SY09
Science des Données

Gérard Govaert, Thierry Denœux, Benjamin Quost et Sylvain Rousseau

Table des matières

Notations

Terme	Description
$\det A$	le déterminant de la matrice carrée A
$\mathrm{diag}\,(A)$	le vecteur colonne défini par la diagonale de A si A est une carrée et la matrice diagonale de diagonale A si A est un vecteur
D_p	la matrice diagonale de dimension n des pondérations p_i
d_p	le vecteur colonne de dimension n des pondérations p_i
I_n	la matrice carrée identité de dimension n
$\mathbb{1}_n$	le vecteur colonne de dimension n rempli de 1
X	la matrice des données d'un tableau individus–variables de taille (n, p)
U_n	la matrice carrée de dimension n remplie de 1
$\mathrm{Tr}\,(A)$	la trace de la matrice carrée A
A^T	la matrice transposée de A

Première partie

Méthodes non supervisées

Chapitre 1

Introduction

Statistique et analyse de données

La Statistique est une discipline scientifique ayant pour objectif de rassembler et d'étudier des données chiffrées recueillies sur un sujet afin d'en tirer des informations. Le mot statistique est aussi utilisé pour désigner ces données chiffrées (exemple : les statistiques de la natalité). La statistique fait partie des *sciences du hasard* et son histoire est très liée à celle de la théorie des probabilités. Avant l'apparition, au 17^{e} siècle, de cette nouvelle science, la statistique resta essentiellement *descriptive*. Elle était utilisée, par exemple, par les États pour connaître leur population (richesse, activité,...) afin d'établir les impôts. Pour caractériser et résumer de tels tableaux de données, les outils utilisés sont variés : représentation graphique (carte géographique, histogramme,...), valeurs typiques (qui deviendront plus tard des paramètres de positionnement, de dispersion et de forme), ajustement, corrélation, indices.

Au début des années 1900, on voit se développer une nouvelle discipline scientifique à part entière : la *statistique mathématique*. Cette nouvelle discipline repose essentiellement sur la théorie des probabilités mais s'en distingue par ses objectifs : la théorie des probabilités, comme toutes les mathématiques, s'appuie sur un raisonnement purement déductif ; à partir d'axiomes, le calcul des probabilités établit un certain nombre de résultats. Par contre, la statistique mathématique cherche à *inférer* à partir des données la loi sous-jacente à ces observations. Parmi les principales méthodes développées en statistique, on peut citer les méthodes d'*estimation*, les *tests d'hypothèses*, la *régression*, la *discrimination* et l'*analyse de la variance*.

Parallèlement à ce développement et constatant le désintérêt des théoriciens pour les techniques descriptives, au début du siècle des chercheurs provenant d'autres disciplines, comme Spearman et Burt de l'école psychométrique américaine, développent des méthodes d'analyse qui cherchent à extraire des données l'« information pertinente » sans supposer aucun modèle probabiliste. Des méthodes comme l'analyse en composantes principales sont alors développées et constituent les premières méthodes d'*analyse des données*.

Data mining

L'apparition des moyens informatiques a eu un impact fondamental sur le développement de l'analyse des données et de la statistique. Les moyens de calcul ainsi disponibles ont permis, par exemple de rendre opérationnelle l'analyse en composantes principales qui, nécessitant la diagonalisation de matrices de grandes dimensions, n'était praticable que sur des petits jeux de données et au prix de longs calculs. Les outils de visualisation ont permis l'utilisation et la réalisation de graphiques en tout genre. Enfin, l'explosion

des données disponibles (données comptables, Web, téléphonie, données fournies par des appareils de mesure comme les images satellitaires ou capteurs de pollution,...la constitution d'entrepôt de données (*data warehouse*) ont encore accentué ce besoin d'analyse. Le perfectionnement des interfaces offre aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples des logiciels. Cette évolution ainsi que la popularisation de nouvelles méthodes algorithmiques (réseaux de neurones, *support vector machine*,...) et de moyens graphiques ont conduit au développement et à la commercialisation de logiciels intégrant des méthodes statistiques et algorithmiques sous la terminologie de *data mining*, quelquefois traduit par fouille de données.

L'objectif du *data mining* est l'analyse de grands jeux de données pour en extraire des informations pertinentes généralement dans une perspective d'aide à la décision. Domaine situé à l'intersection de la statistique et de l'informatique, le *data mining* s'appuie sur différentes familles de méthodes comme la statistique multivariée, l'analyse de données, l'apprentissage statistique (*Machine learning*), supervisé ou non supervisé, ou encore la reconnaissance des formes statistique.

Voici quelques exemples de problèmes abordés par le *data mining* : prédire si un patient, hospitalisé pour une crise cardiaque, aura une seconde crise à partir de données comme l'âge, le poids, la taille, les habitudes alimentaires et de mesures cliniques comme des analyses de sang ; estimer le taux de glucose dans le sang à partir d'un spectre d'absorption du sang ; prédire le prix d'une matière première à partir de données économiques et climatiques ; reconnaître le code postal sur une enveloppe à partir d'une image digitalisée ; identifier les facteurs de risque d'un cancer de la prostate à partir de variables ; détecter au plus vite des défaillances en contrôle de qualité ; gérer la relation client en marketing ; prévoir le marché pour une meilleure gestion des stocks ; recherche de « niche » ; détection de fraude bancaire ; analyse du comportement des internautes (*Web mining*).

Objectif de la première partie du cours

Il est classique de distinguer deux phases : une phase exploratoire et une phase d'apprentissage ou phase décisionnelle. La première partie de ce cours portera essentiellement sur la phase exploratoire dont les principaux objectifs sont la vérification de la cohérence du tableau de données (erreurs, valeurs manquantes, valeurs atypiques (*outliers*,...), la sélection de variables, le codage des variables (choix des unités de mesure, transformation de variables, par exemple, pour obtenir une distribution symétrique, découpage en classe,...), la recherche de relations intéressantes entre les variables et la recherche de typologie. Les principaux outils utilisés sont les résumés numériques, les représentations graphiques, la construction de variables synthétiques et l'identification de groupes homogènes dans la population étudiée.

Enfin, pour terminer cette introduction, on peut citer un certain nombre d'ouvrages généraux pouvant être utiles pour une bonne compréhension de ce cours : Duda et al. (2001), Flury (1997), Govaert (2003), Govaert (2009), Hastie et al. (2001), Lebart et al. (1995) et Saporta (2006).

Chapitre 2

Principaux types de données

Les données se présentent généralement sous forme de tableaux rectangulaires de nombres. Les plus courants sont les tableaux *individus–variables* et les tableaux de proximités.

1 Les données individus–variables

Dans ce premier paragraphe, les données à traiter, regroupées dans un tableau numérique de dimension (n, p) et représentées dans la figure 2.1, correspondent à un ensemble Ω de n *individus* pour lesquels on connaît la valeur de p *variables*.

	variable 1	$\cdots$	variable j	$\cdots$	variable p
individu 1	x_{11}		x_{1j}		x_{1p}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu i	x_{i1}		x_{ij}		x_{ip}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
individu n	x_{n1}		x_{nj}		x_{np}

FIGURE 2.1 – Tableau de données

On notera $X = (x_{ij})$ la matrice réelle à n lignes et p colonnes associée aux données et $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ et $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ les vecteurs colonnes associés aux individus et aux variables.

En statistique, un tel tableau de données peut être vu comme la réalisation d'un échantillon de taille n d'un vecteur aléatoire de dimension p . Ce vecteur aléatoire de dimension p défini par les variables aléatoires $X_1, \ldots, X_p$ sera noté $\boldsymbol{X} = (X_1, \ldots, X_p)^T$. L'étude de ces vecteurs aléatoires fera l'objet du chapitre A de l'annexe.

Suivant les valeurs que peuvent prendre ces variables, on distingue les variables *quantitatives* et les variables *qualitatives*.

1.1 Variables quantitatives

La variable est dite *quantitative* lorsqu'il s'agit d'une application de l'ensemble des individus Ω dans l'ensemble des réels $\mathbb{R}$ et que la notion de somme, de produit par un réel et d'ordre a un sens pour les valeurs de cette variable. Par exemple, la taille, le poids ou la teneur en minerai vérifient ces propriétés. On classe généralement aussi dans cette catégorie des variables qui ne présentent pourtant pas toutes ces propriétés, par exemple la température pour laquelle la notion de produit par un réel n'a pas toujours de sens.

Une distinction est souvent faite entre *variable continue* (ou *mesure*) et *variable discrète*

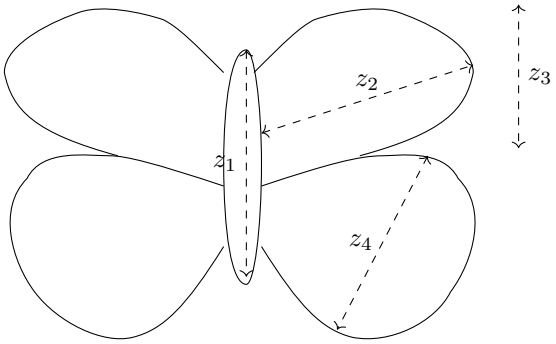


FIGURE 2.2 – Variables

	z_1	z_2	z_3	z_4
0	22	35	24	19
1	24	31	21	22
2	27	36	25	15
3	27	36	24	23
4	21	33	23	18
5	26	35	23	32
6	27	37	26	15
7	22	30	19	20
8	25	33	22	22
9	30	41	28	17
10	24	39	27	21
11	29	39	27	17
12	29	40	27	17
13	28	36	23	24
14	22	36	24	20
15	23	30	20	20
16	28	38	26	16
17	25	34	23	14
18	26	35	24	15
19	23	37	25	20
20	31	42	29	18
21	26	34	22	21
22	24	38	26	21

TABLE 2.2 – Jeu de données

suivant que les valeurs prises par la variable appartiennent à un intervalle réel ou à un sous-ensemble fini ou dénombrable de $\mathbb{R}$.

Le tableau 2.1 correspond à un exemple de données regroupant les notes obtenues par 9 élèves en mathématiques, sciences, français, latin et dessin-musique.

	math	scie	fran	lati	d-m
jean	6.00	6.00	5.00	5.50	8
alin	8.00	8.00	8.00	8.00	9
anni	6.00	7.00	11.00	9.50	11
moni	14.50	14.50	15.50	15.00	8
didi	14.00	14.00	12.00	12.50	10
andr	11.00	10.00	5.50	7.00	13
pier	5.50	7.00	14.00	11.50	10
brig	13.00	12.50	8.50	9.50	12
evel	9.00	9.50	12.50	12.00	18

TABLE 2.1 – Les données Notes

Le tableau 2.2 a été établi par les enfants d'une classe élémentaire : après avoir collecté 23 papillons, les enfants ont reporté dans un tableau les mesures de 4 longueurs (z_1, z_2, z_3 et z_4) mesurées en mm sur chacun des papillons (voir figure 2.2). Ces papillons appartenaient à différentes espèces et l'objectif est d'essayer de retrouver ces espèces à partir uniquement des 4 mesures.

1.2 Variables qualitatives

Cette fois, on suppose que l'espace d'arrivée est un ensemble fini. Les éléments de cet ensemble sont appelés *modalités*. Le numéro de département, la catégorie socio-professionnelle sont des exemples de variables qualitatives.

Lorsqu'il y a une *relation d'ordre* sur l'ensemble des modalités, on parle de variables qualitatives *ordinales*. Par opposition, les premières sont appelées variables qualitatives *nominales*. Par exemple, dans un sondage d'opinion, lorsque l'on demande de caractériser un produit en répondant « très bon », « bon », « moyen », « mauvais », « très mauvais », on obtient une variable qualitative ordinale à cinq modalités.

La distinction entre ces deux types de variables qualitatives est importante : en effet, utiliser des méthodes prévues pour des variables nominales sur des variables ordinales

conduira à négliger une partie de l'information ; au contraire, utiliser des méthodes prévues pour des variables ordinales sur des variables nominales conduira à ajouter de l'information incorrecte.

