mathcal{X}$ consiste à définir une distance entre ses éléments.

Définition 4.1.1 (Distance (métrique))**.** *On appelle* distance *sur $\mathcal{X}$ toute application*

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$$

telle que pour tous $x, y, z \in \mathcal{X}$:

1. ***Séparation :*** $d(x, y) = 0 \iff x = y$;
2. ***Symétrie :*** $d(x, y) = d(y, x)$;
3. ***Inégalité triangulaire :*** $d(x, z) \leq d(x, y) + d(y, z)$.

Le couple $(\mathcal{X}, d)$ est alors appelé espace métrique.

Une distance induit une topologie sur $\mathcal{X}$ via les *boules ouvertes*.

Définition 4.1.2 (Boule ouverte et topologie induite)**.** *Soit $(\mathcal{X}, d)$ un espace métrique. Pour $x \in \mathcal{X}$ et $r > 0$, la boule ouverte de centre x et de rayon r est*

$$B(x, r) = \{y \in \mathcal{X} \ : \ d(x, y) < r\}.$$

On appelle ouverts *de la topologie induite par d les ensembles $U \subset \mathcal{X}$ tels que pour tout $x \in U$, il existe $r > 0$ vérifiant $B(x, r) \subset U$.*

Cette construction formalise l'idée intuitive suivante : un point x est "intérieur" à un ensemble U si l'on peut se déplacer *un peu* autour de x sans sortir de U. C'est exactement ce dont ont besoin de nombreux algorithmes : comparer localement des points selon la proximité définie par d.

4.1.2 Distances de Minkowski sur $\mathbb{R}^p$

Lorsque les données sont quantitatives, l'espace naturel est $\mathbb{R}^p$. Les distances les plus utilisées proviennent des normes ℓ_q.

Définition 4.1.3 (Distance de Minkowski)**.** *Pour $q \geq 1$, la distance de Minkowski d'ordre q sur $\mathbb{R}^p$ est définie, pour $x = (x^1, \ldots, x^p)$ et $y = (y^1, \ldots, y^p)$, par*

$$d_q(x, y) = \left(\sum_{j=1}^{p} |x^j - y^j|^q \right)^{1/q}.$$

Quelques cas particuliers sont fondamentaux en pratique :

— $q = 1$: distance de Manhattan

$$d_1(x, y) = \sum_{j=1}^{p} |x^j - y^j|;$$

— $q = 2$: distance euclidienne

$$d_2(x, y) = \left(\sum_{j=1}^{p} (x^j - y^j)^2 \right)^{1/2};$$

— $q = \infty$: distance du maximum (ou ℓ_∞)

$$d_\infty(x, y) = \max_{1 \leq j \leq p} |x^j - y^j|.$$

Exemple 4.1.1 (Distance euclidienne en dimension p)**.** *Dans $\mathbb{R}^p$, la distance euclidienne mesure la longueur du segment reliant x à y. Elle est adaptée lorsque les variables sont comparables (mêmes unités ou variables préalablement normalisées).*

Remarque (importance du prétraitement). Lorsque les variables n'ont pas la même échelle (par exemple une variable en euros et une autre en années), la distance euclidienne peut être dominée par la variable de plus grande amplitude. Cela justifie en pratique des normalisations (centrage-réduction) avant d'utiliser une distance de type Minkowski.

4.1.3 Similarités

Dans certains modèles (méthodes à noyaux, graphes de voisinage, spectral clustering), on préfère travailler avec une *similarité* plutôt qu'avec une distance : deux observations sont d'autant plus proches que leur similarité est grande.

Définition 4.1.4 (Similarité)**.** *On appelle* similarité *sur $\mathcal{X}$ toute application*

$$s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$$

telle que $s(x, y)$ est d'autant plus grande que x et y sont jugés "semblables" (au sens choisi par l'utilisateur ou le modèle).

Transformer une distance en similarité. Une manière classique de construire une similarité consiste à composer une distance avec une fonction décroissante. Deux constructions usuelles sont :

— **Similarité gaussienne (noyau RBF)** :

$$s(x,y) = \exp\left(-\frac{d(x,y)^2}{2h^2}\right),$$

où $h > 0$ est un paramètre d'échelle. Cette transformée se révèle en pratique assez pertinente pour plusieurs raison :

(i) elle renvoie toujours une valeur dans $(0,1]$;
(ii) elle vaut 1 lorsque $x = y$, traduisant une similarité maximale ;
(iii) elle décroît rapidement quand la distance augmente, ce qui favorise un comportement "local".

— **Similarité rationnelle** :

$$s(x,y) = \frac{1}{1+d(x,y)}.$$

Cette transformation est plus douce : la décroissance est plus lente et peut être utile lorsque l'on souhaite conserver l'influence de points plus éloignés.

Choix du paramètre h. Dans la similarité gaussienne, h joue un rôle clé : un petit h rend la similarité très locale (seuls les points très proches comptent), un grand h lisse davantage la notion de proximité. Ce paramètre se règle généralement par validation.

4.2 Topologie et proximités pour des données catégorielles

Les variables catégorielles apparaissent très fréquemment en pratique : sexe, groupe sanguin, stade d'une maladie, catégorie socio-professionnelle, type de produit, etc. Le problème est qu'il n'existe pas toujours de notion naturelle de distance entre catégories.

Par exemple, la catégorie "cadre de la fonction publique" est-elle plus proche de "artisan" ou de "employé de commerce" ? Sans information externe (hiérarchie, sémantique, ontologie), ce type de proximité est ambigu.

4.2.1 Distance minimale : la métrique discrète

Lorsque l'on ne dispose que de l'information "égal / différent", on peut utiliser la distance la plus simple possible.

Définition 4.2.1 (Distance discrète)**.** *Soit $\mathcal{X}$ un ensemble (par exemple un ensemble de catégories). La* distance discrète *est définie par*

$$d_{disc}(x,y) = \begin{cases} 0 & \text{si } x = y, \\ 1 & \text{sinon.} \end{cases}$$

Cette distance formalise l'idée que deux catégories ne sont proches que si elles sont identiques. Elle est souvent suffisante lorsque les catégories sont nominales sans structure additionnelle.

Variables catégorielles multiples. Si l'observation possède plusieurs attributs catégoriels, par exemple $X = (X^1, \dots, X^p)$ où chaque X^j prend des valeurs dans un ensemble fini, une distance naturelle est la distance de Hamming :

$$d_{\mathrm{Ham}}(x,y) = \sum_{j=1}^{p} \mathbb{1}_{\{x^j \neq y^j\}},$$

qui compte le nombre de coordonnées différentes.

4.2.2 Encodage *one-hot* et retour aux distances euclidiennes

Pour appliquer des méthodes qui exigent des vecteurs réels (régression, SVM linéaire, k-NN avec norme ℓ_q, réseaux de neurones), on transforme souvent les variables catégorielles en variables numériques via un encodage.

Définition 4.2.2 (Encodage *one-hot*)**.** *Soit C le nombre de modalités possibles d'une variable catégorielle $Z \in \{1, \dots, C\}$. On appelle* encodage one-hot *la représentation de Z par un vecteur binaire $e(Z) \in \{0,1\}^C$ défini par*

$$e(Z) = (e_1, \dots, e_C) \quad \text{où} \quad e_k = \mathbb{1}_{\{Z=k\}}.$$

Ainsi, une modalité est représentée par un vecteur qui contient un unique 1 à la position correspondant à la modalité, et des 0 ailleurs.

Exemple 4.2.1 (Encodage one-hot)**.** *On considère la variable catégorielle* `Couleur` *pouvant prendre $C = 3$ modalités :*

$$\{\textit{Rouge},\ \textit{Vert},\ \textit{Bleu}\}.$$

On fixe l'ordre (Rouge, Vert, Bleu). L'encodage one-hot donne :

$$e(\textit{Rouge}) = (1,0,0), \qquad e(\textit{Vert}) = (0,1,0), \qquad e(\textit{Bleu}) = (0,0,1).$$

Remarque (distance induite par l'encodage). Une fois encodées, les catégories vivent dans $\mathbb{R}^C$ et l'on peut utiliser une norme ℓ_q. Par exemple, avec la distance euclidienne :

$$d_2\big(e(a), e(b)\big) = \begin{cases} 0 & \text{si } a = b, \\ \sqrt{2} & \text{si } a \neq b, \end{cases}$$

ce qui revient à distinguer seulement "égal / différent", mais dans un espace vectoriel compatible avec les méthodes numériques.

Message clé. La topologie (via une distance ou une similarité) n'est pas un choix secondaire : elle encode la notion de proximité que les algorithmes vont exploiter. Pour des variables quantitatives, les normes ℓ_q fournissent un cadre naturel. Pour des variables catégorielles, on utilise soit des distances discrètes (ou de Hamming), soit des encodages (one-hot) permettant de réutiliser des outils géométriques sur $\mathbb{R}^p$.

Chapitre 5

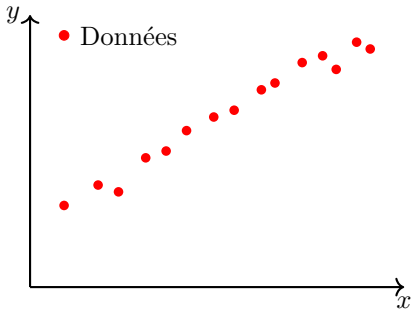
Comprendre le processus d'apprentissage d'un modèle

Avant d'entrer dans le détail des méthodes spécifiques au traitement automatique du langage naturel (NLP), il est indispensable de rappeler les principes fondamentaux qui gouvernent l'apprentissage d'un modèle statistique. En effet, tous les systèmes prédictifs en apprentissage automatique – de la régression linéaire aux réseaux de neurones profonds – reposent sur une même structure conceptuelle : définir un modèle, choisir une *fonction de perte* qui quantifie ses erreurs, puis ajuster ses paramètres afin de minimiser cette perte.

Ce chapitre introduit ces idées essentielles de manière simple mais rigoureuse, avant de les appliquer ultérieurement à des données textuelles.

5.1 Comprendre un modèle, c'est comprendre sa fonction de perte

Supposons que l'on dispose d'un ensemble de points de données observés, représentés dans le plan ci-dessous.



Un modèle peut être vu comme une fonction mathématique

$$f_\theta : \mathcal{X} \to \mathcal{Y},$$

où θ désigne l'ensemble de ses paramètres (poids, biais, etc.). Étant donné un jeu de données $\{(x_i, y_i)\}_{i=1}^n$, l'objectif de l'apprentissage consiste à trouver des paramètres θ tels que $f_\theta(x_i)$ se rapproche le plus possible de la valeur réelle y_i.

Exemple simple. Supposons que le nuage de points observé suggère une relation approximativement linéaire entre deux variables x et y. Un modèle naturel consiste alors à considérer une fonction affine :

$$\hat{y} = \theta_0 + \theta_1 x.$$

La qualité de cette approximation est mesurée par l'*énergie résiduelle*, c'est-à-dire la moyenne des carrés des écarts entre les valeurs prédites et les valeurs observées :

$$\mathcal{L}(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

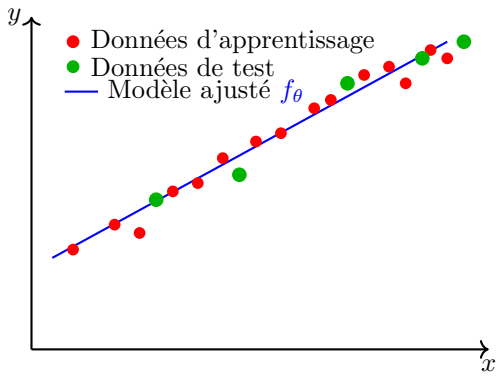
Cette quantité est appelée *fonction de perte*. Minimiser cette perte revient à déterminer la droite qui s'ajuste au mieux aux données au sens des moindres carrés.

Illustration : séparation apprentissage / test et évaluation de la perte. En apprentissage supervisé, les données disponibles sont généralement séparées en deux sous-ensembles :

— un **jeu d'apprentissage** (environ 80 %) utilisé pour estimer les paramètres du modèle ;

— un **jeu de test** (environ 20 %) utilisé pour évaluer sa capacité de généralisation.

La figure suivante illustre ce principe : les points rouges correspondent aux données d'apprentissage, tandis que les points verts représentent des données jamais vues par le modèle. La droite ajustée minimise la perte uniquement sur les données d'apprentissage.



La fonction de perte quantifie ainsi l'écart entre les prédictions du modèle et la réalité observée. Le modèle appris est celui qui minimise cette perte sur les données d'apprentissage.

5.2 Apprentissage et généralisation

Un enjeu central de l'apprentissage automatique n'est pas seulement de bien ajuster les données disponibles, mais de *généraliser* à de nouvelles données. Un modèle qui reproduit parfaitement les données d'apprentissage mais échoue sur des données inédites est dit *sur-ajusté* (*overfitting*). À l'inverse, un modèle trop simple qui ne parvient pas à capturer la structure des données est dit *sous-ajusté* (*underfitting*).

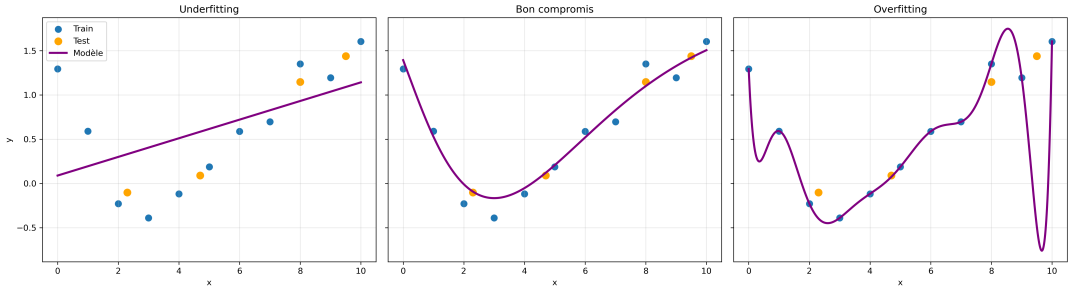


FIGURE 5.1 – Illustration du compromis biais–variance. À gauche, le modèle est sous-ajusté (biais élevé). Au centre, l'ajustement est équilibré. À droite, le modèle est sur-ajusté (variance élevée).

Compromis biais–variance. Ce phénomène est formalisé par le **compromis biais–variance**. Un modèle trop rigide présente un biais élevé, tandis qu'un modèle trop flexible présente une variance élevée. L'objectif de l'apprentissage statistique est de trouver un équilibre permettant de minimiser l'erreur globale de prédiction.

5.2.1 Train–test *split*

Afin de limiter le risque de sur-apprentissage et d'évaluer de manière réaliste les performances d'un modèle, il est d'usage de scinder le jeu de données initial en deux sous-ensembles disjoints : un *jeu d'apprentissage* (*training set*) et un *jeu de test* (*test set*). Cette séparation peut être effectuée selon différentes proportions, par exemple 50%–50%, 70%–30% ou 80%–20%, en fonction de la taille du jeu de données et de la complexité du modèle étudié.

Le jeu d'apprentissage est utilisé exclusivement pour estimer les paramètres du modèle, tandis que le jeu de test est réservé à l'évaluation finale de ses performances. Cette dissociation est cruciale : évaluer un modèle sur les données qui ont servi à son apprentissage conduit généralement à une estimation trop optimiste de ses capacités prédictives, en particulier pour des modèles flexibles susceptibles de mémoriser les données.

D'un point de vue théorique, si $\mathbb{P}$ désigne la distribution réelle des données et $\mathcal{L}$ une fonction de perte donnée, la qualité intrinsèque d'un modèle paramétré par θ est mesurée par le *risque*

$$R(\theta) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathcal{L}(f_\theta(x), y)\big].$$

Ce risque correspond à l'erreur moyenne que l'on commettrait sur une infinité de nouvelles observations issues de la même population. En pratique, la distribution $\mathbb{P}$ est inconnue, et le risque ne peut être évalué qu'à partir des données disponibles.

On introduit alors des estimateurs empiriques du risque, calculés séparément sur le jeu d'apprentissage et sur le jeu de test :

$$\hat{R}_{\text{train}}(\theta) = \frac{1}{n_{\text{train}}} \sum_{i\in\mathcal{I}_{\text{train}}} \mathcal{L}\big(f_\theta(x_i), y_i\big), \qquad \hat{R}_{\text{test}}(\theta) = \frac{1}{n_{\text{test}}} \sum_{i\in\mathcal{I}_{\text{test}}} \mathcal{L}\big(f_\theta(x_i), y_i\big).$$

Lorsque le modèle généralise correctement, ces deux quantités sont de même ordre de grandeur. À l'inverse, un écart important entre l'erreur d'apprentissage et l'erreur de test est le signe d'un sur-apprentissage.

Rôle des métriques d'évaluation. Le jeu de test joue également un rôle central dans le choix et la comparaison des modèles, à travers l'introduction de *métriques d'évaluation*. Une métrique est une fonction qui quantifie la qualité des prédictions produites par un modèle, indépendamment du processus d'apprentissage. Le choix de la métrique dépend de la nature du problème : en régression, on utilise par exemple l'erreur quadratique moyenne ou le coefficient de détermination R^2 ; en classification, on considère des mesures telles que l'exactitude (*accuracy*), la précision, le rappel ou l'aire sous la courbe ROC.

L'évaluation de ces métriques sur le jeu de test permet de comparer objectivement différents modèles ou différentes configurations d'un même modèle. Elle constitue une étape essentielle du processus de modélisation, car elle fournit une estimation non biaisée de la performance attendue sur des données nouvelles, et guide ainsi le choix du modèle final à retenir.

Illustration du découpage des données.

Définition 5.2.1 (Train–test split)**.** *On partitionne le jeu de données initial $\mathcal{D}$ en deux sous-ensembles disjoints :*

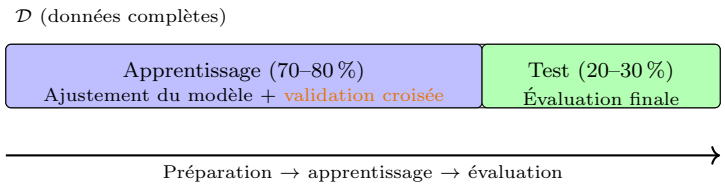
$$\mathcal{D} = \mathcal{D}_{train} \cup \mathcal{D}_{test}, \qquad \mathcal{D}_{train} \cap \mathcal{D}_{test} = \varnothing.$$

Le jeu d'apprentissage $\mathcal{D}_{train}$ est utilisé pour ajuster les paramètres du modèle, tandis que le jeu de test $\mathcal{D}_{test}$ est réservé à l'évaluation de ses performances prédictives sur des données non observées.

Dans un cadre général de modélisation statistique ou d'apprentissage automatique :

— $\mathcal{D}_{\text{train}}$ sert à estimer les paramètres du modèle (par exemple des coefficients de régression ou des poids neuronaux), ainsi qu'à ajuster d'éventuels hyperparamètres ;

— $\mathcal{D}_{\text{test}}$ permet de mesurer objectivement la qualité de prédiction du modèle à l'aide de métriques appropriées (erreur quadratique, précision, rappel, AUC, calibration, etc.).



Remarque 2 (Bonnes pratiques). Dans les problèmes de classification déséquilibrée ou de survie, il est recommandé de réaliser un découpage *stratifié*, afin de préserver des proportions comparables de classes ou d'évènements entre les jeux d'apprentissage et de test. Cette précaution garantit une évaluation plus fiable et évite des biais artificiels dans les métriques de performance.

5.2.2 Apprentissage comme problème d'optimisation et lecture biais–variance

Décomposition biais–variance pour la perte quadratique. Dans de nombreux problèmes de régression, on choisit comme fonction de perte l'écart quadratique

$$\mathcal{L}\big(f_\theta(x), y\big) = \big(y - f_\theta(x)\big)^2.$$

Ce choix est particulièrement naturel pour deux raisons. D'une part, il pénalise plus fortement les erreurs importantes qu'une perte linéaire, ce qui correspond souvent à l'intuition qu'une grande erreur est *beaucoup* plus grave qu'une petite erreur. D'autre part, sur le plan mathématique, la perte quadratique est convexe et différentiable, ce qui la rend compatible avec des méthodes d'optimisation efficaces (notamment la descente de gradient) et conduit, dans de nombreux cas, à des formules analytiques simples. Dans ce cadre, le risque théorique associé à un modèle f_θ s'écrit

$$R(\theta) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}\Big[\Big(Y - f_\theta(X)\Big)^2\Big].$$

Lorsque l'on s'intéresse à la capacité de généralisation, il est fréquent d'analyser l'erreur attendue *en un point* x fixé, en considérant la quantité

$$\mathbb{E}\Big[(Y - \widehat{f}(x))^2\Big],$$

où $\widehat{f}$ désigne le modèle appris à partir d'un échantillon (et est donc une variable aléatoire : un autre échantillon conduirait en général à un autre modèle). Cette écriture met en évidence que l'erreur provient de deux sources distinctes : l'aléa sur les données (X, Y) et l'aléa lié au processus d'apprentissage.

Pour formaliser cela, introduisons d'abord deux notions fondamentales.

Définition 5.2.2 (Biais et variance d'un prédicteur en un point)**.** *Soit $\widehat{f}(x)$ la prédiction fournie par un modèle appris sur un échantillon aléatoire. On définit :*

$$\text{Biais}\big(\widehat{f}(x)\big) = \mathbb{E}\left[\widehat{f}(x)\right] - f^\star(x), \qquad \text{Var}\big(\widehat{f}(x)\big) = \mathbb{E}\Big[\big(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]\big)^2\Big],$$

où $f^\star(x) = \mathbb{E}[Y \mid X = x]$ est la fonction de régression *(la meilleure prédiction possible au sens de la perte quadratique).*

La fonction $f^\star$ joue un rôle central : elle correspond à la limite théorique que l'on ne peut pas dépasser, car même avec un modèle parfait, la variable Y conserve une part d'aléa conditionnel lorsqu'on connaît $\{X = x\}$.

Proposition 5.2.1 (Décomposition biais–variance sous perte quadratique)**.** *On a la décomposition suivante :*

$$\mathbb{E}\Big[(Y-\widehat{f}(x))^2\Big] = \underbrace{\mathrm{Var}\big[Y \mid X=x\big]}_{\text{bruit irréductible}} + \underbrace{\mathrm{Var}\Big[\widehat{f}(x)\Big]}_{\text{variance du modèle}} + \underbrace{\mathrm{Biais}\Big(\widehat{f}(x)\Big)^2}_{\text{biais}^2}. \tag{5.1}$$

Preuve. On écrit d'abord la décomposition

$$Y-\widehat{f}(x) = \big(Y-f^\star(x)\big) + \big(f^\star(x)-\widehat{f}(x)\big).$$

En élevant au carré :

$$(Y-\widehat{f}(x))^2 = \big(Y-f^\star(x)\big)^2 + \big(\widehat{f}(x)-f^\star(x)\big)^2 + 2\big(Y-f^\star(x)\big)\big(f^\star(x)-\widehat{f}(x)\big).$$

On prend l'espérance. Conditionnellement à $\{X=x\}$, on a $\mathbb{E}[Y-f^\star(x) \mid X=x]=0$ par définition de $f^\star(x)=\mathbb{E}[Y \mid X=x]$. De plus, $\widehat{f}(x)$ dépend uniquement de l'échantillon d'apprentissage et ne dépend pas de la réalisation ponctuelle de Y au point x (dans l'expérience de généralisation). Ainsi, le terme croisé s'annule à l'espérance, et l'on obtient :

$$\mathbb{E}\Big[(Y-\widehat{f}(x))^2\Big] = \mathbb{E}\Big[(Y-f^\star(x))^2\Big] + \mathbb{E}\Big[\big(\widehat{f}(x)-f^\star(x)\big)^2\Big].$$

Le premier terme vaut $\mathrm{Var}[Y \mid X=x]$ (c'est le bruit irréductible). Pour le second terme, on applique l'identité

$$\mathbb{E}\Big[\big(\widehat{f}(x)-f^\star(x)\big)^2\Big] = \mathbb{E}\Big[\big(\widehat{f}(x)-\mathbb{E}[\widehat{f}(x)]\big)^2\Big] + \big(\mathbb{E}[\widehat{f}(x)]-f^\star(x)\big)^2,$$

c'est-à-dire

$$\mathbb{E}\Big[\big(\widehat{f}(x)-f^\star(x)\big)^2\Big] = \mathrm{Var}\big(\widehat{f}(x)\big) + \mathrm{Biais}\big(\widehat{f}(x)\big)^2.$$

En combinant les deux égalités, on obtient la décomposition annoncée. □

Cette identité est fondamentale : elle montre que même si l'on disposait d'un modèle parfait, une partie de l'erreur resterait inévitable (le terme $\mathrm{Var}[Y \mid X=x]$). Le rôle de l'apprentissage statistique est donc de contrôler au mieux les deux autres termes. Les modèles très flexibles tendent à réduire le biais mais augmentent la variance, tandis que les modèles plus contraints font l'inverse. C'est précisément ce mécanisme qui explique l'existence d'un compromis biais–variance et justifie l'usage de techniques de régularisation.

Optimisation et compromis. De manière générale, apprendre un "meilleur modèle" revient à formuler et résoudre un problème d'optimisation. On se place dans une famille de modèles paramétriques $\{f_\theta : \theta \in \Theta\}$, où θ désigne l'ensemble des paramètres à estimer. Selon le contexte, θ peut représenter les coefficients d'une droite de régression, les poids d'un réseau de neurones, ou plus généralement les paramètres gouvernant la forme de la fonction de prédiction.

Dans un cadre de régression quantitative, le critère le plus fréquemment retenu est la perte quadratique. L'apprentissage consiste alors à choisir le paramètre θ qui minimise le risque associé :

$$\theta^\star \in \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}\Big[\Big(Y-f_\theta(X)\Big)^2\Big].$$

Comme on l'a vu précédemment, sous la perte quadratique, ce risque peut être décomposé en une somme de trois termes : un bruit irréductible, indépendant du modèle, une variance et un biais au carré. Le terme de bruit étant constant par rapport au choix de θ, il n'intervient pas dans la recherche du minimum. Ainsi, le problème d'apprentissage peut s'interpréter comme la minimisation du compromis

$$\operatorname*{argmin}_{\theta\in\Theta} \Big\{\mathrm{Var}\big(\widehat{f}_\theta(x)\big) + \mathrm{Biais}\big(\widehat{f}_\theta(x)\big)^2\Big\}.$$

Cette écriture est particulièrement éclairante sur le plan conceptuel. Elle montre que l'apprentissage cherche simultanément à contrôler deux sources d'erreur de nature différente : la variance, qui mesure la sensibilité du modèle aux fluctuations de l'échantillon d'apprentissage, et le biais, qui quantifie l'erreur systématique introduite par une modélisation trop restrictive. Le critère de minimisation attribue un poids identique à ces deux composantes, ce qui correspond à une intuition naturelle : un bon modèle doit être à la fois suffisamment flexible pour approcher la fonction cible, et suffisamment stable pour produire des prédictions robustes sur de nouvelles données.

Cette lecture biais–variance permet ainsi de comprendre pourquoi l'optimisation d'une simple perte empirique conduit, en pratique, à des choix de modèles et de paramètres qui reflètent un équilibre subtil entre complexité et robustesse. Elle constitue un cadre conceptuel unificateur pour analyser aussi bien les modèles linéaires que les modèles plus complexes, tels que les arbres de décision ou les réseaux de neurones.

5.3 Optimisation par descente de gradient

Une fois la fonction de perte définie, l'apprentissage du modèle consiste à la minimiser. Dans la majorité des cas, cette minimisation est réalisée à l'aide de la **descente de gradient**.

Principe. Soit $\theta \in \mathbb{R}^d$ le vecteur des paramètres et $\mathcal{L}(\theta)$ une fonction différentiable. À partir d'une valeur initiale $\theta^{(0)}$, la descente de gradient itérative s'écrit :

$$\theta^{(t+1)} = \theta^{(t)} - \eta \, \nabla_\theta \mathcal{L}(\theta^{(t)}),$$

où $\eta > 0$ est le *taux d'apprentissage*. L'algorithme est itéré jusqu'à convergence, c'est-à-dire jusqu'à ce que les paramètres estimés se stabilisent. Plus précisément, on arrête les itérations lorsque la variation entre deux itérations successives devient suffisamment faible, par exemple lorsque $\|\theta^{(t+1)} - \theta^{(t)}\| \leq \varepsilon$ où $\varepsilon > 0$ est un seuil de tolérance fixé à l'avance. En pratique, afin de garantir l'arrêt de l'algorithme même en l'absence de convergence stricte, on impose également un nombre maximal d'itérations.

Intuition géométrique. La descente de gradient peut être interprétée comme le mouvement d'une bille roulant le long d'un paysage d'énergie, cherchant à atteindre le point le plus bas.

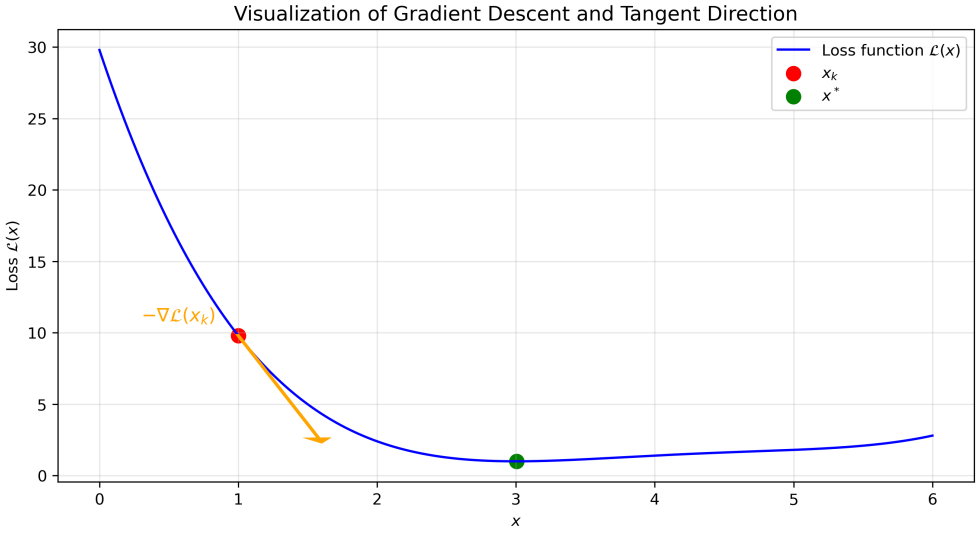


FIGURE 5.2 – Illustration de la descente de gradient : le vecteur $-\nabla\ell(x_k)$ pointe vers le minimum x^*.

La descente de gradient est l'un des algorithmes d'optimisation les plus fondamentaux en apprentissage automatique. Elle permet d'optimiser des fonctions complexes et de grande dimension, pour lesquelles aucune solution analytique n'est disponible. Apprendre un modèle revient à minimiser une énergie, jusqu'à atteindre un état d'équilibre correspondant aux paramètres optimaux.

Chapitre 6

La régression linéaire

6.1 Contexte et motivation

La régression linéaire occupe une place centrale en statistique et en modélisation des données. Elle constitue bien souvent le premier modèle quantitatif étudié, non seulement parce qu'elle est mathématiquement accessible, mais surtout parce qu'elle formalise une idée extrêmement naturelle : expliquer une variable d'intérêt à partir d'une ou plusieurs autres variables observées.

Dans de très nombreuses situations concrètes, on cherche en effet à comprendre comment une grandeur Y évolue en fonction d'une autre grandeur X. Il peut s'agir, par exemple, de relier un salaire à un niveau d'études, une consommation énergétique à une température extérieure, ou encore un rendement agricole à une quantité d'engrais. La régression linéaire propose un cadre mathématique simple pour analyser ce type de relations, avec un formalisme que la littérature mathématique maîtrise grandement.

Données observées. On suppose disposer d'un échantillon de données constitué de n observations indépendantes, que l'on note

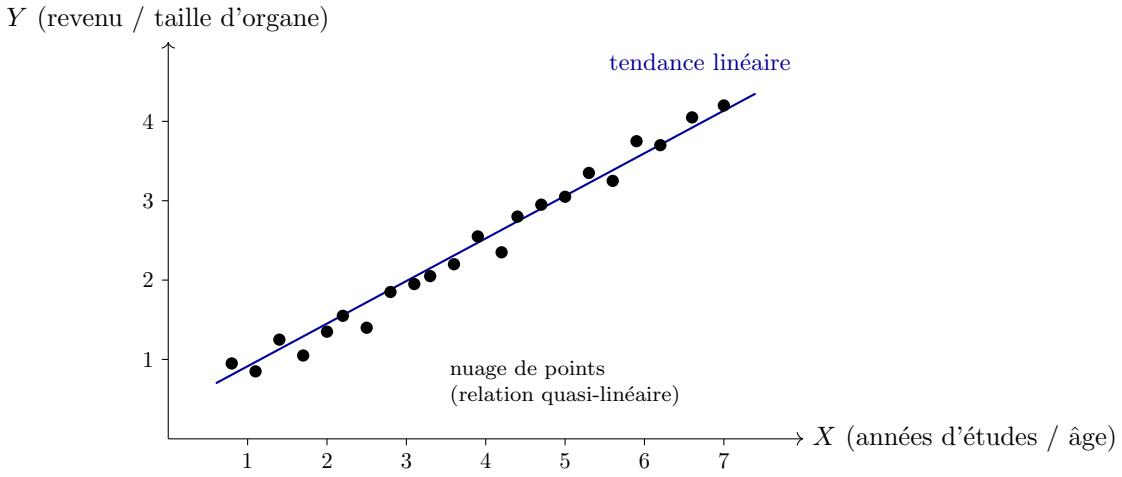
$$\big\{(X_1, Y_1), \dots, (X_n, Y_n)\big\}.$$

Pour chaque individu ou unité statistique observée, la variable $X_i \in \mathbb{R}$ représente une variable explicative quantitative, tandis que $Y_i \in \mathbb{R}$ correspond à la variable réponse que l'on cherche à analyser ou à expliquer. La taille n de l'échantillon joue un rôle fondamental, car elle conditionne à la fois la précision des estimations et la fiabilité des conclusions que l'on pourra tirer du modèle. On discutera plus tard de la notion d'inférence statistique qui permet de comprendre comment la quantité de données disponibles influence la variabilité des estimations et, par conséquent, la confiance que l'on peut accorder aux résultats du modèle.

Avant toute modélisation, une étape indispensable consiste à explorer graphiquement les données. La représentation la plus naturelle dans ce contexte est le *nuage de points*, obtenu en plaçant les couples (X_i, Y_i) dans le plan. Cette visualisation permet de se faire une première idée de la relation existant entre les deux variables.

Dans de nombreux cas pratiques, le nuage de points met en évidence une tendance globale : même si les points ne sont pas parfaitement alignés, on observe qu'ils semblent se répartir autour d'une droite. Cette observation empirique constitue le point de départ de la régression linéaire.

Exemple 6.1.1. *En économie, on observe fréquemment que le revenu moyen augmente avec le nombre d'années d'études. En biologie, la taille d'un organe peut croître de manière approximativement linéaire avec l'âge sur une certaine période. Dans ces situations, le nuage de points suggère clairement une relation linéaire entre X et Y.*



Face à une telle configuration, on se demande naturellement : peut-on résumer cette relation par une droite, et si oui, comment choisir cette droite de manière optimale ? Autrement dit, laquelle représente le mieux la structure observée dans les données ?