Les deux types de variables (quantitatives et qualitatives) seront souvent présentes dans un tableau de données. Par exemple, les données de la table 2.3 proposées par Fisher pour illustrer les méthodes de discrimination, définies à partir de 150 iris provenant de 3 espèces différentes, Virginia, Versicolor et Setosa, sur lesquelles ont été mesurées les longueurs et les largeurs du sépale et du pétale, sont constituées d'une variable qualitative nominale à 3 modalités et de 4 variables quantitatives.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
...					
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
...					
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
...					
150	5.9	3.0	5.1	1.8	virginica

TABLE 2.3 – Les données Iris

1.3 Variables binaires

L'ensemble d'arrivée est maintenant un ensemble à deux éléments souvent codés 0 et 1. Il s'agit donc d'une variable qualitative particulière.

Là aussi, on peut rencontrer deux situations : les deux modalités sont parfaitement symétriques (par exemple, féminin ou masculin) ou, au contraire, il existe une relation d'ordre entre les deux modalités (par exemple, dans les tableaux de présence-absence, la présence est souvent considérée comme une information plus importante que l'absence). Dans le premier cas, il faudra utiliser des méthodes d'analyse qui traitent de manière symétrique les deux modalités. Par contre, dans le second cas on pourra faire jouer un rôle différent aux 2 modalités.

La distinction entre les types de variables peut être quelquefois un peu arbitraire. Les notes scolaires en sont un exemple : si celles-ci peuvent en effet être clairement considérées comme des variables qualitatives lorsque l'on utilise les notes A, B, C, D et E et comme des variables quantitatives lorsque l'on utilise, par exemple, les notes entre 0 et 20 avec une précision de 0.1, on peut s'interroger sur la nature de cette note lorsqu'elle appartient à un ensemble plus restreint, comme les entiers de 0 à 20.

1.4 Transformation de variables

Pour étudier simultanément plusieurs variables, il est souvent nécessaire de faire des prétraitements. En voici quelques exemples.

Variable quantitative en variable quantitative

Pour rendre homogènes plusieurs variables quantitatives, les transformations les plus utilisées sont le *centrage* qui soustrait la moyenne à chaque valeur, la *réduction* qui divise chaque valeur par l'écart-type ou encore le *centrage-réduction* qui enchaîne ces deux transformations.

On peut aussi créer une nouvelle variable quantitative en effectuant une combinaison linéaire des variables initiales. Par exemple, la note finale à un examen est obtenue en faisant la somme, pondérée par des coefficients, des notes de chaque matière.

Variable quantitative en variable qualitative

Les principales méthodes statistiques supposent que les variables sont toutes de même type. Or, généralement, les données comportent à la fois des variables quantitatives et qualitatives. Il est alors nécessaire d'effectuer des transformations pour obtenir des variables de même nature.

Pour transformer une variable quantitative en variable qualitative, la méthode la plus utilisée consiste à découper l'ensemble d'arrivée de la variable quantitative en un ensemble de m intervalles consécutifs. On obtient alors une variable qualitative ordinale à m modalités. La difficulté porte sur la définition de ce découpage. Plusieurs techniques peuvent être utilisées :

— découpage défini *a priori* : par exemple, on remplace l'âge par une des valeurs 1, 2, 3 ou 4 suivant les intervalles : 0–18 ans, 19–40 ans, 41–65 ans, plus de 65 ans ;

— découpage défini en utilisant la « forme » de l'histogramme (recherche de modes) ;

— découpage en intervalles de même longueur : il suffit de préciser le nombre d'intervalles et les bornes ;

— découpage en intervalles d'effectifs égaux : il suffit de préciser le nombre d'intervalles.

Variable qualitative en variable binaire

Pour passer d'une variable qualitative à une variable binaire, la transformation la plus utilisée, appelée *codage disjonctif complet*, consiste à remplacer la variable qualitative par les indicatrices de chaque modalité. Dans l'exemple suivant, une variable qualitative à 3 modalités a été remplacée par 3 variables binaires.

	v
1	3
2	1
3	3
4	2
5	1

	v1	v2	v3
1	0	0	1
2	1	0	0
3	0	0	1
4	0	1	0
5	1	0	0

Remarquons que si la variable est qualitative ordinale, l'ordre des modalités est perdu. Dans ce cas, on peut utiliser le *codage additif*. Pour le même exemple, le résultat est maintenant le suivant :

	v
1	3
2	1
3	3
4	2
5	1

	v1	v2	v3
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	0
5	1	0	0

2 Tableaux de proximités

Un tableau de proximité est un tableau carré de nombres mesurant une ressemblance ou une dissemblance entre les éléments d'un ensemble Ω. On peut citer par exemple les tableaux de distances géographiques, les tableaux de distances routières, les tableaux de durées du trajet par le train, les tableaux de corrélations entre variables.

2.1 Types de proximités

Les mesures de proximités mesurent à quel point deux objets sont proches. Elles se décomposent en deux grandes familles, les mesures de similarité et les mesures de dissimilarité.

Mesures de similarité

Les mesures de similarité sont d'autant plus grandes que les deux objets comparés sont proches. Plus précisément, on adopte la définition suivante.

Définition 1. *Une* mesure de similarité *sur un ensemble* Ω *est une fonction* s *de* $\Omega \times \Omega$ *dans* $\mathbb{R}^+$ *vérifiant*

(1) $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \quad s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x})$ *(symétrie)*

(2) $\forall \boldsymbol{x} \neq \boldsymbol{y} \in \Omega, \quad s(\boldsymbol{x}, \boldsymbol{x}) = s_{max} \quad et \quad s_{max} \geq s(\boldsymbol{x}, \boldsymbol{y})$

Mesure de dissimilarité

Les mesures de dissimilarité sont les plus souvent rencontrées. Elles correspondent à l'idée intuitive qu'on se fait d'une distance : plus elle est grande, plus les objets sont éloignés. On commence par la dissimilarité la plus restrictive, la dissimilarité euclidienne ou distance euclidienne.

Définition 2. *Une* distance euclidienne d *sur* Ω *est une application de* $\Omega \times \Omega$ *dans* $\mathbb{R}^+$ *telle qu'il existe un entier* k *et un plongement* p *de* Ω *dans* $\mathbb{R}^k$ *tel que*

$$\forall \omega, \omega' \in \Omega, \quad d(\omega, \omega') = \|p(\omega) - p(\omega')\|_2 .$$

En d'autres termes, une distance euclidienne peut être vue comme la distance usuelle dans un espace vectoriel $\mathbb{R}^k$. Les dissimilarités rencontrées sont rarement euclidiennes. Une définition moins restrictive est la notion de distance.

Définition 3. *Une* distance d *sur un ensemble* Ω *est une application de* $\Omega \times \Omega$ *dans* $\mathbb{R}^+$ *vérifiant les propriétés suivantes :*

(1) $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \quad d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ *(séparation)*

(2) $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ *(symétrie)*

(3) $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \Omega, \quad d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$ *(inégalité triangulaire)*

La définition 3 est effectivement moins restrictive comme le montre l'exemple de 4 points présents dans la figure 2.3. Les dissimilarités entre les 4 points en font une distance. En revanche, cette distance ne peut pas être euclidienne. En effet, la longueur des arêtes intérieures vaut au minimum $\sqrt{4/3} \simeq 1.15$ lorsque les 4 points sont coplanaires. Or ces mêmes distances valent 1.1. La distance ne peut donc pas être une distance euclidienne.

En analyse des données, il n'est pas toujours nécessaire d'avoir toutes ces propriétés et une mesure de dissimilarité est souvent suffisante.

Définition 4. *Une* mesure de dissimilarité *sur un ensemble* Ω *est une fonction* d *de* $\Omega \times \Omega$ *dans* $\mathbb{R}^+$ *vérifiant*

(1) $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ *(symétrie)*

(2) $\forall \boldsymbol{x} \in \Omega, \quad d(\boldsymbol{x}, \boldsymbol{x}) = 0$

Enfin, terminons en citant la distance ultramétrique, fondamentale pour l'étude de la classification hiérarchique.

Définition 5. *Une* ultramétrique *sur un ensemble* Ω *est une fonction* d *de* $\Omega \times \Omega$ *dans* $\mathbb{R}^+$ *vérifiant*

(1) $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega, \quad d(\boldsymbol{x}, \boldsymbol{y}) = 0 \iff \boldsymbol{x} = \boldsymbol{y}$ *(séparation)*

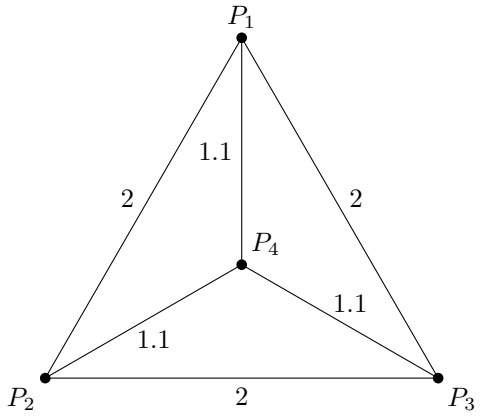


FIGURE 2.3 – Points définissant une distance qui n'est pas euclidienne

(2) $\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega$, $\quad d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ *(symétrie)*

(3) $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \Omega$, $\quad d(\boldsymbol{x}, \boldsymbol{z}) \leq \max(d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{z}))$ *(inégalité ultramétrique)*

Il est facile de montrer que la propriété d'inégalité ultramétrique entraîne la propriété d'inégalité triangulaire. Une ultramétrique est donc une distance. En revanche, une distance ultramétrique n'est pas nécessairement euclidienne.

Notation

Les mesures de proximités portant sur un ensemble fini Ω, plutôt que de noter $d(\omega_i, \omega_j)$, on note souvent d_{ij}.

2.2 Constitution d'un tableau de proximités

Un tableau de proximités peut être issu directement du recueil des données, par exemple les tableaux de distances routières, ou peut être obtenu à partir d'un tableau initial individu–variable. En voici quelques exemples.

Variables quantitatives

Si les variables sont quantitatives, les distances les plus utilisées sont les suivantes :

- distance euclidienne : $d^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_j (x_j - y_j)^2 = (\boldsymbol{x} - \boldsymbol{y})^T I (\boldsymbol{x} - \boldsymbol{y})$;
- distance euclidienne pondérée : $d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T D (\boldsymbol{x} - \boldsymbol{y})$;
- distance de Mahalanobis : $d^2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T S^{-1} (\boldsymbol{x} - \boldsymbol{y})$ où S est la matrice de variance empirique ;
- distance « Manhattan » ou distance L_1 : $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{p} |x_j - y_j|$;
- distance de Chebychev ou distance L_∞ : $d(\boldsymbol{x}, \boldsymbol{y}) = \max_j |x_j - y_j|$;
- distance de Minkowski L_p : $d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{p} (x_j - y_j)^p \right)^{1/p}$ (il s'agit de la généralisation des distances précédentes : L_1=Manhattan, L_2=euclidienne, L_∞=Chebychev).

Enfin la relation

$$d^2(\mathbf{x}_j, \mathbf{x}_{j'}) = 1 - \mathrm{Cor}\,(\mathbf{x}_j, \mathbf{x}_{j'}),$$

permet de définir une distance entre variables à partir de la corrélation linéaire.

Variables qualitatives

Si les variables qualitatives sont nominales, il est possible d'utiliser la distance du χ^2 ou, plus simplement, la distance $d = 1-$ proportion de modalités communes. Cette dernière peut être généralisée en utilisant une table de ressemblance entre modalités.

Si les variables qualitatives sont ordinales, la distance euclidienne sur les rangs renormalisés entre 0 et 1 peut être utilisée.