Deux objectifs majeurs. La régression linéaire répond à cette question en poursuivant deux objectifs complémentaires :

1. Le premier est un objectif de *description*. Il s'agit de résumer la relation moyenne entre Y et X à l'aide d'une formule simple, interprétable et synthétique. La droite de régression fournit alors une approximation globale de la tendance centrale des données, en filtrant le bruit et les fluctuations individuelles.
2. Le second objectif est de nature *inférentielle et prédictive*. Une fois le modèle ajusté, on cherche à quantifier l'effet de la variable explicative X sur la variable réponse Y, c'est-à-dire à mesurer comment une variation de X se traduit en moyenne par une variation de Y. Par ailleurs, le modèle permet de prédire la valeur de Y associée à une nouvelle valeur de X, ce qui constitue un enjeu central dans de nombreuses applications pratiques.

Ces deux motivations – description et inférence – justifient l'importance de la régression linéaire et expliquent pourquoi elle constitue un outil fondamental de l'analyse statistique.

6.2 Régression linéaire simple

Le modèle. On commence par le cas le plus simple : une seule variable explicative.

Définition 6.2.1. *Le modèle de régression linéaire simple est défini par :*

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

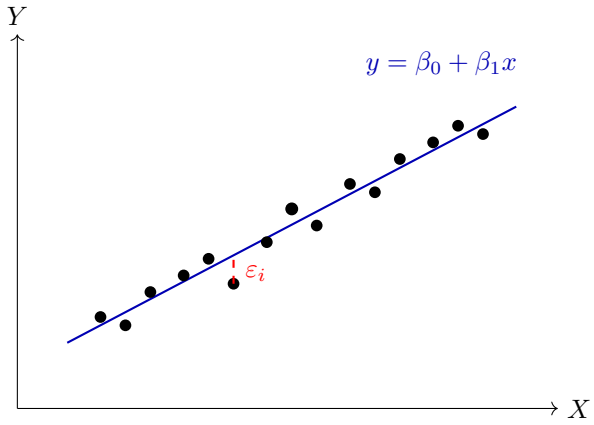
où :

- $\beta_0 \in \mathbb{R}$ *est l'ordonnée à l'origine (intercept),*
- $\beta_1 \in \mathbb{R}$ *est la pente,*
- ε *est un terme d'erreur aléatoire.*

Interprétation géométrique. Le modèle de régression linéaire simple peut être interprété de manière très intuitive à l'aide d'une représentation géométrique dans le plan. Chaque observation (X_i, Y_i) est représentée par un point, et le modèle suppose que la relation moyenne entre X et Y peut être décrite par une droite d'équation

$$y = \beta_0 + \beta_1 x.$$

Graphiquement, cette droite représente la tendance centrale autour de laquelle les données observées se répartissent. En pratique, les points ne sont presque jamais parfaitement alignés : ils se situent au voisinage de la droite, avec une certaine dispersion.



Pour une observation donnée X_i, le modèle fournit une valeur prédite

$$\hat{Y}_i = \beta_0 + \beta_1 X_i,$$

correspondant au point situé sur la droite. La différence entre la valeur observée Y_i et cette valeur prédite est appelée *résidu* et est notée ε_i. Sur la figure, ce résidu est représenté par un segment vertical en pointillé rouge : il mesure l'écart entre le point observé et la droite de régression.

Pourquoi introduire un terme d'erreur ? L'introduction du terme d'erreur ε est essentielle pour rendre le modèle réaliste. En effet, il serait illusoire de supposer que la variable explicative X suffit à elle seule à déterminer parfaitement la valeur de Y. De nombreux facteurs, non observés ou non mesurables, influencent généralement le phénomène étudié. À cela s'ajoutent la variabilité intrinsèque du système ainsi que les inévitables erreurs de mesure.

Le terme ε permet donc de regrouper l'ensemble de ces sources de variabilité dans une composante aléatoire, distincte de la relation déterministe portée par la droite $\beta_0 + \beta_1 X$.

Dans le cadre classique de la régression linéaire, on fait l'hypothèse que ce terme d'erreur suit une loi normale centrée :

$$\varepsilon \sim \mathcal{N}(0, \sigma^2).$$

La condition d'espérance nulle traduit le fait que, en moyenne, la droite décrit correctement la relation entre X et Y, tandis que le paramètre σ^2 mesure l'ampleur de la dispersion des observations autour de cette droite.

Remarque 3. L'hypothèse de normalité est justifiée à la fois par des considérations théoriques – en particulier le théorème central limite – et par sa commodité mathématique, qui permet de développer une théorie statistique complète de l'estimation et de l'inférence.

Le paramètre σ^2, inconnu en pratique, devra être estimé à partir des données, au même titre que les coefficients β_0 et β_1.

6.3 Estimation par la méthode des moindres carrés

Principe. L'idée est de choisir la droite qui minimise l'**énergie des erreurs**, c'est-à-dire la somme des carrés des résidus.

Définition 6.3.1. *Les estimateurs des moindres carrés* $(\hat{\beta}_0, \hat{\beta}_1)$ *sont définis par :*

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \big(Y_i - (\beta_0 + \beta_1 X_i)\big)^2 \right\}.$$

Cette approche pénalise fortement les grandes erreurs et conduit à un problème mathématique bien posé.

Démonstration des estimateurs des moindres carrés. On introduit la *fonction objective* (ou fonction de coût) associée au critère des moindres carrés :

$$J(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(Y_i - (\beta_0 + \beta_1 X_i)\big)^2.$$

Notre objectif est de déterminer le couple $(\hat{\beta}_0, \hat{\beta}_1)$ qui minimise J sur $\mathbb{R}^2$.

Définition 6.3.2 (Moyennes empiriques et centre de gravité)**.** *Soient* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *des observations réelles. On définit les moyennes empiriques des variables* X *et* Y *par*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Le point $(\bar{X}, \bar{Y})$ *est appelé le* centre de gravité *(ou barycentre) du nuage de points associé aux données. Il correspond au barycentre des points* (X_i, Y_i) *lorsque toutes les observations sont affectées du même poids. Cette interprétation géométrique jouera un rôle central dans l'étude de la régression linéaire, notamment parce que la droite de régression passe toujours par ce point.*

Proposition 6.3.1. *Supposons que* $\sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$ *(c'est-à-dire que les* X_i *ne sont pas tous égaux). Alors la fonction* J *admet un unique minimiseur* $(\hat{\beta}_0, \hat{\beta}_1)$*, donné explicitement par*

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Preuve. En plusieurs étapes :

1) Régularité et convexité du critère. La fonction J est un polynôme (somme de carrés) en (β_0, β_1), donc elle est de classe $\mathcal{C}^2$. De plus, comme il s'agit d'une somme de termes quadratiques, on s'attend à une convexité globale. Nous allons de toute façon déterminer explicitement le point critique, puis vérifier qu'il correspond bien à un minimum *(unique sous l'hypothèse* $\sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$*).*

2) Calcul des dérivées partielles. On dérive J par rapport à β_0 puis à β_1.
Pour β_0 :

$$\frac{\partial J}{\partial \beta_0}(\beta_0, \beta_1) = \sum_{i=1}^{n} 2\big(Y_i - (\beta_0 + \beta_1 X_i)\big) \cdot (-1) = -2\sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_i\big).$$

Pour β_1 :

$$\frac{\partial J}{\partial \beta_1}(\beta_0, \beta_1) = \sum_{i=1}^{n} 2\big(Y_i - (\beta_0 + \beta_1 X_i)\big) \cdot (-X_i) = -2\sum_{i=1}^{n} X_i\big(Y_i - \beta_0 - \beta_1 X_i\big).$$

Les conditions d'optimalité du premier ordre (annulation du gradient) imposent donc :

$$\begin{cases} \sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_i\big) = 0, \\ \sum_{i=1}^{n} X_i \big(Y_i - \beta_0 - \beta_1 X_i\big) = 0. \end{cases}$$

On appelle ce système les *équations normales* des moindres carrés.

3) Première équation : apparition du centre de gravité. Développons la première équation :

$$\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i = 0.$$

En divisant par n et en utilisant les définitions de $\bar{X}$ et $\bar{Y}$, on obtient :

$$\bar{Y} - \beta_0 - \beta_1 \bar{X} = 0, \qquad \text{c'est-à-dire} \qquad \beta_0 = \bar{Y} - \beta_1 \bar{X}.$$

Cette relation montre déjà un fait fondamental : *toute droite minimisante doit passer par le point* $(\bar{X}, \bar{Y})$, car si l'on remplace $x = \bar{X}$ dans $y = \beta_0 + \beta_1 x$, on obtient

$$\beta_0 + \beta_1 \bar{X} = \bar{Y}.$$

4) Seconde équation : recentrage et calcul de β_1. Partons de la seconde équation normale :

$$\sum_{i=1}^{n} X_i \big(Y_i - \beta_0 - \beta_1 X_i\big) = 0.$$

On substitue l'expression $\beta_0 = \bar{Y} - \beta_1 \bar{X}$ obtenue précédemment :

$$\sum_{i=1}^{n} X_i \Big(Y_i - (\bar{Y} - \beta_1 \bar{X}) - \beta_1 X_i\Big) = 0,$$

soit

$$\sum_{i=1}^{n} X_i \Big((Y_i - \bar{Y}) + \beta_1 (\bar{X} - X_i)\Big) = 0.$$

En développant :

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y}) + \beta_1 \sum_{i=1}^{n} X_i (\bar{X} - X_i) = 0.$$

Or

$$\sum_{i=1}^{n} X_i (\bar{X} - X_i) = \bar{X} \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i^2 = n\bar{X}^2 - \sum_{i=1}^{n} X_i^2.$$

À ce stade, une simplification décisive consiste à *recentrer* les données autour de $\bar{X}$ et $\bar{Y}$. L'idée est d'écrire $X_i = \big(X_i - \bar{X}\big) + \bar{X}$ et $Y_i = \big(Y_i - \bar{Y}\big) + \bar{Y}$ afin de faire apparaître des sommes de termes centrés, qui ont des propriétés simples (notamment $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$ et $\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0$).

Calcul du terme $\sum_{i=1}^{n} X_i (Y_i - \bar{Y})$. Écrivons $X_i = (X_i - \bar{X}) + \bar{X}$:

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y}) = \sum_{i=1}^{n} \big((X_i - \bar{X}) + \bar{X}\big)(Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) + \bar{X} \sum_{i=1}^{n} (Y_i - \bar{Y}).$$

Or $\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0$ par définition de $\bar{Y}$. Ainsi :

$$\sum_{i=1}^{n} X_i (Y_i - \bar{Y}) = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Calcul du terme $n\bar{X}^2 - \sum_{i=1}^n X_i^2$. On part de l'identité

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n \left(X_i^2 - 2\bar{X}X_i + \bar{X}^2\right) = \sum_{i=1}^n X_i^2 - 2\bar{X}\sum_{i=1}^n X_i + n\bar{X}^2.$$

Comme $\sum_{i=1}^n X_i = n\bar{X}$, on obtient

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - 2\bar{X} \cdot n\bar{X} + n\bar{X}^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2,$$

donc

$$n\bar{X}^2 - \sum_{i=1}^n X_i^2 = -\sum_{i=1}^n (X_i - \bar{X})^2.$$

5) Conclusion : formule explicite $\hat{\beta}_1$. La seconde équation normale devient alors :

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) + \beta_1\Big(-\sum_{i=1}^n (X_i - \bar{X})^2\Big) = 0,$$

c'est-à-dire

$$\beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

Sous l'hypothèse $\sum_{i=1}^n (X_i - \bar{X})^2 > 0$, on peut diviser et on obtient :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Enfin, en reportant dans $\beta_0 = \bar{Y} - \beta_1\bar{X}$, on obtient :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}.$$

6) Unicité du minimiseur. L'hypothèse $\sum_{i=1}^n (X_i - \bar{X})^2 > 0$ signifie que les X_i ne sont pas tous identiques : il existe une véritable variabilité de la variable explicative. Dans ce cas, le critère J est strictement convexe en β_1 (et donc en (β_0, β_1)), ce qui garantit l'unicité du minimiseur. □

Intuition géométrique. Le recentrage en $(\bar{X}, \bar{Y})$ n'est pas un simple artifice de calcul : il révèle la structure du problème. L'estimateur $\hat{\beta}_1$ compare la *co-variation* de X et Y autour de leurs centres de gravité (numérateur) à la *variabilité propre* de X autour de son centre de gravité (dénominateur). En particulier, si X varie beaucoup mais que Y ne varie pas en cohérence avec X, alors la pente estimée est proche de 0.

6.4 Régression linéaire multiple

Jusqu'à présent, nous avons étudié la situation où une seule variable explicative intervient dans la modélisation de la variable réponse. En pratique, cette hypothèse est souvent trop restrictive. De nombreux phénomènes dépendent simultanément de plusieurs facteurs, et ignorer une partie de cette information peut conduire à des modèles incomplets ou biaisés. La régression linéaire multiple vise précisément à étendre le cadre précédent afin de prendre en compte plusieurs variables explicatives agissant conjointement sur la variable d'intérêt.

6.4.1 Extension du modèle

On suppose désormais que, pour chaque observation, on dispose de p variables explicatives. L'échantillon de données s'écrit alors

$$\big\{(X_1^1, \dots, X_1^p, Y_1), \dots, (X_n^1, \dots, X_n^p, Y_n)\big\},$$

où X_i^j désigne la j-ième variable explicative associée à la i-ième observation.

Le modèle de régression linéaire multiple postule que la variable réponse Y dépend linéairement de ces p variables, selon la relation

$$Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon.$$

Le coefficient β_0 joue le rôle d'un terme constant (ou intercept), tandis que chaque coefficient β_j mesure l'effet marginal de la variable X^j sur Y, toutes choses égales par ailleurs. Le terme d'erreur ε regroupe, comme précédemment, l'ensemble des facteurs non pris en compte par le modèle ainsi que la variabilité intrinsèque du phénomène étudié.

Il est important de souligner que, contrairement au cas univarié, il n'existe plus de représentation géométrique simple dans le plan. La droite de régression est remplacée par un *hyperplan* dans un espace de dimension $p + 1$. C'est cette observation qui motive l'introduction d'une écriture matricielle, à la fois plus compacte et plus puissante.

6.4.2 Hypothèses du modèle de régression linéaire multiple

Le modèle de régression linéaire multiple repose sur un certain nombre d'hypothèses structurelles et probabilistes. Ces hypothèses ne sont pas de simples contraintes techniques : elles conditionnent la validité des résultats théoriques, l'interprétation des coefficients estimés et la fiabilité des outils d'inférence statistique associés au modèle.

Il est donc essentiel de les expliciter clairement avant de procéder à l'estimation des paramètres.

Hypothèse de linéarité. La première hypothèse est celle de *linéarité du modèle*. Elle stipule que la variable réponse peut être exprimée comme une combinaison linéaire des variables explicatives, à laquelle s'ajoute un terme d'erreur :

$$Y = \beta_0 + \beta_1 X^1 + \cdots + \beta_p X^p + \varepsilon.$$

Il est important de souligner que cette hypothèse porte sur la forme du modèle *par rapport aux paramètres* β_j, et non sur la nature intrinsèque des variables. En pratique, il est tout à fait possible d'introduire des transformations non linéaires des covariables (par exemple X^2, $\log(X)$, etc.), tant que le modèle reste linéaire en les coefficients.

Hypothèse d'indépendance des observations. On suppose que les observations $(X_i^1, \dots, X_i^p, Y_i)$ sont indépendantes les unes des autres. Cette hypothèse est généralement justifiée lorsque les données proviennent d'un échantillonnage aléatoire et que chaque observation correspond à une unité statistique distincte.

L'indépendance est cruciale pour l'analyse de la variabilité des estimateurs et pour le développement de l'inférence statistique. En présence de dépendances (données temporelles ou spatiales, par exemple), le modèle linéaire classique n'est plus adapté sans ajustements spécifiques.

Hypothèse d'espérance conditionnelle nulle. On suppose que le terme d'erreur vérifie

$$\mathbb{E}[\varepsilon \mid X^1, \dots, X^p] = 0.$$

Cette hypothèse signifie que, conditionnellement aux variables explicatives, le modèle capture correctement la valeur moyenne de Y. Autrement dit, aucune information systématique expliquant Y ne doit être contenue dans le terme d'erreur.

Sur le plan statistique, cette condition garantit que les estimateurs des moindres carrés sont non biaisés.

Hypothèse d'homoscédasticité. Le modèle suppose que la variance du terme d'erreur est constante :

$$\mathrm{Var}[\varepsilon \mid X^1, \ldots, X^p] = \sigma^2.$$

Cette hypothèse, appelée *homoscédasticité*, signifie que la dispersion des erreurs ne dépend pas du niveau des variables explicatives. En pratique, une violation de cette hypothèse se traduit par des résidus dont la variance varie avec X, ce qui peut affecter l'efficacité des estimateurs et fausser les tests statistiques.

Hypothèse de non-colinéarité parfaite. On suppose enfin que les variables explicatives ne sont pas liées par des relations de dépendance linéaire exacte. Mathématiquement, cela revient à supposer que la matrice $\mathbb{X}^\top \mathbb{X}$ est inversible.

Cette hypothèse garantit que chaque variable explicative apporte une information propre au modèle et que les coefficients β_j sont identifiables. En présence de colinéarité parfaite, il devient impossible de distinguer l'effet individuel des variables concernées.

Hypothèse de normalité des erreurs (optionnelle). Dans de nombreux développements, on fait l'hypothèse supplémentaire que

$$\varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Cette hypothèse n'est pas nécessaire pour définir ni calculer les estimateurs des moindres carrés. En revanche, elle joue un rôle central dans l'établissement des intervalles de confiance et des tests statistiques, qui seront abordés ultérieurement.

Remarque 4. Les hypothèses précédentes constituent le cadre classique de la régression linéaire multiple. En pratique, elles doivent être systématiquement confrontées aux données à l'aide d'outils diagnostiques (analyse des résidus, tests, graphiques). Lorsque certaines hypothèses sont violées, des extensions ou des modèles alternatifs peuvent être envisagés.

Exemple 6.4.1 (Exemple concret de régression linéaire multiple)**.** *On considère une étude simplifiée portant sur la réussite universitaire d'étudiants. Pour chaque étudiant, on observe :*

- X^1 *: le nombre d'heures de travail personnel par semaine,*
- X^2 *: le nombre d'heures de cours suivies par semaine,*
- X^3 *: une note moyenne obtenue l'année précédente.*

On dispose donc de $p = 3$ variables explicatives.
La variable réponse est ici bidimensionnelle :

$$Y = (Y^{(1)}, Y^{(2)}) \in \mathbb{R}^2,$$

où :

- $Y^{(1)}$ *représente la moyenne obtenue au premier semestre,*
- $Y^{(2)}$ *représente la moyenne obtenue au second semestre.*

Pour un étudiant donné, une observation s'écrit donc :

$$(X^1, X^2, X^3, Y^{(1)}, Y^{(2)}).$$

Dans le cadre strict de la régression linéaire multiple telle qu'elle est étudiée dans ce chapitre, on traite séparément chaque composante de la variable réponse. Autrement dit, on ajuste deux modèles distincts :

$$Y^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X^1 + \beta_2^{(1)} X^2 + \beta_3^{(1)} X^3 + \varepsilon^{(1)},$$
$$Y^{(2)} = \beta_0^{(2)} + \beta_1^{(2)} X^1 + \beta_2^{(2)} X^2 + \beta_3^{(2)} X^3 + \varepsilon^{(2)}.$$

Chaque composante de Y est donc expliquée par les mêmes variables explicatives, mais avec ses propres coefficients. L'étude conjointe de plusieurs variables réponses relève de modèles plus avancés, qui dépassent le cadre de ce cours.

6.4.3 Écriture matricielle du modèle

Afin de simplifier les notations et de mettre en évidence la structure algébrique du problème, on regroupe l'ensemble des variables explicatives dans une matrice dite *matrice de design*, notée $\mathbb{X}$. Celle-ci est définie par

$$\mathbb{X} = \begin{pmatrix} 1 & X_1^1 & \dots & X_1^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n^1 & \dots & X_n^p \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}.$$

La première colonne est constituée uniquement de 1. Elle permet d'intégrer le terme constant β_0 dans l'écriture matricielle, ce qui évite de traiter l'intercepte séparément des autres coefficients.

On introduit également le vecteur des paramètres inconnus

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1},$$

ainsi que le vecteur des observations

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{et} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Dans ce cadre, le modèle de régression linéaire multiple s'écrit de manière compacte sous la forme

$$Y = \mathbb{X}\boldsymbol{\beta} + \varepsilon.$$

Cette écriture met clairement en évidence la nature linéaire du modèle et permet d'unifier l'étude théorique du cas simple et du cas multiple.

6.4.4 Estimation par la méthode des moindres carrés

Comme dans le cas univarié, les coefficients du modèle sont estimés par la méthode des moindres carrés. L'idée consiste à choisir le vecteur $\boldsymbol{\beta}$ qui minimise la somme des carrés des résidus, c'est-à-dire la norme euclidienne du vecteur des erreurs.

On cherche donc à résoudre le problème d'optimisation :

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \|Y - \mathbb{X}\boldsymbol{\beta}\|_2^2.$$

Ce critère généralise naturellement celui introduit dans le cas de la régression linéaire simple. Il s'agit toujours de minimiser une fonction quadratique strictement convexe, sous réserve que la matrice $\mathbb{X}$ contienne suffisamment d'information.

Théorème 6.4.1. *Si la matrice $\mathbb{X}^\top\mathbb{X}$ est inversible, alors le problème des moindres carrés admet une solution unique, donnée par*

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y.$$

Preuve. On considère la fonction objectif des moindres carrés

$$J(\boldsymbol{\beta}) = \|Y - \mathbb{X}\boldsymbol{\beta}\|_2^2 = (Y - \mathbb{X}\boldsymbol{\beta})^\top(Y - \mathbb{X}\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbb{R}^{p+1}.$$

En développant, on obtient

$$J(\boldsymbol{\beta}) = Y^\top Y - 2\,\boldsymbol{\beta}^\top\mathbb{X}^\top Y + \boldsymbol{\beta}^\top\mathbb{X}^\top\mathbb{X}\,\boldsymbol{\beta}.$$

Cette fonction est de classe $\mathcal{C}^1$ (c'est un polynôme) et son gradient vaut

$$\nabla J(\boldsymbol{\beta}) = -2\,\mathbb{X}^\top Y + 2\,\mathbb{X}^\top \mathbb{X}\,\boldsymbol{\beta}.$$

Un point minimiseur $\hat{\boldsymbol{\beta}}$ doit vérifier la condition d'optimalité du premier ordre

$$\nabla J(\hat{\boldsymbol{\beta}}) = 0 \quad \Longleftrightarrow \quad \mathbb{X}^\top \mathbb{X}\,\hat{\boldsymbol{\beta}} = \mathbb{X}^\top Y.$$

Ce système est appelé *équation normale*. Si $\mathbb{X}^\top\mathbb{X}$ est inversible, il admet une unique solution donnée par

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y.$$

Il reste à justifier qu'il s'agit bien d'un *minimum* (et non d'un maximum ou d'un point selle). La hessienne de J est constante et vaut

$$\nabla^2 J(\boldsymbol{\beta}) = 2\,\mathbb{X}^\top\mathbb{X}.$$

Sous l'hypothèse d'inversibilité, $\mathbb{X}^\top\mathbb{X}$ est définie positive, donc la hessienne est définie positive : J est alors strictement convexe sur $\mathbb{R}^{p+1}$. Par strict convexité, le point critique est nécessairement l'unique minimiseur global de J. □

Commentaires et interprétation. La matrice $\mathbb{X}^\top\mathbb{X}$ joue un rôle central dans la régression linéaire multiple. Elle est appelée *matrice de Gram* associée aux colonnes de la matrice de design $\mathbb{X}$ et regroupe l'ensemble des produits scalaires entre les vecteurs colonnes de $\mathbb{X}$, c'est-à-dire entre les différentes variables explicatives, y compris la colonne de 1 correspondant à l'intercept. En notant $\mathbb{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p)$, où $\mathbf{x}_0$ désigne la colonne de 1 et $\mathbf{x}_j$ la colonne associée à la variable X^j, on a

$$(\mathbb{X}^\top\mathbb{X})_{jk} = \langle \mathbf{x}_j, \mathbf{x}_k \rangle,$$

où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire usuel dans $\mathbb{R}^n$.

D'un point de vue géométrique, cette matrice mesure les relations de dépendance linéaire entre les covariables : lorsque deux colonnes sont fortement corrélées, leur produit scalaire est élevé, ce qui traduit une redondance d'information. L'hypothèse d'inversibilité de $\mathbb{X}^\top\mathbb{X}$ équivaut à l'indépendance linéaire des colonnes de $\mathbb{X}$, c'est-à-dire au fait qu'aucune variable explicative ne peut être exprimée comme une combinaison linéaire exacte des autres. Cette condition est essentielle pour garantir l'identifiabilité du modèle et l'unicité de l'estimateur des moindres carrés.

Sur le plan statistique, la non-inversibilité de $\mathbb{X}^\top\mathbb{X}$ correspond à une situation de *multicolinéarité parfaite*, dans laquelle il devient impossible de distinguer l'effet propre de certaines variables sur la variable réponse. Enfin, la formule

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top Y$$

met en évidence la nature géométrique de l'estimation : elle peut être interprétée comme la projection orthogonale du vecteur des observations Y sur l'espace vectoriel engendré par les colonnes de $\mathbb{X}$, suivie de la résolution d'un système linéaire. Cette expression constitue le cœur de la régression linéaire multiple et servira de point de départ à de nombreuses extensions du modèle linéaire, notamment les méthodes de régularisation étudiées ultérieurement.

6.5 Qualité du modèle

Ajuster une régression linéaire (simple ou multiple) revient à construire une fonction de prédiction

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i^1 + \cdots + \hat{\beta}_p X_i^p$$

qui approxime au mieux les valeurs observées Y_i. Une question essentielle est alors la suivante : *dans quelle mesure le modèle explique-t-il réellement les données ?* Pour répondre à cette question, on introduit des grandeurs qui mesurent la variabilité de Y et la part de cette variabilité capturée (ou non) par le modèle.

6.5.1 Sommes des carrés et décomposition de la variabilité

On note $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ la moyenne empirique de la variable réponse. Comme on l'a déjà souligné, $\bar{Y}$ représente une valeur centrale : si l'on devait prédire Y sans utiliser aucune covariable, la prédiction la plus naturelle (au sens des moindres carrés) serait précisément la constante $\bar{Y}$.

Somme totale des carrés.

Définition 6.5.1 (Somme totale des carrés)**.** *On appelle* somme totale des carrés *(en anglais* Total Sum of Squares*) la quantité*

$$\mathrm{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

où

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

désigne la moyenne empirique des observations.

La somme totale des carrés mesure la variabilité globale de la variable réponse autour de sa moyenne. Elle constitue une référence absolue : elle correspond à l'erreur que l'on commettrait si l'on cherchait à prédire toutes les observations par une constante unique, égale à $\bar{Y}$. Une valeur élevée de SST traduit une forte dispersion des données, tandis qu'une valeur faible indique que les observations sont proches de leur moyenne.

Somme des carrés résiduels.

Définition 6.5.2 (Somme des carrés résiduels)**.** *Après ajustement du modèle, on définit les résidus par*

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i,$$

où $\hat{Y}_i$ désigne la valeur prédite par le modèle pour l'observation i. La somme des carrés résiduels *(en anglais* Residual Sum of Squares*) est alors définie par*

$$\mathrm{SSR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}.$$

La quantité SSR mesure l'erreur globale commise par le modèle sur les données observées. Elle évalue la part de la variabilité des données qui n'est pas expliquée par le modèle. Plus SSR est petite, plus l'ajustement est précis. Dans le cadre de la méthode des moindres carrés, l'estimation des paramètres vise précisément à minimiser cette somme.

Somme des carrés expliqués.

Définition 6.5.3 (Somme des carrés expliqués)**.** *On appelle* somme des carrés expliqués *(en anglais* Explained Sum of Squares*) la quantité*

$$\mathrm{SSE} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2,$$

où $\hat{Y}_i$ sont les valeurs prédites par le modèle et $\bar{Y}$ la moyenne empirique des observations.

La somme des carrés expliqués mesure la variabilité des prédictions produites par le modèle autour de la moyenne. Elle représente la part de la variabilité totale des données que le modèle parvient à capturer à l'aide des variables explicatives. Plus SSE est grande, plus le modèle explique une part importante de la structure présente dans les données.

Proposition 6.5.1 (Décomposition de la variabilité)**.** *Dans le cadre des moindres carrés (avec intercept), on a la décomposition fondamentale*

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}.$$

Preuve. On écrit

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}).$$

En élevant au carré puis en sommant, on obtient

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}).$$

Il reste à montrer que le dernier terme est nul. Or, dans la régression aux moindres carrés avec intercept, le vecteur des résidus est orthogonal à l'espace engendré par les colonnes de $\mathbb{X}$, donc en particulier orthogonal au vecteur $(\hat{Y}_1 - \bar{Y}, \ldots, \hat{Y}_n - \bar{Y})$ qui appartient à cet espace. Ainsi

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0,$$

ce qui donne bien $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$. □

Remarque 5. Cette décomposition est l'analogue, en statistiques, d'un théorème de Pythagore : SST mesure l'énergie totale, SSR l'énergie résiduelle (non expliquée), et SSE l'énergie expliquée par le modèle.

Lien entre décomposition empirique et décomposition biais–variance. Il existe un lien conceptuel profond entre la décomposition empirique

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

et la décomposition théorique du risque quadratique, rappelée en (5.1). Ces deux écritures n'opèrent toutefois pas au même niveau : la première est définie sur un échantillon fini et observé, tandis que la seconde s'exprime en espérance, relativement à la distribution inconnue des données et au caractère aléatoire du processus d'apprentissage.

Cette correspondance peut être résumée par le tableau suivant :

Niveau empirique (échantillon)	**Niveau théorique (espérance)**
$\mathrm{SST} = \sum(Y_i - \bar{Y})^2$	$\mathrm{Var}[Y]$
$\mathrm{SSE} = \sum(\hat{Y}_i - \bar{Y})^2$	$\mathrm{Biais}(\widehat{f})^2$
$\mathrm{SSR} = \sum(Y_i - \hat{Y}_i)^2$	$\mathrm{Var}\left[\,\widehat{f}\,\right] + \mathrm{Var}[Y \mid X]$

Plus précisément, on peut montrer que, en moyenne sur les échantillons d'apprentissage,

$$\mathbb{E}[\mathrm{SSR}] \;\approx\; n\Big(\mathrm{Var}\big(\widehat{f}(X)\big) + \mathrm{Var}[Y \mid X]\Big),$$

tandis que la quantité SSE est associée à l'erreur systématique du modèle, c'est-à-dire au biais. Ainsi, la somme des carrés expliqués joue un rôle analogue au biais au carré, tandis que la somme des carrés résiduels agrège la variance du modèle et le bruit irréductible.

Lecture géométrique unificatrice. Cette analogie s'éclaire par une interprétation géométrique commune aux deux décompositions. Dans les deux cas, il s'agit d'une *projection orthogonale dans un espace de fonctions*. Empiriquement, la régression des moindres carrés correspond à la projection du vecteur des observations $(Y_1, \dots, Y_n)$ sur l'espace engendré par les prédictions possibles du modèle. La décomposition $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$ découle alors directement du théorème de Pythagore.

Du point de vue théorique, la décomposition biais–variance repose sur la même idée : la fonction de régression optimale $f^\star(x) = \mathbb{E}[Y \mid X = x]$ joue le rôle de projection de Y sur l'espace des fonctions de X, tandis que l'algorithme d'apprentissage fournit une approximation aléatoire de cette projection.

En synthèse, la décomposition empirique des sommes de carrés et la décomposition biais–variance du risque quadratique constituent deux manifestations d'un même principe fondamental. La première offre une lecture concrète et calculable sur les données observées, tandis que la seconde fournit une compréhension théorique et probabiliste des sources d'erreur. Leur mise en regard permet de relier directement les performances observées d'un modèle à ses propriétés statistiques de généralisation.

6.5.2 Coefficient de détermination R^2

Définition.

Définition 6.5.4 (Coefficient de détermination)**.** *On appelle* coefficient de détermination *le nombre*

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

Cette formule s'interprète immédiatement : R^2 compare l'erreur du modèle (SSR) à l'erreur du modèle le plus naïf possible (prédire $\bar{Y}$ pour tout le monde), qui correspond à SST. Ainsi :

- si SSR est très petite devant SST, alors le modèle explique bien les données et R^2 est proche de 1 ;
- si SSR est proche de SST, alors le modèle n'apporte presque rien par rapport à la simple moyenne et R^2 est proche de 0.

Propriété 6.5.1. *Dans une régression linéaire estimée par moindres carrés avec intercept, on a toujours*

$$0 \leq R^2 \leq 1.$$

Exemples d'interprétation.

Exercice 6.5.1 (Modèle peu informatif). Supposons que $R^2 = 0.05$. Cela signifie que le modèle n'explique qu'environ 5% de la variabilité de Y. Autrement dit, même en connaissant les covariables, l'incertitude sur Y reste très grande : le modèle est probablement trop simple, les covariables sont peu pertinentes, ou bien la relation n'est pas linéaire.

Exercice 6.5.2 (Modèle très explicatif). Si $R^2 = 0.90$, alors le modèle explique environ 90% de la variabilité de Y. Cela indique une forte capacité descriptive : la prédiction $\hat{Y}$ suit étroitement les données, et les résidus sont globalement faibles.

Remarque 6. Un R^2 élevé ne garantit pas à lui seul que le modèle est "bon" au sens scientifique : il peut résulter d'un sur-ajustement (trop de variables), d'une relation accidentelle sur l'échantillon, ou encore d'un phénomène déterministe dans les données. L'analyse des résidus et des hypothèses du modèle reste indispensable.

6.5.3 Le R^2 ajusté

Un point important mérite d'être compris : lorsque l'on ajoute des variables explicatives, la somme des carrés résiduels SSR ne peut qu'*diminuer* (ou rester constante), car on augmente la flexibilité du modèle. Par conséquent, R^2 ne peut qu'augmenter (ou rester constant). Cela rend R^2 peu fiable pour comparer des modèles de tailles différentes, car il favorise mécaniquement les modèles plus complexes.

Pour corriger cet effet, on introduit une version pénalisée : le R^2 ajusté.

Définition 6.5.5 (R^2 ajusté)**.** *Dans un modèle de régression linéaire multiple comportant p variables explicatives (et donc $p+1$ paramètres en comptant l'intercept), on définit*

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

Interprétation. Le facteur $\frac{n-1}{n-p-1}$ est supérieur à 1 et augmente avec p. Ainsi, à R^2 fixé, plus le modèle contient de variables, plus la pénalisation est forte. Le R^2_{adj} ne récompense donc l'ajout d'une variable que si cette variable améliore suffisamment l'ajustement pour compenser l'augmentation de complexité.