Variables binaires

Pour les variables binaires, en notant a, b, c et d le nombre de fois où les individus $\boldsymbol{x}$ et $\boldsymbol{y}$ ont répondu respectivement $(1,1)$, $(1,0)$, $(0,1)$ et $(0,0)$ aux variables binaires, alors toute une série de mesures de proximité peuvent être définies. Par exemple :

- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{2a}{2a+b+c}$ (Csekanowski, Sorensen, Dice) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a-(b+c)+d}{a+b+c+d}$ (Hamman) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a}{a+b+c}$ (Jaccard) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a}{a+b}$ (Kulezynsky) ;
- $d(\boldsymbol{x},\boldsymbol{y}) = \frac{a}{[(a+b)(a+c)]^{1/2}}$ (Ochiai).

2.3 Transformation

Il existe de nombreux moyens de passer d'un type de proximités à un autre. Remarquons qu'il est facile de transformer une mesure de similarité s en une mesure de dissimilarité en posant

$$d_{ij} = s_{\max} - s_{ij}.$$

On peut symétriser une proximité p_{ij} avec la transformation

$$p_{ij}^{\text{sym}} = \frac{p_{ij} + p_{ji}}{2}.$$

Une dissimilarité d_{ij} peut être transformée en distance en ajoutant une constante c à d_{ij} pour $i \neq j$ définie par

$$c = \max_{i,j,k} d_{ij} - d_{ik} - d_{kj}.$$

2.4 Utilisation

Les mesures de proximités peuvent être intégrées dans les méthodes (ACP, ACM, méthode des centres-mobiles, discrimination linéaire ou quadratique,...) ou peut être la donnée de base de la méthode (AFTD, MDS, classification hiérarchique). Lorsqu'il n'existe pas de méthode adaptée à un type de données, il est souvent possible de définir une proximité cohérente avec les données et d'appliquer ce dernier type de méthodes.

Chapitre 3

Méthodes exploratoires élémentaires

Avant d'aborder des méthodes de représentation relativement complexes comme l'analyse en composantes principales ou la classification automatique, nous présentons dans ce chapitre les principaux outils de statistique exploratoire (ou descriptive) élémentaire. Remarquons que leur utilisation peut aller très loin et fait même l'objet d'une méthode d'analyse complète appelée EDA (Exploratory data analysis) (Tukey, 1977 ; Chambers et al., 1983 ; Tukey, 1983 ; Cleveland, 1994b,a). Ces outils dépendent de la forme des données qui ont été décrites dans le chapitre précédent. Nous nous sommes limités dans ce chapitre à la description des tableaux *individus-variables*.

1 Description des variables quantitatives

Rappelons que dans ce cas, les données à traiter, regroupées dans un tableau numérique X de dimension (n, p), correspondent à un ensemble Ω de n *individus* pour lesquels on connaît la valeur de p *variables* et que ce tableau peut être considéré comme la réalisation d'un échantillon de taille n du vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_j, \ldots, X_p)$. Chaque vecteur $\boldsymbol{x}_i$ correspond à une réalisation du vecteur aléatoire $\boldsymbol{X}$; chaque vecteur $\mathbf{x}_j$ correspond à une réalisation d'un échantillon de taille n de la variable aléatoire X_j. La description des données peut alors s'appuyer sur les principales statistiques définies à partir de cette matrice X : vecteur moyenne empirique, variance empirique, covariance empirique, coefficient de corrélation linéaire empirique, matrice de variance empirique, matrice de corrélation empirique.

1.1 Description monodimensionnelle

Dans ce paragraphe, l'objectif est de décrire l'ensemble des valeurs $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ correspondant à une variable quantitative.

Statistiques élémentaires

Les statistiques les plus simples sont le minimum et le maximum. D'autres mesurent la valeur centrale comme, par exemple, la moyenne empirique

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij},$$

ou la médiane qui partage les valeurs ordonnées prises par la variable en deux parties égales. On peut généraliser cette notion de médiane en utilisant les quantiles d'ordre p

qui partagent en p quantités égales l'ensemble étudié ; Les quartiles, au nombre de 3, partagent en 4 parties de même effectif la population totale ; le premier quartile q_1 laisse à gauche 25% de la population, le deuxième q_2 est la médiane et le troisième q_3 laisse à gauche 75% de la population. Enfin, certaines statistiques mesurent la dispersion ; par exemple l'étendue (maximum-minimum), la variance empirique

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2,$$

et l'écart-type empirique $s_j = \sqrt{s_j^2}$; ces deux statistiques sont des caractéristiques de dispersion autour de la moyenne ; la largeur de l'intervalle interquartile, ou encore étendue interquartile, définie par la valeur $q_3 - q_1$ contient 50% de la population et constitue une caractéristique de dispersion autour de la médiane

Notons que la moyenne et l'écart-type sont des caractéristiques statistiques sensibles aux valeurs extrêmes, ce qui n'est pas le cas pour la médiane et l'intervalle interquartile. La figure 3.1 fournit quelques unes de ces statistiques pour les variables quantitatives des données `Iris`.

	sepal_length	sepal_width	petal_length	petal_width
count	150.0000	150.0000	150.0000	150.0000
mean	5.8433	3.0573	3.7580	1.1993
std	0.8281	0.4359	1.7653	0.7622
min	4.3000	2.0000	1.0000	0.1000
25%	5.1000	2.8000	1.6000	0.3000
50%	5.8000	3.0000	4.3500	1.3000
75%	6.4000	3.3000	5.1000	1.8000
max	7.9000	4.4000	6.9000	2.5000

FIGURE 3.1 – Description élémentaire des données `Iris`

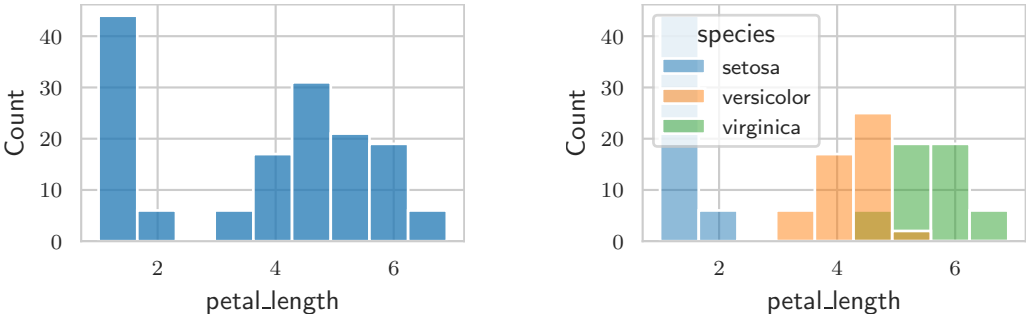
Histogramme

Pour tracer un histogramme, il suffit de découper l'intervalle $[\min, \max]$ en un certain nombre d'intervalles disjoints et d'associer à chaque intervalle un rectangle dont l'aire est proportionnelle à la fréquence des individus ayant pris leur valeur dans cet intervalle. Si la longueur de chaque intervalle est constante, les rectangles ont alors une hauteur proportionnelle à la fréquence ou à la fréquence relative. Le choix du nombre d'intervalles peut avoir une influence assez grande sur la forme de l'histogramme. Il existe un certain nombre de règles empiriques conseillées pour effectuer ce choix. Ainsi, la règle de Sturges recommande de prendre un nombre de classes égal à l'entier immédiatement supérieur ou égal à $1 + \log_2 n$, n étant la taille de l'échantillon. La figure 3.2a correspond à l'histogramme obtenu avec la variable longueur du pétale des données `Iris`. On peut aussi tenir compte de la répartition des iris suivant les 3 espèces à la figure 3.2b. Cette figure montre clairement que la variable longueur du pétale discrimine clairement les 50 premières fleurs des suivantes.

Estimation de la densité par la méthode des noyaux

L'histogramme, lorsqu'il s'appuie sur les fréquences relatives, peut être vu comme une estimation de la fonction de densité d'une variable aléatoire X. Il parait alors souhaitable d'estimer cette fonction de densité par une fonction plus régulière que l'histogramme. Différentes méthodes d'estimation non paramétriques ont été développées mais la plus utilisée est la méthode du noyau définie de la manière suivante : si f est la densité de probabilité de la variable aléatoire X, il est possible de montrer la relation

$$f(x) = \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(x - h/2 \leqslant X \leqslant x + h/2\right).$$



(a) Histogramme de la longueur des pétales.

(b) Histogramme de la longueur des pétales par espèce.

FIGURE 3.2 – Histogrammes des données `Iris`.

Un estimateur simple de cette densité consiste alors à estimer $\mathbb{P}(x - h/2 < X < x + h/2)$ par la proportions des observations $(x_1, \ldots, x_n)$ situées dans l'intervalle $[x-h/2, x+h/2]$:

$$\widehat{f}(x) = \frac{1}{nh}(\text{nombre d'observations situées dans } [x - h/2, h + h/2]).$$

En introduisant la fonction

$$K(x) = \mathbb{1}_{[-0.5,+0.5]}(x),$$

l'estimateur s'écrit alors

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

Comme l'histogramme, ceci conduit à des fonctions discontinues, ce qui n'est pas entièrement satisfaisant pour estimer des fonctions de densité souvent régulières. Pour ceci, il suffit de s'appuyer sur d'autres fonctions K, appelées *fonctions noyaux*, plus régulières ; la condition étant de prendre des fonctions d'intégrale 1 sur $\mathbb{R}$ et positive, c'est-à-dire des fonctions de densité. Les principaux noyaux utilisés sont les suivants :

- noyau rectangulaire : $K(x) = \mathbb{1}_{[-0.5,+0.5]}(x)$;
- noyau triangulaire : $K(x) = (1 - |x|)\mathbb{1}_{[-1,+1]}(x)$;
- noyau gaussien : $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$;
- noyau d'Epanechnikov : $K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,+1]}(x)$;
- noyau de Lejeune : $K(x) = \frac{105}{64}(1 - x^2)^2(1 - 3x^2)\mathbb{1}_{[-1,+1]}(x)$.

En pratique, l'estimateur sera moins sensible au choix du noyau qu'au choix de h qui règlera le degré de régularité. Par ailleurs, le noyau de Lejeune se comporte correctement bien que la condition de positivité ne soit pas remplie. La figure 3.3 représente un échantillon de taille 50 et l'estimation de densité obtenue avec le noyau gaussien et l'histogramme obtenu avec 9 classes.

Diagramme en boîte

La figure 3.4 représente les diagrammes en boîte, encore appelés boîtes à moustaches ou *boxplots*, associés à chacune des 4 variables des données `Iris`.

Chacun de ces graphiques est constitué d'un rectangle et de 2 moustaches. Le rectangle est délimité par les quartiles et partagé en deux par la médiane. Pour définir les moustaches, il est nécessaire de définir tout d'abord la notion d'éléments atypiques (*outliers*) : il s'agit de valeurs relativement éloignées des autres. Ici, elles sont définies en prenant les valeurs distantes de l'intervalle interquartile d'une valeur supérieure à 1.5 fois la longueur de cet intervalle. Les valeurs minimum et maximum de l'échantillon auquel on

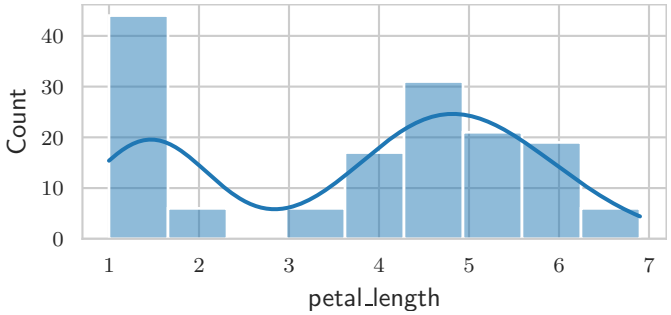


FIGURE 3.3 – Estimation de densité et histogramme

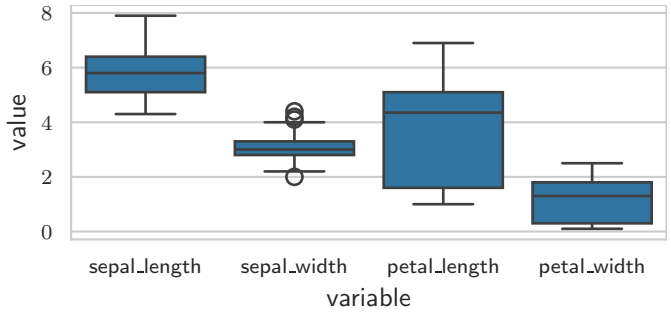


FIGURE 3.4 – Diagramme en boîte pour les données `Iris`

a enlevé ces éléments atypiques ainsi que les éléments atypiques eux-mêmes forment alors la moustache.

Cette première étape devrait déjà permettre de mettre en évidence certaines caractéristiques comme la présence de données aberrantes, l'absence de symétrie de la distribution ou encore la présence de populations hétérogènes.

1.2 Description bidimensionnelle

Cette fois, l'objectif est de décrire les liens pouvant exister entre deux variables $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ et $\mathbf{x}_{j'} = (x_{1j'}, \ldots, x_{nj'})^T$.

Graphique de dispersion

En représentant chaque individu i par le point de coordonnées (x_{i1}, x_{i2}), on obtient un nuage de n points dans le plan. Cette représentation permet de visualiser de manière synthétique et claire les données et de voir rapidement, par exemple, si une relation existe entre ces deux variables. Si les points semblent avoir été disséminés au hasard alors il n'y a aucune relation entre les deux variables. Si les points se regroupent autour d'une droite alors il y a une liaison linéaire entre ces deux variables et cette liaison peut être quantifiée par le coefficient de corrélation. Si les points se regroupent autour d'une fonction non linéaire (par exemple fonction polynomiale, logarithmique,...) alors une transformation de l'une des variables par cette fonction permet d'avoir une liaison linéaire entre cette nouvelle variable et l'autre variable. Par exemple, la figure 3.5 qui représente les variables mathématiques et sciences du tableau de notes permet de visualiser une relation linéaire entre les 2 variables.

Dans le paragraphe suivant, nous verrons comment ce type de représentation peut mettre en évidence respectivement une absence de liaison, une absence de liaison en moyenne mais pas en dispersion, une relation linéaire et enfin une relation non linéaire.

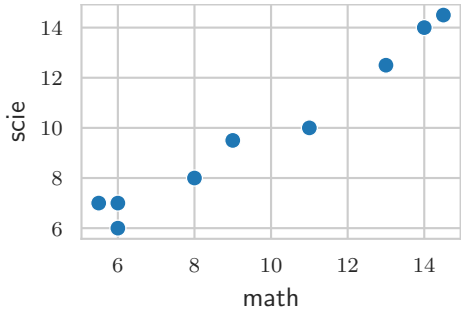


FIGURE 3.5 – Graphique de dispersion pour les données Notes

Covariance et corrélation empiriques

Pour étudier les liens entre deux variables quantitatives, on utilise souvent la covariance empirique

$$s_{jj'} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ij'} - \overline{x}_{j'})$$

et le coefficient de corrélation linéaire empirique

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}}.$$

Remarquons que l'on a $s_{jj} = (s_j)^2$ et $r_{jj} = 1$.

Le coefficient de corrélation linéaire est à utiliser avec prudence : il est en effet difficile avec un seul nombre de caractériser entièrement le lien qui peut exister entre deux variables. Par exemple la figure 3.6, qui représente les graphiques de dispersion correspondant à des échantillons associés à deux variables aléatoires, recouvre différentes situations que le seul coefficient de corrélation ne peut expliquer :

- cas (a) : r petit et indépendance entre les variables ;
- cas (b) : r petit mais variance de Y dépendant de la variable X ;
- cas (c) : r grand et forte dépendance linéaire ;
- cas (d) : r petit et forte dépendance non linéaire.

La figure 3.7 complète ces exemples. La première ligne correspond à un coefficient de corrélation linéaire faible, la seconde ligne à un coefficient de corrélation linéaire fort avec, à chaque fois, des situations très différentes.

Histogramme bidimensionnel

Il s'agit de l'extension de notion d'histogramme à la description de 2 variables.

Estimation de la densité

De la même façon, l'estimation de la fonction de densité d'une variable aléatoire peut être étendue à celui d'un vecteur aléatoire du plan. La fonction de densité estimée est alors une fonction réelle de 2 variables et son graphe une surface. Pour un échantillon $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où les $\boldsymbol{x}_i \in \mathbb{R}^2$, l'estimateur s'écrit

$$\widehat{f}(x) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right),$$

où K est une fonction de densité dans $\mathbb{R}^2$ centrée à l'origine qui pourra être par exemple le noyau gaussien

$$K(\boldsymbol{x}) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right).$$

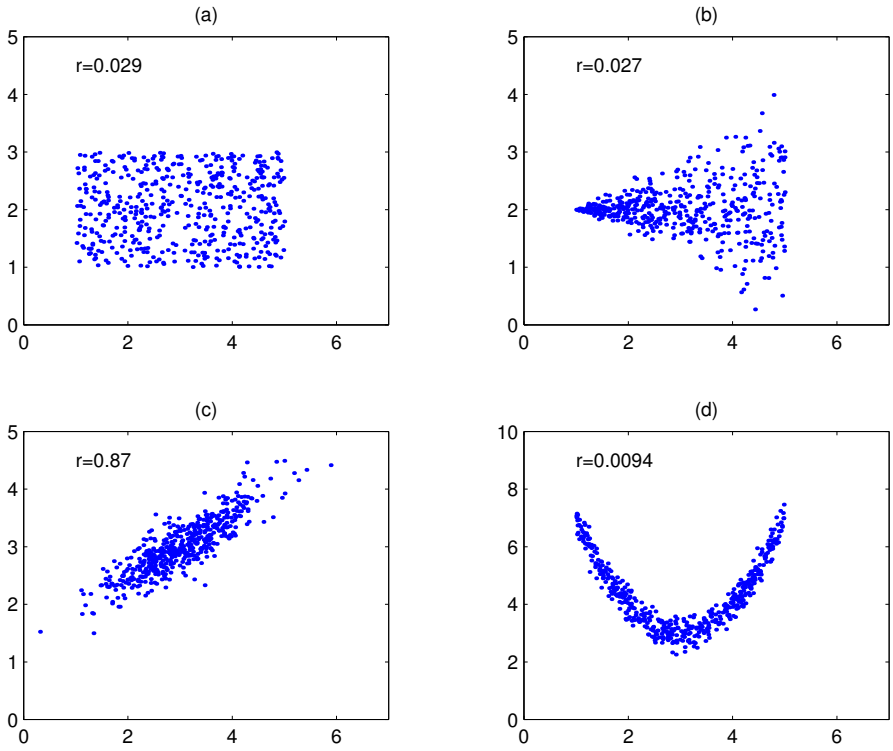


FIGURE 3.6 – Exemples de corrélations

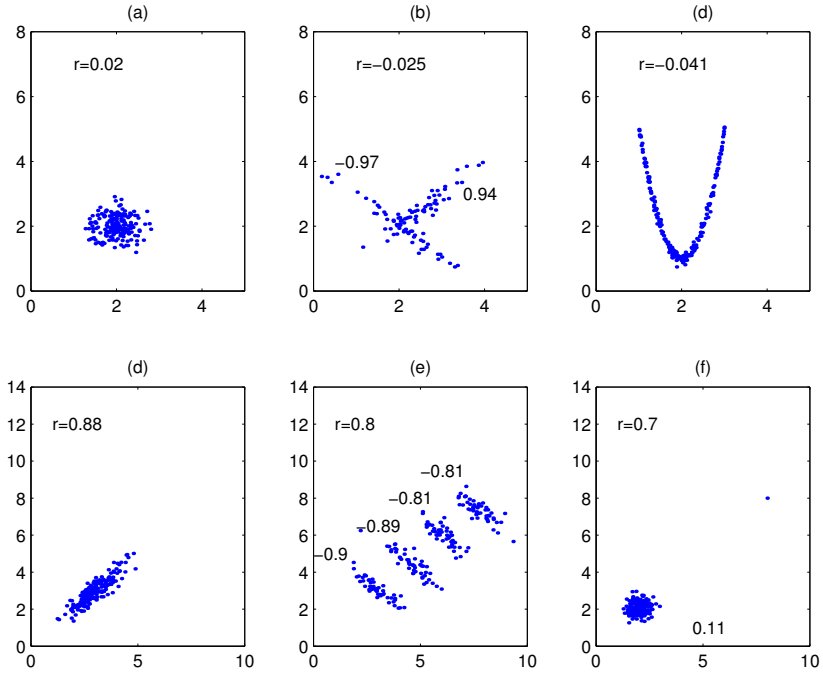


FIGURE 3.7 – Exemples de corrélations

1.3 Description multidimensionnelle

Cette fois, l'objectif est de décrire les liens pouvant exister entre l'ensemble de toutes les variables du tableau X.

Matrice de covariance et de corrélation

Lorsque l'on a plus de deux variables, on peut utiliser la matrice de covariance V de l'échantillon qui regroupent l'ensemble des covariances pour tous les couples de variables :

$$V = (s_{jj'})_{j,j'=1,\dots,p} = \frac{1}{n}(X - \mathbb{1}_n \overline{\boldsymbol{x}}^T)^T (X - \mathbb{1}_n \overline{\boldsymbol{x}}^T) = \frac{1}{n} Y^T Y$$

où $\mathbb{1}_n$ est la matrice de dimension $(n, 1)$ remplie de 1 et Y est la matrice centrée associée à X, et la matrice de corrélation R de l'échantillon qui regroupent l'ensemble des corrélations pour tous les couples de variables :

$$R = (r_{jj'})_{j,j'=1,\dots,p} = D_{1/s_j} V D_{1/s_j}$$

où D_{1/s_j} est la matrice diagonale définie par les valeurs $(1/s_1, \dots, 1/s_p)$.

Remarquons que l'on utilise aussi la matrice de covariance empirique

$$V^* = (s_{jj'})_{j,j'=1,\dots,p} = \frac{1}{n-1}(X - 1_n \overline{\boldsymbol{x}}^T)^T (X - 1_n \overline{\boldsymbol{x}}^T) = \frac{1}{n-1} Y^T Y.$$

La table 3.1 représente les matrices de variance et de corrélation obtenues avec les données `Iris`.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.686	-0.042	1.27	0.52
Sepal.Width	-0.042	0.190	-0.33	-0.12
Petal.Length	1.274	-0.330	3.12	1.30
Petal.Width	0.516	-0.122	1.30	0.58

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.00	-0.12	0.87	0.82
Sepal.Width	-0.12	1.00	-0.43	-0.37
Petal.Length	0.87	-0.43	1.00	0.96
Petal.Width	0.82	-0.37	0.96	1.00

TABLE 3.1 – Matrice de variance et matrice de corrélation des données `Iris`

Multiplot ou graphique matriciel

Lorsqu'on dispose de plus de 2 variables, il est possible, si le nombre de variables n'est pas trop grand, de représenter simultanément tous les plans correspondant aux différents couples de variables dans un seul graphique souvent appelé multiplot. La figure 3.8 représente le graphique obtenu avec les données `Iris`.

Description 3-D

Enfin, terminons ce chapitre en citant quelques descriptions utilisant 3 dimensions, c'est-à-dire permettant de représenter exactement 3 variables. Pour représenter le nuage de points défini par 3 variables, c'est-à-dire un nuage de points dans l'espace, certains logiciels offrent un outil interactif, souvent appelé « Brushing », permettant par rotations et homothéties de se « promener » dans ce nuage et donc de l'analyser. Enfin, des méthodes plus simples permettent aussi de représenter 3 variables. Par exemple, les données peuvent être représentées dans un plan par des cercles dont la position des centres sera définie par les 2 premières variables et le rayon par la troisième.