Exemple 6.5.1 (Comparaison de deux modèles)**.** *Supposons que l'on compare deux modèles sur les mêmes données (même n) :*

- *Modèle A : $p = 2$ variables explicatives, $R^2 = 0.72$.*
- *Modèle B : $p = 6$ variables explicatives, $R^2 = 0.74$.*

Même si R^2 augmente légèrement, le R^2_{adj} du modèle B peut être inférieur à celui du modèle A si les variables supplémentaires n'apportent qu'un gain marginal. Dans ce cas, le modèle A est à préférer : il est plus simple et presque aussi explicatif.

Remarque 7. Le R^2_{adj} constitue un premier outil de comparaison de modèles, mais il ne remplace pas une analyse plus complète (validation croisée, critères AIC/BIC, diagnostic des résidus, etc.). Il a néanmoins l'avantage d'être simple et directement interprétable dans le cadre des régressions linéaires.

Chapitre 7

La régression logistique

La régression logistique constitue une extension naturelle de la régression linéaire lorsque la variable réponse n'est plus quantitative, mais *catégorielle*. Dans ce chapitre, on se concentre sur le cas fondamental où la variable à prédire est binaire, prenant ses valeurs dans $\{0, 1\}$. Ce cadre est omniprésent en pratique : diagnostic médical (malade / non malade), détection de fraude (fraude / non fraude), réussite à un examen (oui / non), etc.

L'objectif est de modéliser la probabilité d'appartenance à l'une des deux classes à partir de variables explicatives quantitatives.

7.1 Contexte et motivation

On considère un échantillon de données

$$\big\{(X_1^1, \dots, X_1^p, Y_1), \dots, (X_n^1, \dots, X_n^p, Y_n)\big\},$$

où $(X_i^1, \dots, X_i^p) \in \mathbb{R}^p$ représentent les variables explicatives et $Y_i \in \{0, 1\}$ la variable réponse.

Puisque Y est binaire, elle suit naturellement une loi de Bernoulli. Plus précisément, conditionnellement aux variables explicatives, on suppose que

$$Y \mid (X^1, \dots, X^p) \sim \mathcal{B}(q),$$

où le paramètre q satisfait

$$q \in [0, 1], \qquad \mathbb{E}[Y \mid X^1, \dots, X^p] = q.$$

Ainsi, le problème fondamental de la régression logistique consiste à estimer la probabilité conditionnelle

$$q(x) = \mathbb{P}(Y = 1 \mid X = x),$$

en fonction des covariables.

7.2 Limites de la régression linéaire

Une première idée naturelle serait de modéliser directement $q(x)$ par une fonction linéaire des covariables :

$$q(x) \approx \beta_0 + \sum_{j=1}^{p} \beta_j x^j.$$

Cependant, cette approche est inadéquate : une fonction linéaire à valeurs réelles peut produire des valeurs négatives ou supérieures à 1, ce qui est incompatible avec une interprétation probabiliste.

Il est donc nécessaire de contraindre la prédiction à appartenir à l'intervalle $[0, 1]$. Pour cela, on introduit une fonction de lien

$$g : \mathbb{R} \longrightarrow [0, 1],$$

qui transforme une combinaison linéaire des covariables en une probabilité valide.

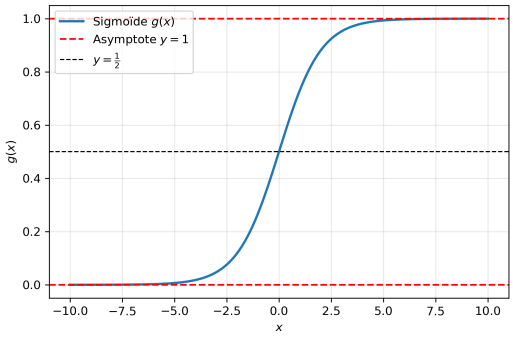
7.3 La fonction sigmoïde

Parmi les fonctions de lien possibles, la plus utilisée est la *fonction logistique*, ou sigmoïde, définie par

$$g(x) = \frac{\mathrm{e}^x}{1 + \mathrm{e}^x}.$$

Cette fonction a le bon goût de posséder plusieurs propriétés essentielles :

- elle est strictement croissante et bijective de $\mathbb{R}$ vers $]0, 1[$;
- elle est symétrique autour du point $x = 0$;
- elle vérifie $g(0) = \frac{1}{2}$, ce qui correspond à une situation d'incertitude maximale ;
- elle possède un point d'inflexion en $x = 0$.



7.4 Le modèle de régression logistique

Définition 7.4.1 (Régression logistique)**.** *Le modèle de régression logistique est défini par*

$$\mathbb{P}(Y = 1 \mid X = x) = g\left(\beta_0 + \sum_{j=1}^{p} \beta_j x^j\right) = \frac{\mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x^j}}{1 + \mathrm{e}^{\beta_0 + \sum_{j=1}^{p} \beta_j x^j}}.$$

Ainsi, le modèle suppose que le *log-rapport de chances* (log-odds)

$$\log\left(\frac{q(x)}{1 - q(x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j$$

est une fonction linéaire des covariables. Cette écriture explique le nom de *régression logistique*.

7.5 Fonction de perte et estimation

7.5.1 Perte d'entropie croisée

Dans ce cadre probabiliste, la perte quadratique n'est plus adaptée. La fonction de perte naturelle est la *perte d'entropie croisée*, définie pour une observation $(y, \hat{q})$ par

$$\ell(y, \hat{q}) = -\big(y \log \hat{q} + (1 - y) \log(1 - \hat{q})\big).$$

Définition 7.5.1 (Perte empirique logistique)**.** *La perte empirique associée au modèle est*

$$\mathcal{L}_n(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \Big(Y_i \log q(X_i) + (1 - Y_i) \log(1 - q(X_i))\Big).$$

Cette fonction pénalise fortement les prédictions très confiantes mais incorrectes, ce qui est particulièrement pertinent en classification.

7.5.2 Lien avec le maximum de vraisemblance

Un point fondamental est que la minimisation de la perte d'entropie croisée est strictement équivalente à l'estimation des paramètres par *maximum de vraisemblance*. En effet, sous l'hypothèse

$$Y_i \mid X_i \sim \mathcal{B}(q(X_i)),$$

la vraisemblance du modèle est

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} q(X_i)^{Y_i} \big(1 - q(X_i)\big)^{1-Y_i},$$

et la maximisation du logarithme de cette quantité conduit exactement à la minimisation de $\mathcal{L}_n(\boldsymbol{\beta})$.

7.6 Optimisation numérique

Contrairement à la régression linéaire, il n'existe pas de formule explicite pour les estimateurs $\hat{\boldsymbol{\beta}}$. La fonction de perte est convexe, mais non quadratique. L'estimation repose donc sur des méthodes numériques, principalement la descente de gradient.

À chaque itération t, les paramètres sont mis à jour selon

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta \nabla \mathcal{L}_n\big(\boldsymbol{\beta}^{(t)}\big),$$

où $\eta > 0$ est le pas d'apprentissage.

7.7 Interprétation et décision

Le modèle fournit une probabilité $\hat{q}(x)$. Pour produire une prédiction binaire, on introduit un seuil, généralement 0.5 :

$$\hat{Y} = \begin{cases} 1 & \text{si } \hat{q}(x) \geq \frac{1}{2}, \\ 0 & \text{sinon.} \end{cases}$$

La frontière de décision correspond alors à l'hyperplan

$$\beta_0 + \sum_{j=1}^{p} \beta_j x^j = 0,$$

ce qui montre que la régression logistique induit une *séparation linéaire* de l'espace des covariables.

7.8 Avantages et limites

Avantages. La régression logistique présente de nombreux atouts :

- elle fournit des probabilités interprétables ;
- elle repose sur un cadre probabiliste rigoureux ;
- elle est simple à mettre en œuvre et à interpréter ;
- elle constitue un modèle de référence en classification binaire.

Limites. Sa principale limitation réside dans l'hypothèse de séparation linéaire des classes. Lorsque la frontière réelle entre les classes est fortement non linéaire, le modèle peut présenter un biais important. De plus, comme tout modèle paramétrique, sa performance dépend fortement de la qualité et de la pertinence des variables explicatives.

Ces limitations motivent l'étude de modèles plus flexibles, tels que les méthodes à noyaux, les arbres de décision ou les réseaux de neurones, qui généralisent les principes introduits ici.

Chapitre 8

Arbre de décision

8.1 Contexte et motivations

Les arbres de décision constituent une famille de modèles prédictifs non paramétriques largement utilisés en statistique et en apprentissage automatique. Leur principe repose sur une idée intuitive : *partitionner l'espace des variables explicatives en régions homogènes*, puis associer à chaque région une prédiction simple.

Contrairement aux modèles linéaires, les arbres de décision ne supposent aucune relation fonctionnelle globale entre la variable réponse et les covariables. Ils construisent une approximation locale, par découpage successif de l'espace des données.

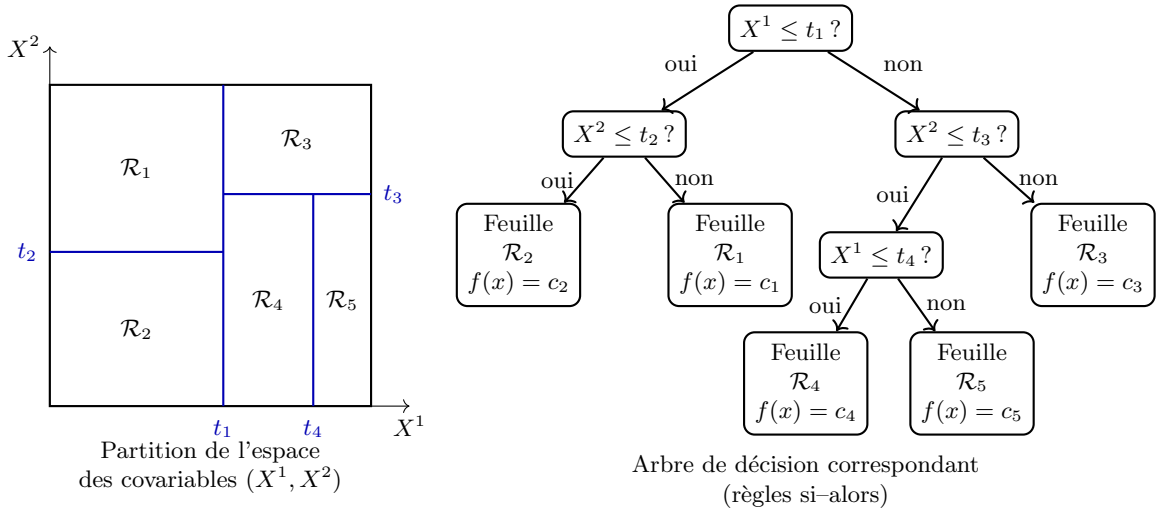


FIGURE 8.1 – Illustration d'un arbre de décision

Arbre de décision en dimension 2. Soit un échantillon de données

$$\big\{(X_1^1, X_1^2, Y_1),\ \ldots,\ (X_n^1, X_n^2, Y_n)\big\},$$

où $(X_i^1, X_i^2) \in \mathbb{R}^2$ désignent les variables explicatives et $Y_i \in \mathbb{R}$ la variable réponse. On cherche à modéliser la relation entre Y et (X^1, X^2) à l'aide d'un arbre de décision de régression.

i. La partie gauche de la figure 8.1 représente une partition de l'espace des covariables (X^1, X^2) obtenue après apprentissage. Chaque région $\mathcal{R}_m$ correspond à une feuille de l'arbre et regroupe des observations considérées comme homogènes du point de vue de la variable réponse. Dans chaque région, la prédiction repose sur l'agrégation locale des valeurs observées de Y (typiquement par moyenne empirique).

ii. La partie droite de la figure fournit la représentation arborescente équivalente de cette partition. Elle explicite la suite de règles de décision de type *si–alors* permettant d'assigner toute nouvelle observation $x = (x^1, x^2)$ à une région donnée. Cette structure met en évidence le caractère hiérarchique et récursif de l'apprentissage par arbres de décision.

Définition 8.1.1 (Forme fonctionnelle d'un arbre de décision)**.** *D'un point de vue formel, le modèle appris par un arbre de décision peut s'écrire sous la forme*

$$f(x) = \sum_{m=1}^{M} c_m \, \mathbb{1}_{\{x \in \mathcal{R}_m\}},$$

où $\{\mathcal{R}_1, \ldots, \mathcal{R}_M\}$ est la partition de l'espace des covariables induite par l'arbre, et où c_m désigne la valeur prédite sur la région (ou feuille) $\mathcal{R}_m$.

Les constantes c_m ne sont pas choisies arbitrairement : elles sont définies comme les solutions de problèmes de minimisation locale de la fonction de perte restreinte à chaque région. Plus précisément, pour une région $\mathcal{R}_m$, on définit

$$c_m = \operatorname*{argmin}_{c \in \mathcal{Y}} \sum_{i:\, X_i \in \mathcal{R}_m} \ell(Y_i, c),$$

où ℓ est la fonction de perte associée au problème considéré.

Remarque 8 (Interprétation de c_m selon la perte). Le choix explicite de c_m dépend de la nature du problème et de la fonction de perte utilisée :

— dans le cas de la régression avec perte quadratique $\ell(y, c) = (y - c)^2$, on obtient

$$c_m = \frac{1}{|\mathcal{R}_m|} \sum_{i:\, X_i \in \mathcal{R}_m} Y_i,$$

c'est-à-dire la moyenne empirique des valeurs observées dans la région $\mathcal{R}_m$;

— dans le cas de la classification avec perte 0–1, la valeur optimale c_m est la classe majoritaire dans la région ;

— pour des critères d'impureté tels que l'entropie ou l'indice de Gini, la prédiction associée à la feuille correspond également à la classe maximisant la probabilité empirique conditionnelle.

Ainsi, un arbre de décision peut être vu comme une approximation locale de la fonction de régression $\mathbb{E}[Y \mid X = x]$ (en régression) ou de la règle de Bayes (en classification).

8.2 Cadre statistique

On se place dans le cadre classique de l'apprentissage supervisé. On dispose d'un échantillon de données

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

où :

— $X_i = (X_i^1, \ldots, X_i^p) \in \mathbb{R}^p$ est le vecteur de covariables,

— $Y_i \in \mathbb{R}$ (régression) ou $Y_i \in \{1, \ldots, K\}$ (classification),

que l'on modélise comme des réalisations i.i.d. de variables aléatoires (X, Y).

L'objectif est de construire une fonction de prédiction

$$f : \mathbb{R}^p \to \mathcal{Y},$$

qui approxime au mieux la relation entre X et Y.

Principe général d'un arbre de décision. Un arbre de décision est une fonction f définie par une suite de règles de type *si–alors*, organisées sous forme arborescente.

Partition de l'espace. L'espace des covariables $\mathbb{R}^p$ est découpé en régions disjointes

$$\mathbb{R}^p = \bigsqcup_{m=1}^{M} \mathcal{R}_m,$$

appelées *feuilles* de l'arbre. Chaque région $\mathcal{R}_m$ est définie par une suite de contraintes simples du type

$$X^j \leq t \quad \text{ou} \quad X^j > t,$$

où $j \in \{1, \ldots, p\}$ et $t \in \mathbb{R}$.

Prédiction locale. À chaque région $\mathcal{R}_m$ est associée une prédiction constante :

$$f(x) = \sum_{m=1}^{M} c_m \, \mathbb{1}_{\{x \in \mathcal{R}_m\}},$$

Ainsi, un arbre de décision est une fonction *par morceaux constante*.

8.3 Apprentissage d'un arbre de décision

8.3.1 Fonction de perte associée

L'apprentissage d'un arbre de décision consiste à construire une fonction de prédiction f qui approxime au mieux la relation entre les variables explicatives X et la variable réponse Y. Cette approximation est formulée de manière quantitative à l'aide d'une *fonction de perte*, qui mesure l'écart entre les valeurs prédites par le modèle et les observations réelles.

Dans le cadre des arbres de décision, l'apprentissage repose sur la minimisation d'une *perte empirique*, calculée à partir des données disponibles. Le choix de la fonction de perte dépend de la nature de la variable réponse : régression ou classification.

Cas de la régression. Lorsque la variable réponse est réelle ($Y \in \mathbb{R}$), on adopte classiquement la perte quadratique

$$\ell(y, \hat{y}) = (y - \hat{y})^2.$$

Cette perte pénalise fortement les grandes erreurs et conduit naturellement à des prédictions de type moyenne. Étant donné une fonction de prédiction f, la perte empirique associée s'écrit

$$\mathcal{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2.$$

Dans le cas des arbres de régression, on rappelle que la fonction f est constante sur chaque région $\mathcal{R}_m$. Minimiser la perte quadratique sur une région donnée conduit alors à choisir, comme valeur prédite c_m, la moyenne empirique des Y_i associés à cette région. Cette propriété explique le rôle central de la moyenne dans les feuilles d'un arbre de régression.

Cas de la classification. Lorsque la variable réponse est discrète, par exemple $Y \in \{1, \ldots, K\}$, l'objectif n'est plus d'approcher une valeur numérique, mais d'assigner chaque observation à une classe. Une perte naturelle est alors la perte 0–1, définie par

$$\ell(y, \hat{y}) = \mathbb{1}_{\{y \neq \hat{y}\}}.$$

Cependant, cette perte est peu adaptée à l'apprentissage des arbres, car elle ne fournit pas une mesure fine de la qualité d'une partition intermédiaire. On lui préfère donc des *critères d'impureté*, qui quantifient le degré d'hétérogénéité des classes au sein d'une région.

Entropie. Soit $\mathcal{R}$ une région contenant un sous-ensemble des données. On note

$$p_k = \mathbb{P}(Y = k \mid X \in \mathcal{R})$$

la proportion d'observations de la classe k dans cette région. L'entropie de $\mathcal{R}$ est définie par

$$H(\mathcal{R}) = -\sum_{k=1}^{K} p_k \log(p_k),$$

avec la convention $0 \log 0 = 0$.