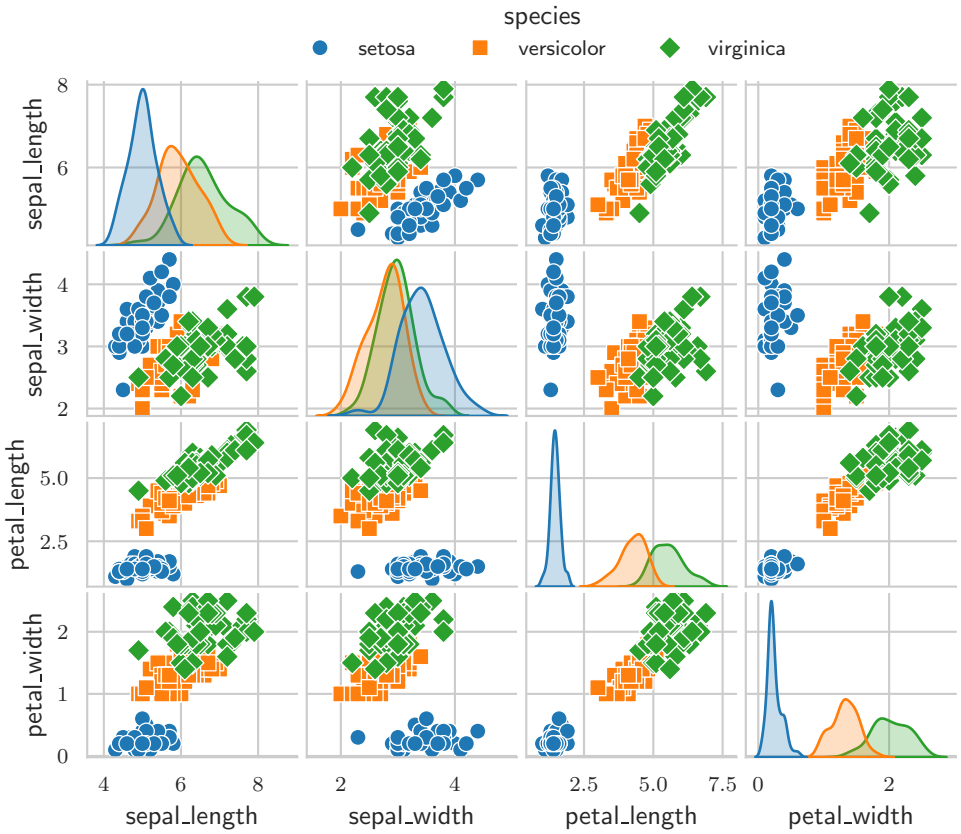


FIGURE 3.8 – Graphique matriciel

Quelques difficultés

Fléau de la dimension Dans les espaces de grande dimension, les calculs sont très similaires à ceux effectués dans le plan mais en réalité, il est difficile de généraliser et de se faire une idée claire de tels espaces. Ce problème est connu sous le nom de « fléau de la dimension », *curse of dimensionality* en anglais, et correspond au fait que les espaces de grande dimension sont **vides** : par exemple la sphère de rayon 0.74 de $\mathbb{R}^{10}$ ne contient que 5% des points d'un cube encadrant cette sphère et parallèle aux axes que l'on aurait remplie uniformément ; autrement dit, tous les points sont proches de la surface de la sphère. Il est donc difficile de généraliser certains outils comme les histogrammes.

Problème lié à la projection Un autre aspect porte sur l'interprétation des projections qui dans de tels espaces peut être quelquefois délicate. Par exemple, la figure 3.9 correspond à la projection de points répartis au hasard dans 15 plans parallèles ; la projection de droite correspond à un plan orthogonal aux 15 plans parallèles contenant tous les points alors que la projection de gauche correspond à un plan formant un angle de 5 degrés avec le plan de la projection précédente.

2 Descriptions des variables qualitatives

2.1 Description monodimensionnelle

Une distribution de n observations associée à une variable qualitative peut être présentée sous forme d'un tableau de fréquences où figure pour chaque modalité ξ_k, le nombre n_k (appelé effectif ou fréquence) d'observations ayant la valeur ξ_k, la fréquence relative

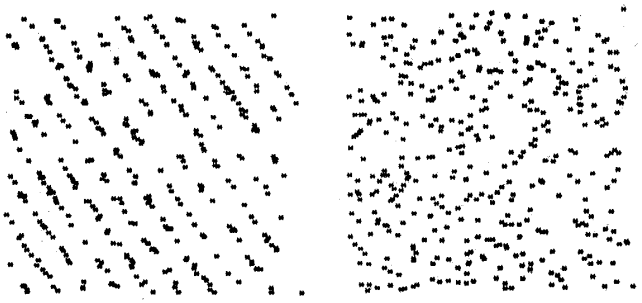


FIGURE 3.9 – Deux projections voisines

$f_k = n_k/n$ correspondante.

Cette information peut être représentée graphiquement sous forme d'un *diagramme en bâtons* dans lequel est associée à chaque modalité une barre de longueur proportionnelle à sa fréquence dans l'échantillon. L'exemple de la figure 3.10 représente les descriptions ainsi obtenues de la variable `Species` pour le sous-échantillon des iris dont la longueur du sépale est supérieure à 5.

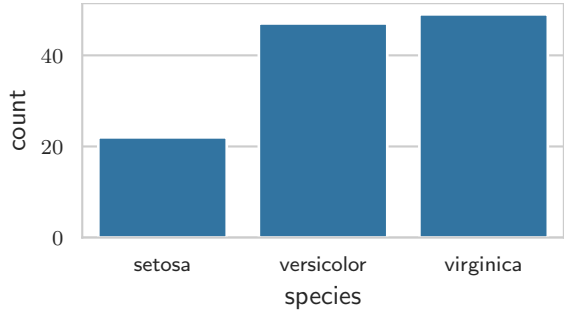


FIGURE 3.10 – Diagramme en bâtons de la variable `Species` pour les données `Iris` dont la longueur du sépale est supérieure à 5

2.2 Description bidimensionnelle : tableaux de contingence

Étant données deux variables qualitatives I et J, le tableau de contingence (I, J) associe à chaque couple de modalités (i, j) le nombre n_{ij} de fois où les 2 modalités sont présentes simultanément. La table 3.2 fournit un exemple d'un tableau de contingence obtenu en croisant deux variables ayant respectivement 2 et 3 modalités.

TABLE 3.2 – Deux variables qualitatives et leur tableau de contingence

I	J
1	1
2	3
1	2
2	2
2	3
2	1

(a) Deux variables qualitatives

	1	2	3
1	1	1	0
2	1	1	2

(b) Tableau de contingence

	prof	*tran*	*mena*	*enfa*	*cour*	*toil*	*repa*	*somm*	*tele*	*lois*
haus	610	140	60	10	120	95	115	760	175	315
faus	475	90	250	30	140	120	100	775	115	305
fnau	10	0	495	110	170	110	130	785	160	430
hmus	615	141	65	10	115	90	115	765	180	305
fmus	179	29	421	87	161	112	119	776	143	373
hcus	585	115	50	0	150	105	100	760	150	385
fcus	482	94	196	18	141	130	96	775	132	336
hawe	652	100	95	7	57	85	150	807	115	330
fawe	510	70	307	30	80	95	142	815	87	262
fnaw	20	7	567	87	112	90	180	842	125	367
hmwe	655	97	97	10	52	85	152	807	122	320
fmwe	168	22	529	69	102	83	174	825	119	392
hcwe	642	105	72	0	62	77	140	812	100	387
fcwe	389	34	262	14	92	97	147	848	84	392
hayo	650	140	120	15	85	90	105	760	70	365
fayo	560	105	375	45	90	90	95	745	60	235
fnay	10	10	710	55	145	85	130	815	60	380
hmyo	650	145	112	15	85	90	105	760	80	357
fmyo	260	52	576	59	116	85	117	775	65	295
hcyo	615	125	95	0	115	90	85	760	40	475
fcyo	413	89	318	23	112	96	102	774	45	409
haes	650	142	122	22	76	94	100	764	96	334
faes	578	106	338	42	106	94	52	752	64	228
fnae	24	8	594	72	158	92	128	840	86	398
hmes	652	133	134	22	68	94	102	762	122	310
fmes	434	77	431	60	117	88	105	770	73	229
hces	627	148	68	0	88	92	86	770	58	463
fces	433	86	296	21	128	102	94	758	58	379

TABLE 3.3 – Tableau « budgets–temps »

Remarques

— On a la propriété $\sum_{i,j} n_{ij} = \text{card}(\Omega)$.

— Dans ce type de tableau, les deux ensembles I et J mis en correspondance, tout en restant différents, sont de même nature, contrairement aux tableaux individus-variables.

Plus généralement, tout tableau regroupant le résultat d'un décompte, de façon à ce que l'addition du contenu des cellules d'une ligne ou d'une colonne ait un sens, peut aussi être considéré comme un tableau de contingence. Ainsi le tableau 3.3 qui regroupe le nombre d'heures passées à pratiquer une des 10 classes d'activité (profession, transport, ménage, enfants, courses, toilette, repas, sommeil, television, loisirs) par un ensemble de 28 types de population caractérisée par le sexe (h ou f), le pays (USA, Ouest, Est ou Yougoslavie), l'activité professionnelle (actif ou non actif) et le mariage (marié ou célibataire) durant une période donnée, peut être vu comme un tableau de contingence.

Chapitre 4

Représentation euclidienne des données

1 Les données

Soit X un tableau de dimension (n, p) correspondant à la mesure de p variables quantitatives effectuées sur n individus. Rappelons que l'on peut associer à chaque individu i un vecteur $\boldsymbol{x}_i$ de dimension p et à chaque variable j un vecteur $\mathbf{x}_j$ de dimension n. On suppose en outre qu'à chaque individu est associé une pondération $p_i > 0$ vérifiant $\sum_i p_i = 1$ et à chaque variable une pondération $q_j > 0$. On notera D_p la matrice diagonale $\mathrm{diag}(p_1, \ldots, p_n)$ et M la matrice diagonale $\mathrm{diag}(q_1, \ldots, q_j)$. Les données sont ainsi caractérisées par le triplet (X, M, D_p).

Dans le cas le plus simple, on prendra les pondérations $p_i = 1/n$ pour tout i et $q_j = 1$ pour tout j, c'est-à-dire $D_p = \frac{1}{n} I_n$ et $M = I_p$. L'utilisation de pondérations plus générales p_i et q_j permettra d'étendre sans difficulté les résultats de l'analyse en composantes principales à d'autres situations telle que l'analyse des correspondances.

2 Nuages associés

On peut alors définir le *nuage des individus*

$$\mathcal{N}(\Omega) = \{(\boldsymbol{x}_i, p_i), i = 1, \ldots, n\},$$

inclus dans $\mathbb{R}^p$ muni de la métrique euclidienne M, c'est-à-dire définie par le produit scalaire

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T M \boldsymbol{y}.$$

On a alors

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T M \boldsymbol{x}} \quad \text{et} \quad d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{y} - \boldsymbol{x})^T M (\boldsymbol{y} - \boldsymbol{x})}.$$

Cette représentation généralise la notion de graphe de dispersion (*scatter plot*) utilisé pour visualiser les données. Dans cette représentation, les points correspondent aux individus et les axes correspondent aux variables.

De façon symétrique, on peut définir le *nuage des variables* :

$$\mathcal{N}(V) = \{(\mathbf{x}_j, q_j), j = 1, \ldots, p\},$$

inclus dans $\mathbb{R}^n$ muni de la métrique euclidienne D_p souvent appelée « métrique des poids ». Cette fois, dans cette représentation les points correspondent aux variables et les axes correspondent aux individus.