L'entropie mesure l'incertitude associée à la distribution des classes dans une région. Elle est minimale (égale à 0) lorsque la région est *pure*, c'est-à-dire lorsque toutes les observations appartiennent à une même classe. À l'inverse, elle est maximale lorsque les classes sont réparties de manière uniforme. Ainsi, une faible entropie indique une région bien séparée du point de vue de la classification.

Indice de Gini. Un autre critère largement utilisé est l'indice de Gini, défini par

$$G(\mathcal{R}) = 1 - \sum_{k=1}^{K} p_k^2.$$

L'indice de Gini admet une interprétation probabiliste simple : il correspond à la probabilité qu'en tirant aléatoirement deux observations dans la région $\mathcal{R}$, celles-ci appartiennent à des classes différentes.

Comme l'entropie, l'indice de Gini est nul lorsque la région est pure et augmente avec l'hétérogénéité des classes. Il présente l'avantage d'être plus simple à calculer et conduit souvent à des partitions très proches de celles obtenues par l'entropie.

Principe de minimisation. Lors de la construction de l'arbre, chaque coupure candidate est évaluée en comparant l'impureté de la région avant et après la séparation. Le critère retenu (perte quadratique, entropie ou indice de Gini) est choisi de manière à *réduire au maximum* l'impureté globale, pondérée par la taille des régions. Ce processus glouton permet de construire récursivement un arbre qui rend les régions de plus en plus homogènes.

Remarque 9. Bien que les critères d'entropie et de Gini soient différents dans leur expression, ils poursuivent le même objectif : produire des régions aussi pures que possible. En pratique, le choix de l'un ou de l'autre a souvent peu d'impact sur la structure finale de l'arbre.

8.3.2 Processus d'apprentissage

L'apprentissage d'un arbre de décision consiste à construire, à partir des données, une partition de l'espace des covariables qui rende la variable réponse aussi homogène que possible à l'intérieur de chaque région. Concrètement, il s'agit de déterminer :

— les variables explicatives utilisées pour les coupures,

— les seuils de séparation associés à ces variables,

— la structure globale de l'arbre, notamment le nombre et la disposition des feuilles,

de manière à minimiser une fonction de perte empirique adaptée au problème considéré (régression ou classification).

Découpage récursif de l'espace. Les arbres de décision sont construits selon un principe de *découpage récursif* de l'espace des covariables. L'algorithme d'apprentissage est de nature *gloutonne* : à chaque étape, on choisit la meilleure coupure locale sans revenir sur les décisions précédentes.

Plus précisément, le processus s'effectue comme suit :

1. on initialise l'arbre avec une seule région $\mathcal{R}_0$ contenant l'ensemble des observations ;
2. à partir d'une région $\mathcal{R}$, on considère toutes les coupures possibles de la forme

$$X^j \leq t \quad \text{ou} \quad X^j > t,$$

où $j \in \{1, \ldots, p\}$ et $t \in \mathbb{R}$;

3. chaque coupure scinde la région $\mathcal{R}$ en deux sous-régions, notées $\mathcal{R}_{\text{gauche}}$ et $\mathcal{R}_{\text{droite}}$;
4. on retient la coupure qui conduit à la plus forte diminution de la perte empirique.

Ce procédé est appliqué récursivement à chaque nouvelle région, jusqu'à ce qu'un critère d'arrêt soit atteint (profondeur maximale, taille minimale des feuilles, ou absence de gain significatif).

Critère optimal local. Soit $\mathcal{R}$ une région intermédiaire contenant un sous-échantillon des données. Pour une coupure candidate (j, t), on évalue la qualité de la séparation en comparant la perte avant et après la coupure. D'un point de vue général, on cherche à minimiser

$$\mathcal{L}(\mathcal{R}_{\text{gauche}}) + \mathcal{L}(\mathcal{R}_{\text{droite}}),$$

où $\mathcal{L}(\cdot)$ désigne la perte empirique restreinte à une région.

Dans le cas de la régression, cette perte correspond à la somme des carrés des résidus. En classification, la perte est remplacée par une mesure d'impureté, telle que l'entropie ou l'indice de Gini.

Apprentissage en classification avec perte entropique. Dans un problème de classification, on suppose que la variable réponse prend des valeurs dans un ensemble fini $\{1, \dots, K\}$. Soit $\mathcal{R}$ une région donnée, et

$$p_k(\mathcal{R}) = \mathbb{P}(Y = k \mid X \in \mathcal{R})$$

la proportion empirique d'observations de la classe k dans cette région. L'entropie associée à $\mathcal{R}$ est

$$H(\mathcal{R}) = -\sum_{k=1}^{K} p_k(\mathcal{R}) \log p_k(\mathcal{R}).$$

Lorsqu'une région $\mathcal{R}$ est scindée en deux sous-régions $\mathcal{R}_{\text{gauche}}$ et $\mathcal{R}_{\text{droite}}$, la qualité de la coupure est évaluée par l'entropie pondérée

$$\frac{|\mathcal{R}_{\text{gauche}}|}{|\mathcal{R}|} H(\mathcal{R}_{\text{gauche}}) + \frac{|\mathcal{R}_{\text{droite}}|}{|\mathcal{R}|} H(\mathcal{R}_{\text{droite}}).$$

La coupure optimale est celle qui minimise cette quantité, ou de manière équivalente, celle qui maximise la *réduction d'entropie*, aussi appelée *gain d'information*.

Choix de la prédiction dans une feuille. Une fois une région $\mathcal{R}_m$ devenue une feuille de l'arbre, on lui associe une prédiction constante. Dans le cas de la régression, la minimisation de la perte quadratique conduit à choisir

$$c_m = \frac{1}{|\mathcal{R}_m|} \sum_{i : X_i \in \mathcal{R}_m} Y_i,$$

c'est-à-dire la moyenne empirique des valeurs observées dans la région. Chaque feuille fournit ainsi un estimateur local de l'espérance conditionnelle $\mathbb{E}[Y \mid X \in \mathcal{R}_m]$.

En classification, la prédiction associée à une feuille est la classe majoritaire dans la région, ce qui correspond au choix minimisant la perte 0–1. Dans les deux cas, la structure de l'arbre permet d'approximer la relation entre X et Y par une succession de décisions simples, conduisant à une prédiction locale.

8.3.3 Arbres de décision avec variables catégorielles

Jusqu'à présent, nous avons principalement considéré le cas où les variables explicatives sont quantitatives. Les arbres de décision se prêtent toutefois particulièrement bien au traitement de *variables catégorielles*, tant du côté des covariables que de la variable réponse. Cette capacité constitue l'un de leurs principaux atouts par rapport aux modèles paramétriques classiques.

On propose ci-dessous un exemple précis pour deux variables explicatives et une réponse, qui peut se généraliser très facilement à autant de variables explicatives que voulues.

Cadre du problème. On considère un échantillon de la forme

$$\left\{(X_i^1, X_i^2, Y_i)\right\}_{i=1}^n,$$

où :

— X^1 et X^2 sont des variables catégorielles finies (par exemple : couleur, type, catégorie socio-professionnelle, etc.) ;

— Y est une variable catégorielle, que l'on supposera ici prendre ses valeurs dans un ensemble fini $\mathcal{Y} = \{1, \ldots, C\}$.

L'objectif est de construire un modèle prédictif permettant d'associer à toute observation (X^1, X^2) une classe prédite pour Y.

Principe des coupures pour variables catégorielles. Contrairement au cas quantitatif, une variable catégorielle ne se prête pas naturellement à une coupure de type seuil. Les règles de décision prennent alors la forme de tests d'appartenance à un sous-ensemble de modalités. Par exemple :

$$X^1 \in \{\mathrm{A}, \mathrm{C}\} \quad \text{vs} \quad X^1 \in \{\mathrm{B}\}.$$

Chaque nœud interne de l'arbre correspond ainsi à une partition de l'ensemble des modalités d'une variable explicative, induisant une séparation de l'échantillon en sous-groupes plus homogènes du point de vue de la variable réponse.

Fonction de perte : l'entropie. Dans le cadre de la classification, l'apprentissage de l'arbre repose sur la minimisation d'une mesure d'impureté des régions. La plus couramment utilisée est l'*entropie de Shannon*.

Définition 8.3.1 (Entropie d'une région)**.** *Soit $\mathcal{R}$ une région contenant un sous-échantillon. On note*

$$p_k = \mathbb{P}(Y = k \mid X \in \mathcal{R}), \qquad k = 1, \ldots, C,$$

les proportions empiriques des classes dans cette région. L'entropie de $\mathcal{R}$ est définie par

$$H(\mathcal{R}) = -\sum_{k=1}^{C} p_k \log p_k.$$

L'entropie mesure l'incertitude associée à la distribution des classes dans une région :

— elle est minimale (nulle) lorsque la région est *pure*, c'est-à-dire composée d'une seule classe ;

— elle est maximale lorsque les classes sont réparties de manière uniforme.

Critère de coupure optimale. Lorsqu'une région $\mathcal{R}$ est scindée en deux sous-régions $\mathcal{R}_{\text{gauche}}$ et $\mathcal{R}_{\text{droite}}$, la qualité de la coupure est évaluée par la diminution d'entropie, appelée *gain d'information* :

$$\text{Gain} = H(\mathcal{R}) - \frac{|\mathcal{R}_{\text{gauche}}|}{|\mathcal{R}|} H(\mathcal{R}_{\text{gauche}}) - \frac{|\mathcal{R}_{\text{droite}}|}{|\mathcal{R}|} H(\mathcal{R}_{\text{droite}}).$$

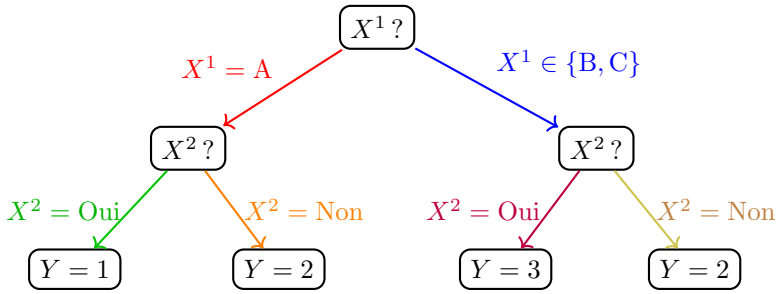
L'algorithme choisit, de manière gloutonne, la variable et la partition des modalités maximisant ce gain.

Prédiction dans une feuille. Une fois l'arbre construit, chaque feuille $\mathcal{R}_m$ correspond à une région terminale dans laquelle la prédiction est constante. Dans le cas de la classification, la valeur prédite est la classe majoritaire :

$$c_m = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y = k \mid X \in \mathcal{R}_m).$$

Ainsi, l'arbre fournit une approximation par morceaux de la probabilité conditionnelle $\mathbb{P}(Y \mid X)$.

Illustration : arbre de décision catégoriel.



Arbre de décision avec variables catégorielles
(règles conditionnelles sur les modalités)

Cet exemple illustre la capacité des arbres de décision à manipuler directement des variables catégorielles, sans nécessiter de transformation préalable (encodage numérique, variables indicatrices, etc.). Les règles de décision sont explicites, interprétables et proches d'un raisonnement logique humain. En contrepartie, lorsque le nombre de modalités est élevé, le nombre de partitions possibles croît rapidement, ce qui renforce la nécessité de mécanismes de régularisation pour éviter le sur-apprentissage.

8.3.4 Sur-apprentissage et régularisation

Les arbres de décision présentent une très grande capacité d'adaptation aux données. En particulier, lorsqu'un arbre est autorisé à croître sans contrainte, il peut continuer à scinder l'espace des covariables jusqu'à isoler presque chaque observation dans une feuille distincte. Dans une telle situation, le modèle reproduit quasi parfaitement les données d'apprentissage, mais devient extrêmement sensible aux fluctuations de l'échantillon : on parle alors de *sur-apprentissage* (*overfitting*).

Ce phénomène s'interprète naturellement à la lumière du compromis biais–variance. Un arbre très profond possède un biais faible, car il est capable d'approximer des relations complexes et non linéaires entre les variables. En revanche, cette flexibilité s'accompagne d'une variance élevée : de légères modifications des données peuvent conduire à une structure d'arbre très différente, et donc à des prédictions instables sur de nouvelles observations.

Principe de la régularisation. Afin de contrôler le sur-apprentissage et d'améliorer les performances de généralisation, on introduit des mécanismes de *régularisation*. Leur objectif est de limiter la complexité effective de l'arbre, en empêchant des découpages trop fins de l'espace des covariables. Parmi les stratégies les plus courantes, on peut citer :

- la limitation de la profondeur maximale de l'arbre, qui borne le nombre de décisions successives ;
- l'imposition d'un nombre minimal d'observations par feuille, afin d'éviter des régions trop petites et peu représentatives ;
- la pénalisation explicite de la complexité de l'arbre, par exemple en ajoutant un terme de coût lié au nombre de feuilles ou de nœuds internes.

Ces contraintes reviennent à accepter un biais légèrement plus élevé en échange d'une réduction significative de la variance, ce qui conduit généralement à une diminution de l'erreur de prédiction sur des données non observées.

Positionnement conceptuel. Les arbres de décision occupent ainsi une place particulière parmi les modèles statistiques et d'apprentissage automatique. Ils offrent :

- une grande flexibilité, leur permettant de capturer des interactions complexes entre variables ;
- une interprétabilité directe, grâce à leur structure hiérarchique et à leurs règles de décision explicites ;
- une absence d'hypothèses paramétriques fortes sur la forme de la relation entre X et Y.

En contrepartie, lorsqu'ils sont utilisés isolément, les arbres de décision souffrent d'une variance élevée, en particulier dans des contextes de données bruitées ou de taille modérée. Cette limitation motive le développement de méthodes d'agrégation, telles que les forêts aléatoires ou les méthodes de boosting, qui combinent plusieurs arbres afin de réduire la variance tout en conservant leur capacité de modélisation. Ces approches seront étudiées ultérieurement.

Chapitre 9

Les forêts aléatoires

Avant d'introduire formellement les forêts aléatoires, il est essentiel de comprendre la motivation statistique qui sous-tend leur construction. Cette motivation repose sur une idée simple mais fondamentale : *réduire la variance d'un modèle instable en combinant plusieurs modèles appris sur des échantillons légèrement différents*. Cette approche est connue sous le nom de *bagging* (*bootstrap aggregating*).

9.1 Motivation : le bagging

Les arbres de décision présentent de nombreux avantages. Ils sont interprétables (*white box*), peu exigeants en prétraitement des données, capables de gérer naturellement des interactions complexes et utilisables aussi bien en régression qu'en classification. Toutefois, ces qualités s'accompagnent d'une faiblesse majeure : les arbres de décision sont des modèles à *forte variance*.

En pratique, une légère modification du jeu de données peut conduire à un arbre très différent, tant dans sa structure que dans ses prédictions. Cette instabilité rend les arbres particulièrement sensibles au sur-apprentissage, notamment lorsque les données sont bruitées ou déséquilibrées.

Le but du bagging consiste donc à corriger ce défaut en construisant plusieurs arbres, puis en combinant leurs prédictions.

9.1.1 Principe général du bagging

Supposons que l'on dispose d'un modèle d'apprentissage μ, par exemple un arbre de décision. Le bagging construit plusieurs versions de ce modèle à partir de versions légèrement différentes du jeu de données initial, puis à agréger ces modèles pour former un prédicteur plus stable.

Le but est double :

— améliorer la capacité de généralisation du modèle ;

— réduire la variance sans augmenter significativement le biais.

Cette stratégie repose sur une technique clé de rééchantillonnage : le *bootstrap*.

9.1.2 Le bootstrap

On considère un échantillon d'apprentissage de taille N,

$$\mathcal{D} = \left\{(X_i^1, \ldots, X_i^p, Y_i)\right\}_{i=1}^N,$$

où Y est la variable réponse expliquée par les variables explicatives $X^1, \ldots, X^p$.

Définition 9.1.1 (Échantillon bootstrap)**.** *Un échantillon bootstrap est obtenu en tirant,* avec remise*, N observations dans $\mathcal{D}$. Certaines observations peuvent apparaître plusieurs fois, tandis que d'autres peuvent ne pas être sélectionnées.*

On répète cette opération B fois, ce qui conduit à B jeux de données bootstrap indépendants conditionnellement aux données initiales.

Pour chaque échantillon bootstrap $k \in \{1, \dots, B\}$, on entraîne un modèle μ_k (par exemple un arbre de décision).

Agrégation des modèles. Une fois les B modèles appris, on définit le modèle agrégé μ^+ par :

$$\mu^+(x) = \begin{cases} \dfrac{1}{B} \sum_{k=1}^{B} \mu_k(x), & \text{en régression,} \\ \text{vote majoritaire des } \big(\mu_k(x)\big)_{k=1}^{B}, & \text{en classification.} \end{cases}$$

Le formalisme associé au vote majoritaire est rédigé dans l'équation (9.1). Cette agrégation constitue le cœur du bagging.

9.1.3 Réduction de la variance par agrégation

L'efficacité du bagging repose sur un principe statistique fondamental : la moyenne de plusieurs estimateurs aléatoires est moins variable qu'un estimateur pris isolément.

Proposition 9.1.1. *Soient $Z_1, \dots, Z_B$ des variables aléatoires indépendantes et identiquement distribuées, d'espérance m et de variance σ^2. Alors*

$$\mathrm{Var}\left(\frac{1}{B}\sum_{k=1}^{B} Z_k\right) = \frac{\sigma^2}{B}.$$

Preuve. Par linéarité de la variance pour des variables **indépendantes**,

$$\mathrm{Var}\left(\frac{1}{B}\sum_{k=1}^{B} Z_k\right) = \frac{1}{B^2}\sum_{k=1}^{B} \mathrm{Var}(Z_k) = \frac{1}{B^2} \cdot B\sigma^2 = \frac{\sigma^2}{B}.$$

□

L'agrégation de prédicteurs indépendants permet ainsi, en théorie, de faire décroître la variance comme $1/B$. C'est précisément ce mécanisme que le bagging cherche à exploiter.