3 Tableau X centré en colonne

Pour simplifier les calculs, on supposera dans la suite que le nuage des individus est centré, c'est-à-dire que son centre de gravité est à l'origine ou encore que la moyenne de chaque variable est nulle ; on dit alors que le tableau est *centré en colonne*. Si ce n'est pas le cas, il est facile de s'y ramener en soustrayant à chaque colonne sa moyenne. Centrer en colonne revient, dans l'espace des individus, à prendre comme nouvelle origine le centre de gravité. On peut alors montrer que l'inertie du nuage des individus et l'inertie portée par un axe $\Delta_{\mathbf{u}}$ de vecteur unitaire $\mathbf{u}$ s'écrivent respectivement

$$\mathcal{I} = \mathrm{Tr}\,(X^T D_p X M) \quad \text{et} \quad \mathcal{I}_{\Delta_{\mathbf{u}}^{\perp}} = \mathbf{u}^T M X^T D_p X M \mathbf{u}.$$

4 Interprétation statistique

Les représentations géométriques des données qui viennent d'être définies permettent de *visualiser* un certain nombre de propriétés statistiques. En voici quelques exemples.

4.1 Centre de gravité et moyenne

Le centre de gravité du nuage des individus, notion géométrique, a pour coordonnées les moyennes des p variables, notions statistiques.

4.2 Inertie et variance

Si on suppose que $D_p = \frac{1}{n} I_n$, le produit scalaire et la norme définis dans l'espace des variables s'expriment alors respectivement comme la covariance et la variance. On peut en effet écrire les relations suivantes :

$$\mathrm{Cov}(\mathbf{x}_j, \mathbf{x}_{j'}) = \frac{1}{n} \sum_i x_{ij} x_{ij'} = \sum_i p_i x_{ij} x_{ij'} = \mathbf{x}_j^T D_p \mathbf{x}_{j'} = \langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle_{D_p}$$

$$\mathrm{Var}(\mathbf{x}_j) = \mathrm{Cov}(\mathbf{x}_j, \mathbf{x}_{j'}) = \langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle_{D_p} = \|\boldsymbol{x}_j\|^2_{D_p}.$$

Dans ce cas, la matrice de variance V s'écrit donc $X^T D_p X$ et l'inertie du nuage $\mathcal{I} = \mathrm{Tr}\,(V)$.

En utilisant les résultats précédents, on a donc aussi

$$\mathrm{Cor}(\mathbf{x}_j, \mathbf{x}_{j'}) = \frac{\langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle_{D_p}}{\|\mathbf{x}_j\|_{D_p} \|\mathbf{x}_{j'}\|_{D_p}}$$

La corrélation s'interprète comme le cosinus de l'angle des deux vecteurs $\mathbf{x}_j$ et $\mathbf{x}_{j'}$ dans l'espace des variables $\mathbb{R}^n$ et l'orthogonalité de deux variables s'interpréte comme la non corrélation linéaire entre les deux variables. On peut aussi exprimer les relations précédentes en terme de métrique

$$d^2(\mathbf{x}_j, \mathbf{x}_{j'}) = \mathrm{Var}(\mathbf{x}_j) + \mathrm{Var}(\mathbf{x}_{j'}) - 2\,\mathrm{Cov}(\mathbf{x}_j, \mathbf{x}_{j'})$$

$$d^2(0, \mathbf{x}_j) = \mathrm{Var}(\mathbf{x}_j).$$

4.3 Cercles des corrélations et variables normées

Si on suppose en outre que le tableau X, déjà centré, est réduit, c'est-à-dire si la variance de chaque variable $\mathbf{x}_j$ est égale à 1, on a alors

$$d^2(0, \mathbf{x}_j) = \|\mathbf{x}_j\|^2_{D_p} = \mathrm{Var}\,(\mathbf{x}_j) = 1,$$

et dans l'espace des variables $\mathbb{R}^n$, les variables sont donc toutes situées sur une hypersphère de centre 0 et de rayon 1, appelée « cercle des corrélations ». Par ailleurs, on obtient les relations suivantes :

$$\mathrm{Cor}(\mathbf{x}_j, \mathbf{x}_{j'}) = \langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle_{D_p}, \quad \text{et} \quad d^2(\mathbf{x}_j, \mathbf{x}_{j'}) = 2(1 - \mathrm{Cor}(\mathbf{x}_j, \mathbf{x}_{j'})).$$

Remarquons que dans cette situation les matrices de covariance et de corrélation sont les mêmes ; l'expression $X^T D_p X$ représente donc aussi dans ce cas la matrice de corrélation.

Chapitre 5

L'analyse en composantes principales

1 Introduction

Les *méthodes factorielles* ont pour objectif de visualiser, et plus généralement, d'analyser des données multidimensionnelles, c'est-à-dire des données regroupant souvent un grand nombre de variables. La prise en compte simultanée de ces variables est un problème difficile ; heureusement, l'information apportée par ces variables est souvent redondante et toutes ces méthodes vont exploiter cette caractéristique pour tenter de remplacer les variables initiales par un nombre réduit de nouvelles variables sans perdre trop d'information. Remarquons que la construction de variables synthétiques, consistant à résumer plusieurs variables par une seule, est une démarche habituelle (moyenne à l'école, QI, répartition des hommes politiques sur un axe droite-gauche). Il y a mieux à faire. C'est ce qu'ont proposé les psychologues américains Spearman, Burt et Thurstone en résumant les résultats de nombreux tests psychologiques par un facteur général d'aptitude et un nombre très limité de facteurs spécifiques comme la mémoire ou l'intelligence.

Lorsque les variables sont toutes quantitatives, l'analyse en composantes principales (ACP) va chercher à résoudre ce problème en considérant que les nouvelles variables sont des combinaisons linéaires des variables initiales et, qu'en plus, elles doivent être non corrélées linéairement. Si l'on représente les données initiales à l'aide d'un nuage de points, on peut montrer que ce problème revient à chercher les droites, les plans et de manière plus générale les variétés linéaires proches du nuage initial. Nous utiliserons ce point de vue géométrique dans ce chapitre. Cette méthode a d'abord été développée par K.Pearson (1900) pour deux variables, puis par H. Hotelling (1933) qui l'a étendue à un nombre quelconque de variables. L'ouvrage de Jackson (1991) constitue un panorama très complet et assez récent de l'ACP.

Les méthodes factorielles, dont l'ACP est l'exemple le plus connu, varient suivant la forme des données mais utilisent toutes les mêmes bases mathématiques. Il faut les distinguer des méthodes regroupées sous le terme « factor analysis » par les anglo-saxons qui sont des méthodes de statistiques inférentielles s'appuyant sur un modèle statistique et qui sont assez peu utilisées en France. En dehors de l'ACP destinée aux tableaux de variables quantitatives, les principales méthodes factorielles sont l'analyse factorielle des correspondances (AFC) pour les tableaux de contingence, l'analyse des correspondances multiples (ACM) pour les tableaux de variables qualitatives, l'analyse factorielle d'un tableau de distances (AFTD) pour les tableaux de proximités et l'analyse factorielle discriminante qui permet de mettre en évidence les différences entre des individus issus de plusieurs classes.

Dans tout ce chapitre, on utilisera les représentations géométriques (nuage des individus

et nuage des variables) associées à un tableau de variables quantitatives décrites dans le chapitre précédent et on supposera que le tableau est centré en colonne.

2 Axes principaux d'inertie

2.1 Formulation mathématique

L'objectif est d'obtenir une représentation fidèle du nuage $\mathcal{N}(\Omega)$ de $\mathbb{R}^p$ en le projetant sur un espace de faible dimension. Pour ceci, on cherche à minimiser les « écarts » entre les points de $\mathcal{N}(\Omega)$ et leurs projections. Les espaces de représentation choisis sont les variétés linéaires (droite, plan,…). La formulation mathématique de l'ACP est alors la suivante : *Trouver la variété linéaire E_k de dimension k ($k < p$) tel que $\mathcal{I}_{E_k}$, l'inertie du nuage $\mathcal{N}(\Omega)$ par rapport à E_k, soit minimum.*

Rappelons que l'on a

$$\mathcal{I}_{E_k} = \frac{1}{n} \sum_i d^2(\boldsymbol{x}_i, E_k).$$

En utilisant la version 2 du théorème de Huygens rappelé à la page 167, on peut en déduire que l'espace E_k minimisant $\mathcal{I}_{E_k}$ contient nécessairement le centre de gravité du nuage $\mathcal{N}(\Omega)$, c'est-à-dire ici l'origine O puisqu'on a supposé le tableau X centré en colonne. E_k est un donc un sous-espace vectoriel. D'autre part, nous savons que dans ce cas, l'inertie totale du nuage $\mathcal{I}$ se décompose en une somme $\mathcal{I}_{E_k} + \mathcal{I}_{E_k^\perp}$ où $\mathcal{I}_{E_k^\perp}$ est l'inertie expliquée par E_k. En conséquence, le problème peut s'écrire maintenant : *Trouver le sous-espace vectoriel E_k de dimension k ($k < p$) tel que l'inertie expliquée $\mathcal{I}_{E_k^\perp}$ par E_k soit maximum.*

2.2 Résultats préalables

Théorème 1 (Emboîtement des solutions)**.** *Si E_{k-1} est un sous-espace vectoriel optimal de dimension $k-1$, alors la recherche d'un sous-espace optimal de dimension k peut se faire parmi l'ensemble des sous-espaces vectoriels de dimension k contenant E_{k-1}.*

Preuve. Soit F_k un sous-espace quelconque de dimension k de $\mathbb{R}^p$.
Le sous-espace $F_k \cap E_{k-1}^\perp$ ne peut être réduit au vecteur nul sinon le sous-espace $F_k \oplus E_{k-1}^\perp$ serait de dimension $p+1$. Il existe donc $\mathbf{v} \neq 0 \in F_k \cap E_{k-1}^\perp$. Soit $\Delta_\mathbf{v}$ l'axe correspondant et G l'espace supplémentaire M-orthogonal à $\Delta_\mathbf{v}$ dans F_k (on a donc $F_k = G \oplus \Delta_\mathbf{v}$).

Si on note $H = E_{k-1} \oplus \Delta_\mathbf{v}$, on a

$$\mathcal{I}_{F_k^\perp} = \mathcal{I}_{G^\perp} + \mathcal{I}_{\Delta_\mathbf{v}^\perp} \quad \text{car} \quad G \perp \Delta_\mathbf{v}$$

$$\mathcal{I}_{H^\perp} = \mathcal{I}_{E_{k-1}^\perp} + \mathcal{I}_{\Delta_\mathbf{v}^\perp} \quad \text{car} \quad E_{k-1} \perp \Delta_\mathbf{v}.$$

Mais par hypothèse, E_{k-1} est optimal. On a donc :

$$\mathcal{I}_{E_{k-1}^\perp} \geq \mathcal{I}_{G^\perp} \Rightarrow \mathcal{I}_{H^\perp} \geq \mathcal{I}_{F_k^\perp}$$

On peut donc restreindre la recherche d'un sous-espace optimal aux sous-espaces contenant E_{k-1}. □

Remarquons qu'on n'affirme pas dans ce théorème l'existence d'espaces optimaux.

Théorème 2. *La recherche d'un sous-espace vectoriel optimal E de dimension k contenant un sous-espace F de dimension $k-1$ est équivalente à la recherche d'un axe $\Delta_\mathbf{v}$ M-orthogonal à F et maximisant $\mathcal{I}_{\Delta_\mathbf{v}^\perp}$.*

Preuve. On a une décomposition $E = F \oplus \Delta_\mathbf{v}$ avec $\Delta_\mathbf{v} \perp F$. On a donc $\mathcal{I}_{E^\perp} = \mathcal{I}_{F^\perp} + \mathcal{I}_{\Delta_\mathbf{v}^\perp}$.
Maximiser $\mathcal{I}_{E^\perp}$ est donc équivalent à maximiser $\mathcal{I}_{\Delta_\mathbf{v}^\perp}$. □

2.3 Résolution du problème

On suppose dans la suite que les vecteurs $\mathbf{u}_j$ sont unitaires, c'est-à-dire que $\|\mathbf{u}_j\|^2 = \langle \mathbf{u}_j, \mathbf{u}_j \rangle_M = \mathbf{u}_j^T M \mathbf{u}_j = 1$. En outre, on sait que pour tout vecteur unitaire $\mathbf{u}$, $\mathcal{I}_{\Delta_{\mathbf{u}}^\perp}$ est égale à $\langle \mathbf{u}, VM\mathbf{u} \rangle_M = \mathbf{u}^T MVM\mathbf{u}$, où V est la matrice de variance empirique associée à X.