9.1.4 Limites du bagging

En pratique, les modèles $\mu_1, \dots, \mu_B$ ne sont pas indépendants. En effet, les échantillons bootstrap sont construits à partir du même jeu de données initial, ce qui induit une corrélation entre les modèles appris.

Pour comprendre l'impact de cette corrélation, regardons la proposition suivante.

Proposition 9.1.2 (Effet de la corrélation sur la variance d'une moyenne)**.** *Soient $Z_1, \dots, Z_B$ des variables aléatoires de même espérance m et de même variance σ^2. On suppose que, pour tout $k \neq \ell$,*

$$\mathrm{Corr}(Z_k, Z_\ell) = \rho.$$

Alors la variance de leur moyenne est donnée par

$$\mathrm{Var}\left(\frac{1}{B}\sum_{k=1}^{B} Z_k\right) = \rho\,\sigma^2 + \frac{1-\rho}{B}\,\sigma^2.$$

Preuve. La variance étant une forme quadratique,

$$\mathrm{Var}\left(\frac{1}{B}\sum_{k=1}^{B} Z_k\right) = \frac{1}{B^2}\,\mathrm{Var}\left(\sum_{k=1}^{B} Z_k\right).$$

Or, pour une somme de variables aléatoires quelconques,

$$\begin{aligned}
\mathrm{Var}\left(\sum_{k=1}^{B} Z_k\right) &= \mathrm{Cov}\left(\sum_{k=1}^{B} Z_k, \sum_{\ell=1}^{B} Z_\ell\right) \\
&= \sum_{k=1}^{B}\sum_{\ell=1}^{B} \mathrm{Cov}(Z_k, Z_\ell) && \text{Bilinéarité de l'opérateur de covariance} \\
&= \sum_{k=1}^{B} \mathrm{Cov}(Z_k, Z_k) \;+ \sum_{\substack{1\le k,\ell\le B \\ k\neq\ell}} \mathrm{Cov}(Z_k, Z_\ell) \\
&= \sum_{k=1}^{B} \mathrm{Var}(Z_k) \;+ \sum_{1\le k<\ell\le B} \mathrm{Cov}(Z_k, Z_\ell) \\
&\quad + \sum_{1\le \ell<k\le B} \mathrm{Cov}(Z_k, Z_\ell) \\
&= \sum_{k=1}^{B} \mathrm{Var}(Z_k) \;+\; 2 \sum_{1\le k<\ell\le B} \mathrm{Cov}(Z_k, Z_\ell), && \text{Symétrie de l'opérateur de covariance}
\end{aligned}$$

Comme $\mathrm{Var}(Z_k) = \sigma^2$ pour tout k et

$$\mathrm{Cov}(Z_k, Z_\ell) = \mathrm{Corr}(Z_k, Z_\ell)\,\sigma^2 = \rho\,\sigma^2,$$

on obtient

$$\mathrm{Var}\left(\sum_{k=1}^{B} Z_k\right) = B\sigma^2 + 2\binom{B}{2}\rho\,\sigma^2 = B\sigma^2 + B(B-1)\rho\,\sigma^2.$$

En divisant par B^2, il en résulte :

$$\mathrm{Var}\left(\frac{1}{B}\sum_{k=1}^{B} Z_k\right) = \frac{\sigma^2}{B} + \frac{B-1}{B}\,\rho\,\sigma^2 = \rho\,\sigma^2 + \frac{1-\rho}{B}\,\sigma^2.$$

□

Cette expression met en évidence une limite fondamentale du bagging : lorsque la corrélation ρ entre les modèles est non négligeable, le premier terme $\rho\,\sigma^2$ ne disparaît pas, même lorsque B est très grand. La réduction de variance est alors partielle.

Conséquence. Le bagging permet de réduire efficacement la variance des modèles instables, comme les arbres de décision, mais son efficacité est limitée par la corrélation entre les modèles agrégés. La question naturelle devient alors : *comment réduire cette corrélation tout en conservant des modèles performants individuellement ?*

C'est précisément à cette question que répondent les *forêts aléatoires*, en introduisant une source supplémentaire d'aléa lors de la construction des arbres, idée développée dans la section suivante.

9.2 Les forêts aléatoires

La forêt aléatoire (*Random Forest*) est une méthode d'apprentissage statistique qui prolonge naturellement le bagging. Elle repose sur la même idée centrale — agréger de nombreux arbres — mais introduit une source supplémentaire d'aléa lors de la construction de chaque arbre afin de réduire la corrélation entre modèles. Ce mécanisme permet d'améliorer la réduction de variance obtenue par l'agrégation et explique en grande partie les très bonnes performances empiriques des forêts aléatoires, en particulier lorsque la dimension p est grande.

9.2.1 Motivations

Pourquoi aller au-delà du bagging ? Le bagging réduit la variance d'un modèle instable (comme un arbre de décision) en moyennant les prédictions de nombreux modèles entraînés sur des échantillons bootstrap. Toutefois, comme discuté précédemment, l'efficacité de cette réduction est limitée par la corrélation entre les modèles agrégés. Si les arbres appris se ressemblent trop, la variance de la moyenne ne décroît pas fortement, à cause du terme de corrélation

$$\mathrm{Var}\left(\frac{1}{B}\sum_{k=1}^{B} Z_k\right) = \rho\,\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

Réduire ρ devient alors un objectif prioritaire.

Idée clé des forêts aléatoires. Le principe des forêts aléatoires est d'introduire un aléa supplémentaire, *au moment des coupures*, en limitant le choix des variables explicatives disponibles à chaque nœud. Cela force les arbres à explorer des structures différentes, même lorsqu'ils sont entraînés sur des échantillons bootstrap proches, et diminue ainsi la corrélation entre arbres.

Caractéristiques pratiques. Les forêts aléatoires sont appréciées car elles offrent généralement :

- de très bons résultats prédictifs, notamment en grande dimension ;
- une mise en œuvre simple et robuste ;
- peu d'hyperparamètres à régler (nombre d'arbres, profondeur, nombre de variables testées à chaque nœud) ;
- une capacité à gérer naturellement des interactions non linéaires et des variables de types variés.

9.2.2 Apprentissage d'une forêt aléatoire

On se place dans le cadre supervisé. On dispose d'un échantillon

$$\mathcal{D} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}, \qquad X_i = (X_i^1, \ldots, X_i^p) \in \mathbb{R}^p,$$

où $Y_i \in \mathbb{R}$ (régression) ou $Y_i \in \{1, \ldots, C\}$ (classification, où C désigne le nombre total de différentes modalité que peut prendre la réponse).

Rééchantillonnage bootstrap

Comme pour le bagging, on construit B échantillons bootstrap. Pour $k \in \{1, \ldots, B\}$, on définit

$$\mathcal{D}^{(k)} = \{(X_1^{(k)}, Y_1^{(k)}), \ldots, (X_N^{(k)}, Y_N^{(k)})\},$$

obtenu en tirant *avec remise* N observations dans $\mathcal{D}$.

On entraîne ensuite un arbre de décision μ_k sur chaque échantillon $\mathcal{D}^{(k)}$.

Aléa sur le choix des variables au moment des coupures

La différence fondamentale avec le bagging apparaît ici.

Dans un arbre de décision standard, lorsqu'on traite un nœud contenant un sous-échantillon, on explore l'ensemble des variables explicatives $\{1, \dots, p\}$ afin de choisir la meilleure coupure (au sens d'un critère d'impureté ou d'erreur).

Dans une forêt aléatoire, à chaque nœud :

1. on tire uniformément au hasard un sous-ensemble $J \subset \{1, \dots, p\}$ de cardinal m (avec $m \ll p$ en général) ;
2. on cherche la meilleure coupure *uniquement* parmi les variables indexées par J ;
3. on effectue la séparation correspondante et l'on poursuit récursivement.

Définition 9.2.1 (Paramètre m_{try})**.** *Le paramètre m (souvent noté m_{try} dans les logiciels) désigne le nombre de variables explicatives tirées aléatoirement à chaque nœud et parmi lesquelles on recherche la meilleure coupure.*

Formalisation du choix de coupure (classification). Soit $\mathcal{R}$ une région (nœud) de l'arbre, et J le sous-ensemble de variables tiré au hasard. Pour une coupure candidate (j, t) avec $j \in J$, on obtient deux sous-régions $\mathcal{R}_{\text{gauche}}$ et $\mathcal{R}_{\text{droite}}$. Si l'on utilise l'entropie comme critère d'impureté, la coupure est évaluée via

$$\frac{|\mathcal{R}_{\text{gauche}}|}{|\mathcal{R}|} H(\mathcal{R}_{\text{gauche}}) + \frac{|\mathcal{R}_{\text{droite}}|}{|\mathcal{R}|} H(\mathcal{R}_{\text{droite}}),$$

et l'on choisit la coupure minimisant cette quantité (équivalemment : maximisant le gain d'information).

Formalisation du choix de coupure (régression). En régression, un critère standard consiste à minimiser l'erreur quadratique intra-région. La coupure est choisie de manière à minimiser la somme des variances (ou des sommes de carrés résiduels) des sous-régions obtenues :

$$\sum_{i:\, X_i \in \mathcal{R}_{\text{gauche}}} \left(Y_i - \bar{Y}_{\text{gauche}}\right)^2 + \sum_{i:\, X_i \in \mathcal{R}_{\text{droite}}} \left(Y_i - \bar{Y}_{\text{droite}}\right)^2.$$

Remarque sur la réduction de corrélation. Même si deux arbres μ_k et μ_ℓ sont entraînés sur des échantillons bootstrap proches, ils peuvent diverger fortement en forêt aléatoire, car les variables candidates aux coupures ne sont pas les mêmes d'un nœud à l'autre. Ce mécanisme produit une diversité structurelle supplémentaire, ce qui réduit la corrélation entre arbres et améliore l'efficacité de l'agrégation.

Agrégation des arbres

Une fois les B arbres entraînés, on définit le prédicteur final μ^+ par agrégation.

Cas de la régression. Chaque arbre μ_k fournit une prédiction réelle $\mu_k(x) \in \mathbb{R}$. Le prédicteur final est la moyenne :

$$\mu^+(x) = \frac{1}{B} \sum_{k=1}^{B} \mu_k(x).$$

La perte associée est typiquement la perte quadratique, ce qui rend l'agrégation par moyenne particulièrement naturelle.

Cas de la classification. Chaque arbre μ_k fournit une classe prédite dans $\{1, \dots, C\}$. Le prédicteur final est défini par vote majoritaire :

$$\mu^+(x) = \underset{c \in \{1,\dots,C\}}{\operatorname{argmax}} \sum_{k=1}^{B} \mathbb{1}_{\{\mu_k(x) = c\}}. \tag{9.1}$$

Cette règle correspond au choix le plus fréquent parmi les prédictions individuelles. Elle est cohérente avec la minimisation empirique d'une perte de type 0–1.

Comparaison rigoureuse avec le bagging

Remarque 10 (Bagging vs forêts aléatoires). Le bagging et les forêts aléatoires partagent deux points communs : l'utilisation du bootstrap et l'agrégation de nombreux arbres. La différence essentielle est la suivante :

- **Bagging :** à chaque nœud, toutes les variables explicatives sont candidates pour choisir la meilleure coupure ;
- **Forêt aléatoire :** à chaque nœud, on restreint aléatoirement les variables candidates à un sous-ensemble de taille m.

Cette restriction augmente la diversité des arbres et tend à diminuer la corrélation entre modèles, ce qui rend la réduction de variance plus efficace lorsque l'on agrège les prédictions.

9.2.3 Avantages et limites

Avantages. Les forêts aléatoires présentent des avantages importants :

- **Excellente performance prédictive** dans de nombreux contextes, notamment en grande dimension ;
- **Robustesse** : peu sensibles au bruit et aux transformations monotones des variables ;
- **Simplicité d'usage** : peu d'hyperparamètres critiques ;
- **Réduction effective de variance** grâce à l'agrégation et à la réduction de corrélation.

Limites. Elles possèdent néanmoins certaines limites :

- **Perte d'interprétabilité** : une forêt de centaines d'arbres est moins lisible qu'un arbre unique ;
- **Coût de calcul** : apprentissage et prédiction plus lourds qu'un arbre seul, surtout si B est grand ;
- **Données très déséquilibrées** : la performance peut nécessiter des ajustements (pondération des classes, rééchantillonnage) ;
- **Modèle non paramétrique** : les garanties théoriques et l'inférence classique sont moins directes que pour les modèles linéaires.

Ces éléments expliquent pourquoi les forêts aléatoires constituent souvent un excellent compromis entre performance et robustesse, tout en servant de point d'entrée vers d'autres méthodes d'ensemble plus sophistiquées (boosting, gradient boosting, etc.).

Chapitre 10

La méthode des k plus proches voisins

10.1 Principe du plus proche voisin

Nous avons vu dans le chapitre précédent que les arbres de décision, et par extension les forêts aléatoires, construisent des modèles prédictifs en découpant l'espace des covariables au moyen de conditions explicites portant sur les coordonnées des observations. Ces découpages sont donc intrinsèquement axis-aligned et reposent sur une succession de séparations rectangulaires de l'espace.

Nous présentons à présent une approche conceptuellement différente, qui consiste à définir des régions de décision à partir d'une notion de *proximité* entre observations. Cette notion découle directement des outils topologiques introduits au chapitre 4. L'idée centrale est que des observations proches dans l'espace des variables explicatives devraient partager des réponses similaires. Ce principe informel est souvent résumé par l'adage « qui se ressemble s'assemble ».

L'algorithme dit du *plus proche voisin* repose entièrement sur cette intuition. Il s'agit d'une méthode non paramétrique, dans laquelle aucune hypothèse n'est faite sur la forme de la fonction reliant les variables explicatives à la variable réponse.

Considérons un échantillon d'apprentissage constitué de n observations

$$\big\{(X_1^1, \dots, X_1^p, Y_1), \dots, (X_n^1, \dots, X_n^p, Y_n)\big\},$$

où $(X_i^1, \dots, X_i^p) \in \mathbb{R}^p$ désignent les variables explicatives et $Y_i \in \{0, 1\}$ la variable réponse associée. Plus généralement, les variables explicatives vivent dans un espace $\mathcal{X}$ muni d'une distance d, au sens de la définition 4.1.1.

Définition 10.1.1. *On appelle* algorithme du plus proche voisin *l'algorithme qui, à une nouvelle observation $x \in \mathcal{X}$, associe la réponse de l'observation du jeu d'apprentissage la plus proche de x pour la distance d. Autrement dit, le prédicteur est défini par*

$$\hat{f}(x) = Y_{i^\star} \quad \text{où} \quad i^\star = \operatorname*{argmin}_{i=1,\dots,n} d(x, X_i).$$

Cette règle de décision induit une partition naturelle de l'espace $\mathcal{X}$. Pour chaque observation X_i du jeu d'apprentissage, on définit la région

$$V_i = \big\{x \in \mathcal{X} \ : \ \forall\, j \neq i,\ d(x, X_i) \leq d(x, X_j)\big\}.$$

Ces régions forment une partition de $\mathcal{X}$, appelée *diagramme de Voronoï*, au sens où

$$\mathcal{X} = \bigcup_{i=1}^{n} V_i, \quad V_i \cap V_j = \varnothing \text{ si } i \neq j.$$

Chaque région regroupe l'ensemble des points de l'espace dont l'observation X_i est le plus proche voisin.

La figure 10.1 dans le cas bidimensionnel permet d'illustrer géométriquement le mécanisme de cette partition. On pourra représenter des observations de différentes classes dans le plan (X^1, X^2) et visualiser les cellules de Voronoï associées, colorées selon l'étiquette Y_i de leur générateur. Ceci met en évidence la complexité potentielle des frontières de décision, qui ne sont ni linéaires ni nécessairement régulières.

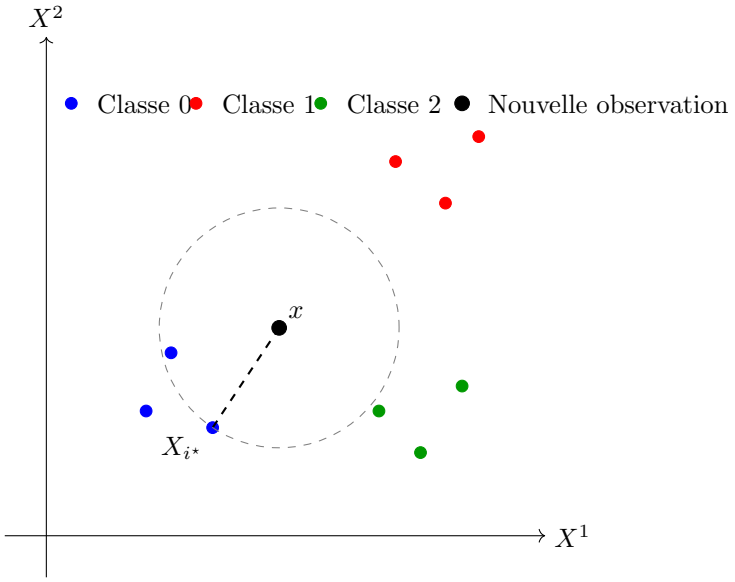


FIGURE 10.1 – Illustration du principe du plus proche voisin. La prédiction associée à x est celle de l'observation $X_{i^\star}$ minimisant la distance $d(x, X_i)$.

10.2 Motivation des k voisins

Bien que conceptuellement simple, la méthode du plus proche voisin présente une faiblesse majeure : elle est extrêmement sensible au bruit. En effet, si une observation du jeu d'apprentissage est mal étiquetée ou aberrante, alors l'ensemble de la cellule de Voronoï associée héritera de cette erreur. Dans un contexte réel, où les données sont presque toujours bruitées, ce phénomène peut conduire à des prédictions très instables.