À partir des deux théorèmes précédents, il est alors facile de voir que le problème de l'ACP se ramène au problème suivant :

- rechercher un axe $\Delta_{\mathbf{u}_1}$ maximisant l'inertie $\mathcal{I}_{\Delta_{\mathbf{u}_1}^\perp} = \langle \mathbf{u}_1, VM\mathbf{u}_1 \rangle_M$, on note $E_1 = \Delta_{\mathbf{u}_1}$;
- rechercher un axe $\Delta_{\mathbf{u}_2}$, M-orthogonal à E_1 maximisant l'inertie $\mathcal{I}_{\Delta_{\mathbf{u}_2}^\perp} = \langle \mathbf{u}_2, VM\mathbf{u}_2 \rangle_M$, on note $E_2 = E_1 \oplus \Delta_{\mathbf{u}_2}$;
- …
- rechercher un axe $\Delta_{\mathbf{u}_k}$, M-orthogonal à E_{k-1} maximisant l'inertie $\mathcal{I}_{\Delta_{\mathbf{u}_k}^\perp} = \langle \mathbf{u}_k, VM\mathbf{u}_k \rangle_M$, on note $E_k = E_{k-1} \oplus \Delta_{\mathbf{u}_k}$.

En posant $B = VM$ et $Q = M$, le théorème 16 de décomposition d'une matrice énoncé dans l'annexe B fournit une réponse à notre problème : les vecteurs propres normés de la matrice VM ordonnés suivant les valeurs propres décroissantes fournissent les axes $\Delta_{\mathbf{u}_1}, \ldots, \Delta_{\mathbf{u}_k}$, appelés *axes factoriels* ou encore *axes principaux d'inertie* et les inerties $\mathcal{I}_{\Delta_{\mathbf{u}_k}^\perp}$ portées ou expliquées par ces axes sont égales aux valeurs propres λ_k.

L'espace $E_k = \Delta_{\mathbf{u}_1} \oplus \cdots \oplus \Delta_{\mathbf{u}_k}$ est donc la solution du problème et on obtient du même coup toutes les solutions pour les dimensions inférieures à k.

Par ailleurs, en utilisant les propriétés de la décomposition en valeurs propres et vecteurs propres, on obtient les relations suivantes

$$VM\mathbf{u}_k = \lambda_k \mathbf{u}_k, \tag{5.1}$$

$$\mathcal{I}_{\Delta_{\mathbf{u}_k}^\perp} = \lambda_k.$$

En notation matricielle, si on note $U = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ la matrice des p vecteurs propres normés rangés en colonne et L la matrice diagonale des valeurs propres associées (et donc rangées dans le même ordre), on a

$$UMU^T = U^T MU = MUU^T = I_p, \tag{5.2}$$

qui traduit le fait que les vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_p$ forment une base M-orthonormale de $\mathbb{R}^p$ et

$$VMU = UL, \tag{5.3}$$

qui est la forme condensée des relations (5.1).

2.4 Résultats pratiques

Si $\mathbf{u}_1, \ldots, \mathbf{u}_p$ sont les vecteurs propres normés ordonnés suivant les valeurs propres décroissantes de la matrice VM, les solutions pour les différentes valeurs de k sont les suivantes :

- $k = 1 : E_1 = \Delta_{\mathbf{u}_1}$;
- $k = 2 : E_2 = E_1 \oplus \Delta_{\mathbf{u}_2}$;
- …
- $k : E_k = E_{k-1} \oplus \Delta_{\mathbf{u}_k}$.

On a en outre $\mathcal{I}_{\Delta_{\mathbf{u}_k}^\perp} = \lambda_k$ pour tout k.

2.5 Inerties expliquées

Proposition 3. *L'inertie expliquée par le sous-espace E_k est la somme des k plus grandes valeurs propres.*

$$\mathcal{I}_{E_k^\perp} = \lambda_1 + \ldots + \lambda_k.$$

Preuve. Les vecteurs propres $\mathbf{u}_\alpha$ sont M-orthogonaux (car la matrice V est M-symétrique). L'espace E_k se décompose donc en une somme directe de sous-espaces M-orthogonaux $\Delta_{\mathbf{u}_\alpha}$, on sait alors que

$$\mathcal{I}_{E_k^\perp} = \sum_{\alpha=1}^{k} \mathcal{I}_{\Delta \mathbf{u}_\alpha^\perp}$$

Puisque $\mathcal{I}_{\Delta \mathbf{u}_\alpha^\perp} = \lambda_\alpha$, le résultat est démontré. □

Remarque 1. *En prenant $k = p$, on retrouve $\mathcal{I} = \operatorname{Tr} VM$. De plus, si r est le rang de la matrice X ($r \leq \min(p, n)$), on a*

$$\lambda_1, \ldots, \lambda_r > 0 \quad et \quad \lambda_{r+1}, \ldots, \lambda_p = 0,$$

et par suite

$$\mathcal{I}_{E_r^\perp} = \mathcal{I}.$$

Finalement, l'inertie du nuage est totalement expliquée par le sous-espace vectoriel E_r ce qui veut dire que le nuage est contenu dans E_r engendré par les r premiers axes factoriels.

2.6 Choix du nombre d'axes à retenir

Pour choisir le nombre d'axes à retenir, on s'appuie généralement sur les *pourcentages d'inertie expliquée* par les différents sous-espaces E_α :

- % d'inertie expliquée par $E_1 = \frac{\lambda_1}{\sum_{\alpha=1}^{p} \lambda_\alpha} \times 100 = \frac{\lambda_1}{\operatorname{Tr}(VM)} \times 100$;
- % d'inertie expliquée par $E_2 = \frac{\lambda_1+\lambda_2}{\sum_{\alpha=1}^{p} \lambda_\alpha} \times 100 = \frac{\lambda_1+\lambda_2}{\operatorname{Tr}(VM)} \times 100$;
- $\vdots$
- % d'inertie expliquée par $E_k = \frac{\lambda_1+\lambda_2+\ldots\lambda_k}{\sum_{\alpha=1}^{p} \lambda_\alpha} \times 100 = \frac{\lambda_1+\lambda_2+\ldots\lambda_k}{\operatorname{Tr}(VM)} \times 100$.

3 Composantes principales

3.1 Définition

Rappelons que le problème de départ était d'obtenir une représentation du nuage $\mathcal{N}(\Omega)$ dans des espaces de petite dimension. On connait maintenant les axes définissant ces espaces. Pour pouvoir obtenir les différentes représentations, il suffit de déterminer les coordonnées de la projection de tous les points du nuage sur chaque axe factoriel. On notera $c_{1\alpha}, \ldots, c_{n\alpha}$ les n coordonnées ainsi obtenues avec l'axe α, $\mathbf{c}_\alpha$ le vecteur $(c_{1\alpha}, \ldots, c_{n\alpha})^T$, appelé α^{e} composante principale et C la matrice obtenue en rangeant en colonne les vecteurs $\mathbf{c}_\alpha$. On peut alors obtenir la projection du nuage $\mathcal{N}(\Omega)$ dans un plan factoriel quelconque $(\mathbf{u}_\alpha, \mathbf{u}_\beta)$ grâce aux composantes principales $\mathbf{c}_\alpha$ et $\mathbf{c}_\beta$. Par exemple, la représentation dans le premier plan factoriel est obtenue grâce à $\mathbf{c}_1$ et $\mathbf{c}_2$.

Pour les indices α strictement supérieurs au rang r, les valeurs propres λ_α sont nulles ce qui entraîne que les inerties expliquées $\mathcal{I}_{\Delta \mathbf{u}_\alpha^\perp}$ et, en conséquence, les composantes principales $\mathbf{c}_\alpha$ sont aussi nulles.

Enfin, en exprimant l'inertie expliquée par l'axe α dans la relation $\lambda_\alpha = \mathcal{I}_{\Delta \mathbf{u}_\alpha^\perp}$, on obtient la relation

$$\lambda_\alpha = \frac{1}{n} \sum_{i \in \Omega} (c_{i\alpha})^2.$$

3.2 Calcul des composantes principales

Proposition 4. *Les composantes principales vérifient les relations*

$$\mathbf{c}_\alpha = XM\mathbf{u}_\alpha \qquad \forall\alpha \tag{5.4}$$

qui s'expriment matriciellement par la relation

$$C = XMU. \tag{5.5}$$

Preuve. Les axes principaux $\mathbf{u}_1, \ldots, \mathbf{u}_p$ forment une base orthonormée, il suffit donc de projeter les $\boldsymbol{x}_i$ sur les vecteurs de base :

$$\begin{aligned} c_{i\alpha} = \langle \boldsymbol{x}_i, \mathbf{u}_\alpha \rangle_M = \boldsymbol{x}_i^T M \mathbf{u}_\alpha \qquad \forall i, \alpha \\ \mathbf{c}_\alpha = XM\mathbf{u}_\alpha \qquad \forall\alpha \\ C = XMU. \end{aligned}$$

□

On peut aussi démontrer cette proposition de la manière suivante.

Preuve. Les composantes principales peuvent être obtenues aussi par changement de base. Si on note $\mathbf{c}_i$ les vecteurs lignes transposés de C, on obtient

$$\begin{aligned} \boldsymbol{x}_i = U\mathbf{c}_i \qquad \forall i \\ U^T M \boldsymbol{x}_i = U^T MU\mathbf{c}_i = \mathbf{c}_i \text{ car } U^T MU = I \qquad \forall i \\ \mathbf{c}_i^T = \boldsymbol{x}_i^T MU \qquad \forall i \end{aligned}$$

et donc

$$C = XMU.$$

□

3.3 Composantes principales : nouvelles variables

Une composante principale associe à chaque individu $\boldsymbol{x}_i$ de Ω un nombre réel. On peut donc la considérer comme une nouvelle variable. Comme les variables initiales $\mathbf{x}_j$, cette variable appartient à l'espace $\mathbb{R}^n$. Quelques propriétés de ces nouvelles variables peuvent alors être établies :

Proposition 5. *Les composantes principales sont des combinaisons linéaires des variables* $\mathbf{x}_j$.

Preuve. On a $\mathbf{c}_\alpha = XM\mathbf{u}_\alpha = X(M\mathbf{u}_\alpha) = \sum_{j=1}^p a_{\alpha j}\mathbf{x}_j$ si on note $\mathbf{a}_\alpha$ le vecteur $M\mathbf{u}_\alpha$. □

Proposition 6. *Les composantes principales* $\mathbf{c}_\alpha$ *sont centrées, de variance* λ_α *et non corrélées 2 à 2.*

Preuve. Une combinaison linéaire de variables centrées est centrée. D'autre part, on a

$$\begin{aligned} \mathrm{Cov}(\mathbf{c}_\alpha, \mathbf{c}_\beta) &= \langle \mathbf{c}_\alpha, \mathbf{c}_\beta \rangle_{D_p} \\ &= \mathbf{c}_\alpha^T D_p \mathbf{c}_\beta \\ &= \mathbf{u}_\alpha^T M X^T D_p X M \mathbf{u}_\beta \\ &= \mathbf{u}_\alpha^T M (X^T D_p X) M \mathbf{u}_\beta \\ &= \mathbf{u}_\alpha^T M V M \mathbf{u}_\beta \\ &= \mathbf{u}_\alpha^T M (V M \mathbf{u}_\beta) \\ &= \lambda_\beta \mathbf{u}_\alpha^T M \mathbf{u}_\beta \\ &= \lambda_\beta \langle \mathbf{u}_\alpha, \mathbf{u}_\beta \rangle_M. \end{aligned}$$

On en déduit

$$\begin{cases} \mathrm{Var}(\mathbf{c}_\alpha) = \lambda_\alpha & \text{si } \alpha = \beta \\ \mathrm{Cov}(\mathbf{c}_\alpha, \mathbf{c}_\beta) = 0 & \text{si } \alpha \neq \beta. \end{cases}$$

□

Dans la nouvelle base, la matrice de variance est donc diagonale : l'ACP revient à diagonaliser la matrice de variance. On peut ainsi poser le problème de l'ACP de manière différente : trouver k nouvelles variables, combinaisons linéaires normées des p variables centrées initiales, non corrélées deux à deux et de variance maximum.