Afin de rendre la méthode plus robuste, il est naturel de ne plus se fier à l'opinion d'un seul voisin, mais de considérer simultanément plusieurs observations proches de x. Cette idée conduit à l'algorithme des *k plus proches voisins*, souvent abrégé en KNN pour *k nearest neighbors*.

Définition 10.2.1. *Soit $K \in \mathbb{N}^*$. L'algorithme des* k plus proches voisins *consiste à associer à une observation $x \in \mathcal{X}$ une réponse construite à partir des réponses des K observations du jeu d'apprentissage les plus proches de x pour la distance d.*

On note $\mathcal{N}_K(x)$ l'ensemble des indices des K observations du jeu d'apprentissage les plus proches de x, défini par

$$\mathcal{N}_K(x) = \Big\{ i \in \{1, \dots, n\} \ : \ d(x, X_i) \leq d\big(x, X_{(K)}\big) \Big\},$$

où $d\big(x, X_{(K)}\big)$ désigne la K-ième plus petite valeur parmi les distances

$$\big\{ d(x, X_1), \dots, d(x, X_n) \big\}.$$

- *Dans un problème de classification, la règle de décision usuelle est le vote majoritaire. Le*

prédicteur s'écrit alors

$$\hat{f}(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i \in \mathcal{N}_K(x)} \mathbb{1}_{\{Y_i = c\}},$$

où $\mathbb{1}_{\{\cdot\}}$ *désigne la fonction indicatrice.*

- *Dans un problème de régression, la prédiction associée à* x *est définie comme la moyenne des réponses de ses* K *plus proches voisins :*

$$\hat{f}(x) = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x)} Y_i.$$

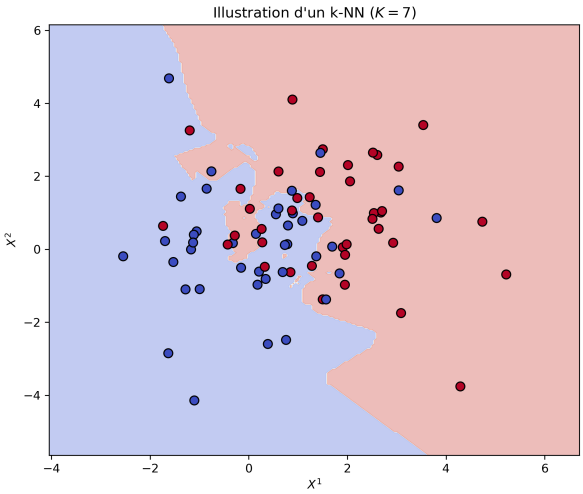


FIGURE 10.2 – Illustration de l'algorithme des k plus proches voisins ($K = 7$) pour une réponse binaire. La frontière de décision est non linéaire et dépend de la densité locale des observations.

Frontières de décision et effet du paramètre K. La frontière de décision induite par l'algorithme des k plus proches voisins est, en général, non linéaire. Elle est constituée de morceaux de segments ou de surfaces correspondant aux régions où la majorité des voisins change. Pour des valeurs faibles de K, ces frontières peuvent être très irrégulières et fortement influencées par des observations isolées.

Lorsque K augmente, l'effet du vote majoritaire conduit à un lissage progressif de la frontière de décision. Les fluctuations locales dues au bruit sont atténuées, au prix d'une perte de finesse dans la représentation des structures complexes des données. Ce phénomène illustre une fois encore le compromis biais–variance : un petit K engendre une faible biais mais une variance élevée, tandis qu'un grand K produit des prédictions plus stables mais potentiellement trop grossières.

Une illustration graphique dans le plan permet de visualiser cet effet. En représentant les régions de décision pour différentes valeurs de K, on observe clairement la simplification progressive des frontières lorsque K augmente, mettant en évidence le rôle central de ce paramètre dans le comportement de l'algorithme.

10.3 La variante des ε-voisins

Dans l'algorithme des k plus proches voisins, le nombre de voisins utilisés pour effectuer une prédiction est fixé *a priori* par le paramètre K, indépendamment de la densité locale des données. Cette approche peut s'avérer sous-optimale lorsque la distribution des observations est très hétérogène : dans certaines régions de l'espace des covariables, les données sont denses, tandis que dans d'autres elles sont rares.

10.3.1 Principe des ε-voisins

Une alternative naturelle consiste à ne plus fixer le nombre de voisins, mais à considérer tous les exemples d'apprentissage suffisamment proches de l'observation à étiqueter, au sens d'une distance absolue. L'idée sous-jacente est que les prédictions fondées sur des observations réellement proches sont intuitivement plus fiables que celles reposant sur des observations éloignées, même si ces dernières figurent parmi les plus proches relativement au reste de l'échantillon.

Cette idée conduit à l'algorithme dit des *ε-voisins*, qui repose sur le réglage d'un paramètre $\varepsilon > 0$ définissant un voisinage autour de l'observation x.

Définition 10.3.1. *On appelle* algorithme des ε-voisins*, ou* ε-ball neighbors*, l'algorithme qui associe à une observation $x \in \mathcal{X}$ une réponse construite à partir de l'ensemble des observations du jeu d'apprentissage situées à une distance inférieure ou égale à ε de x, c'est-à-dire*

$$\mathcal{N}_\varepsilon(x) = \big\{ i \in \{1, \dots, n\} \ : \ d(x, X_i) \leq \varepsilon \big\}.$$

- *Dans un problème de régression, la prédiction associée à x est définie comme la moyenne des réponses des observations appartenant à ce voisinage :*

$$\hat{f}(x) = \frac{1}{|\mathcal{N}_\varepsilon(x)|} \sum_{i \in \mathcal{N}_\varepsilon(x)} Y_i,$$

  à condition que $\mathcal{N}_\varepsilon(x)$ soit non vide.

- *Dans un problème de classification, la règle usuelle repose sur un vote majoritaire parmi les observations du voisinage :*

$$\hat{f}(x) = \operatorname*{argmax}_{c \in \mathcal{Y}} \sum_{i \in \mathcal{N}_\varepsilon(x)} \mathbb{1}_{\{Y_i = c\}}.$$

Remarque 11. Contrairement à la méthode des k plus proches voisins, le nombre d'observations utilisées pour la prédiction dépend ici de la densité locale des données. Dans les régions denses, le voisinage ε contient de nombreuses observations, tandis que dans les régions peu peuplées, il peut en contenir très peu, voire aucune.

10.3.2 Pondération par une fonction de similarité

Une limitation naturelle de la version non pondérée des ε-voisins est que toutes les observations situées dans la boule de rayon ε autour de x contribuent de manière identique à la prédiction, quelle que soit leur distance exacte à x. Or, il est raisonnable de penser que les observations les plus proches de x devraient avoir une influence plus importante que celles situées à la frontière du voisinage.

Pour remédier à cela, on introduit des poids dépendant de la distance, interprétés comme des mesures de similarité. Typiquement, on définit des poids

$$w_i(x) = \phi\big(d(x, X_i)\big),$$

où ϕ est une fonction décroissante. Deux choix classiques sont

$$w_i(x) = \frac{1}{d(x, X_i)} \qquad \text{ou} \qquad w_i(x) = \exp\left(-\frac{1}{2} d(x, X_i)^2\right).$$

Ces pondérations présentent un intérêt pratique majeur : elles permettent d'utiliser un voisinage relativement large tout en accordant une importance prédominante aux observations réellement proches de x. Elles offrent ainsi un compromis entre robustesse et précision locale.

Dans un problème de régression, le prédicteur pondéré s'écrit alors

$$\hat{f}(x) = \frac{1}{\displaystyle\sum_{i \in \mathcal{N}_\varepsilon(x)} w_i(x)} \cdot \sum_{i \in \mathcal{N}_\varepsilon(x)} w_i(x)\, Y_i.$$

Dans un problème de classification, la règle de décision devient

$$\hat{f}(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i \in \mathcal{N}_\varepsilon(x)} w_i(x)\, \mathbb{1}_{\{Y_i = c\}}.$$

10.3.3 Discussion et portée du modèle

Les variantes k-NN et ε-voisins mettent en évidence un point fondamental : l'ingrédient essentiel de ces algorithmes est la distance utilisée pour mesurer la proximité entre les observations. Le choix de la métrique, et plus généralement de la représentation des données dans l'espace $\mathcal{X}$, conditionne entièrement la qualité des prédictions.

Il convient toutefois d'être prudent. Ces méthodes sont particulièrement sensibles à la présence de covariables non pertinentes, qui interviennent néanmoins dans le calcul de la distance et peuvent en biaiser l'interprétation. De plus, en grande dimension, elles souffrent du phénomène connu sous le nom de *malédiction de la dimension* : lorsque la dimension de l'espace augmente, toutes les observations tendent à être éloignées les unes des autres, et la notion même de voisinage perd son sens. Dans ce contexte, l'intuition selon laquelle on peut prédire à partir des exemples proches cesse d'être valable, ce qui limite fortement l'efficacité de ces approches.

10.4 Choix d'un nombre élevé de voisins : avantages et limites

On se place dans le cadre où le nombre d'observations n est très grand, ce qui autorise le choix d'un paramètre K élevé sans contrainte liée à la taille de l'échantillon. Dans ce régime, l'algorithme des k plus proches voisins présente des propriétés asymptotiques intéressantes, mais également des limitations structurelles qui doivent être clairement identifiées.

Avantages. Un des premiers avantages que l'on peut soulever dans ce cadre est celui de la rédaction de la variance par moyennage. En effet, le prédicteur k-NN repose sur une moyenne empirique locale, ce qui s'écrit en régression comme tel :

$$\hat{f}(x) = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x)} Y_i.$$

Sous des hypothèses de régularité usuelles, on peut approximer la variance conditionnelle de cet estimateur par

$$\operatorname{Var}\left(\hat{f}(x) \mid X_1, \ldots, X_n\right) \approx \frac{1}{K} \operatorname{Var}(Y \mid X \approx x),$$

ce qui montre que l'augmentation de K entraîne une diminution de la variance de la prédiction. En classification, le vote majoritaire repose sur une estimation empirique de la probabilité conditionnelle

$$\mathbb{P}(Y = c \mid X = x) \approx \frac{1}{K} \sum_{i \in \mathcal{N}_K(x)} \mathbb{1}_{\{Y_i = c\}},$$

dont la stabilité s'améliore lorsque K augmente. Cette réduction de la variance rend les prédictions plus robustes au bruit et aux erreurs d'étiquetage.

Un second avantage à souligner est celui du lissage des frontières dans un cadre asymptotique. Lorsque K augmente, les frontières de décision induites par l'algorithme deviennent plus régulières. Les fluctuations locales dues à des observations isolées sont atténuées, et la frontière reflète davantage la structure moyenne des données. Dans un cadre asymptotique, il est possible de montrer que si

$$K \longrightarrow \infty \quad \text{et} \quad \frac{K}{n} \longrightarrow 0,$$

alors le prédicteur k-NN converge vers le prédicteur optimal, c'est-à-dire vers le classifieur de Bayes en classification ou vers l'espérance conditionnelle $\mathbb{E}[Y \mid X = x]$ en régression.

Limites. Une première limitation connu lorsque l'on cherche à réduire la variance du prédicteur est celle de la hausse du biais qui lui est induite. L'augmentation de K élargit mécaniquement le voisinage utilisé pour estimer la réponse en x. L'estimateur

$$\hat{f}(x) = \frac{1}{K} \sum_{i \in \mathcal{N}_K(x)} Y_i$$

n'approxime alors plus la quantité locale $\mathbb{E}[Y \mid X = x]$, mais une moyenne sur une région étendue de l'espace des covariables. Cette approximation introduit un biais croissant, car les observations éloignées de x peuvent avoir un comportement différent. Le caractère local de l'estimation, qui constitue la force principale de k-NN, est alors progressivement perdu.

Une autre limitation possible est celle d'un lissage excessif des frontières de décision. Lorsque K est trop grand, les frontières de décision deviennent artificiellement simples et peuvent ignorer des structures fines mais réelles des données. Dans le cas extrême où $K = n$, la prédiction devient indépendante de x :

$$\hat{f}(x) = \begin{cases} \mathbb{E}[Y] & \text{(régression)}, \\ \text{classe majoritaire globale} & \text{(classification)}. \end{cases}$$

L'algorithme cesse alors de tirer parti de la proximité entre observations et perd tout intérêt prédictif.

Enfin, un dernière limite à exposer est celle de l'interaction avec la dimension de l'espace. Même lorsque n est très grand, un nombre élevé de voisins ne permet pas de contourner le phénomène de la malédiction de la dimension (*curse of dimensionality*). Lorsque la dimension de l'espace des covariables augmente, les distances deviennent peu discriminantes et les voisins sélectionnés sont souvent éloignés de l'observation à étiqueter. L'augmentation de K peut alors aggraver ce phénomène en intégrant dans la prédiction des observations peu informatives, ce qui limite fortement l'efficacité de l'algorithme en grande dimension.

Ainsi, le paramètre K joue le rôle d'un véritable paramètre de régularisation. Son choix correspond à un compromis entre réduction de la variance et augmentation du biais, et doit être adapté à la structure géométrique et statistique des données plutôt qu'à la seule taille de l'échantillon.

10.5 Fonction de perte et construction des frontières de décision

Contrairement aux modèles paramétriques classiques étudiés précédemment, tels que la régression linéaire ou la régression logistique, les algorithmes de type k plus proches voisins ne reposent pas sur la minimisation explicite d'une fonction de perte globale lors d'une phase d'apprentissage. Il n'existe donc pas, à proprement parler, de *fonction de perte optimisée* au sens usuel du terme.

L'algorithme des k plus proches voisins relève d'une approche dite *paresseuse* (*lazy learning*) : l'apprentissage se limite à la mémorisation du jeu de données, et toute la complexité du modèle est reportée sur la phase de prédiction. Pour chaque nouvelle observation x, la décision est prise localement, en se fondant sur les observations du jeu d'apprentissage situées dans un voisinage de x.

D'un point de vue théorique, cette procédure peut être interprétée comme une approximation locale des prédicteurs optimaux issus de la théorie du risque. En classification, le classifieur de Bayes minimise la perte 0–1 et s'écrit

$$f^*(x) = \operatorname*{argmax}_{c \in \mathcal{Y}} \mathbb{P}(Y = c \mid X = x).$$

L'algorithme k-NN remplace alors la probabilité conditionnelle inconnue par une estimation empirique locale, obtenue à partir des étiquettes des voisins de x. Le vote majoritaire correspond ainsi à une minimisation locale du risque de classification.

De manière analogue, en régression avec perte quadratique, le prédicteur optimal est donné par l'espérance conditionnelle $f^*(x) = \mathbb{E}[Y \mid X = x]$, que k-NN approxime par la moyenne empirique des réponses des voisins de x. Là encore, aucune optimisation globale n'est effectuée : l'estimation est purement locale.

Les frontières de décision associées à k-NN ne sont donc pas construites explicitement. Elles émergent implicitement de la règle de décision locale et de la géométrie de l'espace des covariables.

Dans le cas $K = 1$, la frontière correspond à un diagramme de Voronoï exact, tandis que pour $K > 1$, le vote majoritaire induit un lissage progressif des frontières, dont la forme dépend fortement de la densité des données et du choix de la distance. Ce mécanisme explique à la fois la grande flexibilité de k-NN et sa sensibilité au bruit et à la dimension de l'espace.

10.6 Un algorithme supervisé et non paramétrique

L'algorithme des k plus proches voisins s'inscrit pleinement dans le cadre de l'apprentissage supervisé. En effet, le jeu de données d'apprentissage est constitué de couples (X_i, Y_i), où les variables explicatives X_i sont observées conjointement avec une variable réponse Y_i. La connaissance des étiquettes Y_i est indispensable au fonctionnement de l'algorithme : ce sont elles qui sont utilisées pour prédire la réponse associée à une nouvelle observation x.

Plus précisément, lors de la phase de prédiction, l'algorithme identifie les observations X_i les plus proches de x dans l'espace des covariables, puis exploite les valeurs correspondantes de Y_i pour construire la prédiction. En classification, la prédiction repose sur un vote majoritaire des étiquettes des voisins, tandis qu'en régression elle correspond à une moyenne locale des réponses. Sans l'accès aux valeurs Y_i, aucune règle de décision ne pourrait être définie, ce qui distingue fondamentalement k-NN des méthodes non supervisées telles que le clustering.

En revanche, l'algorithme des k plus proches voisins est dit *non paramétrique*. Contrairement aux modèles paramétriques classiques, tels que la régression linéaire ou la régression logistique, il ne suppose pas l'existence d'une fonction f appartenant à une famille paramétrée de dimension finie reliant X à Y. Il n'y a pas de vecteur de paramètres à estimer, ni de forme fonctionnelle globale imposée au modèle. La complexité du prédicteur n'est pas contrôlée par un nombre fini de paramètres, mais dépend directement de la taille et de la structure du jeu de données.

Cette absence de paramétrisation se traduit par un comportement très flexible : le prédicteur k-NN s'adapte localement à la géométrie des données et peut, en principe, approximer des relations très complexes entre les variables. En contrepartie, cette flexibilité s'accompagne d'une dépendance forte à la distance choisie et à la densité des observations, ainsi que d'un coût computationnel reporté sur la phase de prédiction plutôt que sur l'apprentissage.

Ainsi, k-NN illustre une distinction fondamentale en apprentissage statistique : un algorithme peut être supervisé sans être paramétrique. Il utilise explicitement les étiquettes observées pour apprendre à prédire, tout en s'affranchissant de toute hypothèse de forme globale sur la relation entre les variables explicatives et la variable réponse.