Proposition 7. *Les composantes principales $\mathbf{c}_\alpha$ sont vecteurs propres de la matrice WD_p associées aux valeurs propres λ_α où*

$$W = XMX^T,$$

est la matrice des produits scalaires associés aux vecteurs individus $\boldsymbol{x}_i$.

Preuve. En effet, on a successivement :

$$\begin{aligned} WD_p\mathbf{c}_\alpha &= XMX^TD_p\mathbf{c}_\alpha \\ &= XMX^TD_pXM\mathbf{u}_\alpha \\ &= XMVM\mathbf{u}_\alpha \\ &= \lambda_\alpha XM\mathbf{u}_\alpha \\ &= \lambda_\alpha\mathbf{c}_\alpha, \end{aligned}$$

où on utilise le fait que $\mathbf{c}_\alpha = XM\mathbf{u}_\alpha$ par la proposition 4 et $VM\mathbf{u}_\alpha = \lambda_\alpha\mathbf{u}_\alpha$ d'après 5.1. □

Si on avait posé directement le problème en terme de recherche de variables, nous aurions obtenu ces variables comme vecteurs propres de la matrice WD_p.

4 Formule de reconstitution

La M-orthogonalité des axes principaux se traduit par $U^TMU = I_p$ ce qui se récrit $MUU^T = I_p$. En post-multipliant la relation $XMU = C$ par U^T, on obtient la relation

$$XMUU^T = CU^T, \qquad \text{c'est-à-dire} \qquad X = CU^T$$

souvent appelée formule de reconstitution.

Cette relation permet de tirer plusieurs conséquences :

— Tout d'abord, elle montre que le tableau de données initial X peut être reconstitué à partir des composantes principales et des axes principaux.

— Par ailleurs, en écrivant cette relation sous la forme

$$X = \sum_{\alpha=1}^{r} \mathbf{c}_\alpha\mathbf{u}_\alpha^T$$

on obtient une décomposition de la matrice X en une somme de matrices de rang 1.

— Si on se limite aux k ($k < r$) premiers termes, on obtient une approximation du tableau initial :

$$X \approx \tilde{X} = \sum_{\alpha=1}^{k} \mathbf{c}_\alpha\mathbf{u}_\alpha^T.$$

Cette propriété est quelquefois utilisée pour compresser les données lorsque l'on est prêt à perdre un peu d'information.

— Enfin, on peut en déduire

$$\boldsymbol{x}_i = \sum_{\alpha=1}^{r} c_{i\alpha}\mathbf{u}_\alpha \qquad \forall i,$$

ce qui montre que les vecteurs $\mathbf{x}_j$ sont des combinaisons linéaires des composantes principales $\mathbf{c}_\alpha$.

5 Qualité de la représentation

5.1 Qualité globale

La qualité globale de représentation de l'ensemble initial Ω sur le sous-espace E_k est mesurée par le pourcentage d'inertie pris en compte par E_k :

$$100 \cdot \frac{\lambda_1 + \cdots + \lambda_k}{\mathrm{Tr}(VM)}.$$

5.2 Contribution relative d'un axe à un individu

Sachant que l'inertie totale du nuage $\mathcal{N}(\Omega)$ est $\sum_{i=1}^{n} p_i \left\|\boldsymbol{x}_i\right\|^2$, la quantité $p_i\|\boldsymbol{x}_i\|^2$ représente la part d'inertie apportée par chaque individu i. Après projection sur l'axe $\mathbf{u}_\alpha$, l'inertie restante est donc $p_i c_{i\alpha}^2$. Chacun des termes $p_i c_{i\alpha}^2$ représente donc la part de l'inertie initiale $p_i\|\boldsymbol{x}_i\|^2$ qu'apportait l'individu i, conservée par l'axe α. Le rapport de ces deux quantités est appelée *contribution relative* du α^{e} axe factoriel à l'individu i et elle est notée $COR(i,\alpha)$:

$$COR(i,\alpha) = \frac{c_{i\alpha}^2}{\|\boldsymbol{x}_i\|^2}.$$

Cette quantité représente aussi le carré du cosinus de l'angle formé par l'individu i et par le vecteur $\mathbf{u}_\alpha$. Si $COR(i,\alpha)$ est proche de 1, l'individu est bien représenté par cet axe, si $COR(i,\alpha)$ est au contraire proche de 0, l'individu est très mal représenté par cet axe.

On peut généraliser cette notion en passant d'un axe à un sous-espace E_k. On appelle contribution relative de l'espace vectoriel E_k la quantité :

$$QLT(i,k) = \frac{\sum_{\alpha=1}^{k} c_{i\alpha}^2}{\|\boldsymbol{x}_i\|^2} = \sum_{\alpha=1}^{k} COR(i,\alpha).$$

On a alors $QLT(i,p) = 1$.

5.3 Contribution relative d'un individu à un axe

En partant de la relation $\lambda_\alpha = \sum_{i=1}^{n} p_i c_{i\alpha}^2$, on peut décomposer λ_α, l'inertie conservée par l'axe $\mathbf{u}_\alpha$, selon les individus. On définit alors la contribution relative de l'individu i à l'axe α , notée $CTR(i,\alpha)$: c'est la part d'inertie du α^{e} axe pris en compte (ou expliquée) par l'individu i. Nous avons :

$$CTR(i,\alpha) = p_i \frac{c_{i\alpha}^2}{\lambda_\alpha}.$$

6 Représentation des variables

Dans l'espace des variables, les composantes principales normées $\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{c}_\alpha$ forment un système de vecteurs orthonormés (une base si $n \geq p$). Dans ce système, les coordonnées

des variables initiales normées sont alors simplement les corrélations. La représentation des p variables initiales dans ce système permet de visualiser les liens entre les variables initiales et les liens entre les composantes principales et les variables initiales. Cette représentation est utilisée pour permettre de donner une « interprétation » aux axes. Le calcul de ces coordonnées vérifie donc

$$\mathrm{Cor}(\alpha, j) = \mathrm{Cov}\left(\frac{1}{\sigma_j}\mathbf{x}_j, \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{c}_\alpha\right) = \frac{1}{\sigma_j}\frac{1}{\sqrt{\lambda_\alpha}}\mathbf{x}_j^T D_p \mathbf{c}_\alpha.$$

7 Éléments supplémentaires

Dans toute analyse factorielle, il est possible de projeter sur les sous-espaces factoriels des individus ou des variables n'ayant pas participé à l'analyse. Ces éléments sont appelés éléments illustratifs ou supplémentaires. Inversement les éléments de départ qui ont participé à l'analyse sont appelés *éléments actifs*.

7.1 Individu supplémentaire

Il faut lui appliquer la même transformation géométrique que celle qui a été appliquée à tous les individus initiaux. Rappelons que nous avons centré en colonne le tableau initial, c'est-à-dire ôté à chaque composante j d'un individu la moyenne de la variable j (cette transformation correspond à une translation dans l'espace $\mathbb{R}^p$). Si $\overline{x}_j$ est la moyenne de chaque variable, calculée uniquement sur les individus initiaux, il suffit d'enlever cette valeur à toutes les coordonnées de l'individu supplémentaire. Ainsi, il suffit de transformer l'individu supplémentaire $\boldsymbol{y}_s = (y_{s1}, \ldots, y_{sp})^T$ en $\boldsymbol{x}_s = (y_{s1} - \overline{x}_1, \ldots, y_{sp} - \overline{x}_p)^T$ et de le projeter sur les axes $\mathbf{u}_\alpha$. Les coordonnées sont ainsi obtenues avec la formule $\langle \boldsymbol{x}_s, \mathbf{u}_\alpha \rangle = \boldsymbol{x}_s^T M \mathbf{u}_\alpha$.

7.2 Variable supplémentaire

Cette fois, la transformation précédente, devient une projection dans $\mathbb{R}^n$. Il faut centrer la nouvelle variable $\mathbf{y}_s = (y_{1s}, \ldots, y_{ns})^T$. Par ailleurs, ce sont les variables normées que l'on représente dans cet espace, il faut donc aussi normer cette variable. Finalement, si on note $\overline{y} = \frac{1}{n}\sum_i y_{is}$ et $s = \sqrt{\frac{1}{n}\sum_i (y_{is} - \overline{y})^2}$ la moyenne et l'écart-type de cette variable, on peut obtenir la représentation de la variable supplémentaire sur les axes factoriels en projetant le vecteur $\mathbf{x}_s = \frac{1}{s}(y_{1s} - \overline{y}, \ldots, y_{ns} - \overline{y})^T$ sur les axes $\mathbf{v}_\alpha = \frac{\mathbf{c}_\alpha}{\sqrt{\lambda_\alpha}}$. Les coordonnées sont ainsi obtenues par la relation suivante

$$\begin{aligned}\langle \mathbf{x}_s, \mathbf{v}_\alpha \rangle_{D_p} &= \mathbf{x}_s^T D_p \frac{\mathbf{c}_\alpha}{\sqrt{\lambda_\alpha}} \\ &= \frac{1}{n\sqrt{\lambda_\alpha}} \mathbf{x}_s^T \mathbf{c}_\alpha,\end{aligned}$$

car $D_p = \frac{1}{n} I_n$. Si on note $x_\alpha = \frac{1}{n\sqrt{\lambda_\alpha}} \mathbf{x}_s^T \mathbf{c}_\alpha$ la coordonnée de la variable supplémentaire selon l'axe factoriel $\mathbf{c}_\alpha$, on a alors

$$\mathbf{x}_s = \sum_{\alpha=1}^{p} x_\alpha \mathbf{c}_\alpha.$$

7.3 Importance pratique des éléments supplémentaires

Les éléments supplémentaires permettent, par exemple, la représentation d'individus prenant des valeurs très différentes des autres (valeurs atypiques) et qui auraient pris une part trop prépondérante à la formation des axes s'ils avaient été actifs, la représentation

d'un groupe d'individus par leur centre de gravité et la représentation d'éléments de natures différentes des éléments initiaux (variables actives : notes scolaires et variables supplémentaires : notes de tests psychologiques ou encore individus actifs : malades et individus supplémentaires : personnes saines). Les éléments supplémentaires ne participant pas à la formation des axes factoriels, une situation intéressante de ces éléments par rapport aux axes (par exemple, une variable supplémentaire très corrélée à une composante principale) est très significative.

8 Un exemple d'ACP

Les données

Il s'agit du tableau de notes décrits dans le chapitre 3. Rappelons que ces données regroupent les notes obtenues par neuf élèves dans les matières mathématiques, sciences, français, latin et dessin :

	math	scie	fran	lati	d-m
jean	6.00	6.00	5.00	5.50	8
alin	8.00	8.00	8.00	8.00	9
anni	6.00	7.00	11.00	9.50	11
moni	14.50	14.50	15.50	15.00	8
didi	14.00	14.00	12.00	12.50	10
andr	11.00	10.00	5.50	7.00	13
pier	5.50	7.00	14.00	11.50	10
brig	13.00	12.50	8.50	9.50	12
evel	9.00	9.50	12.50	12.00	18

Dans la suite, effectuant l'ACP classique, nous prendrons $D_p = \frac{1}{n} I_n$ et $M = I_p$.

Centrage du tableau de données

Les moyennes des cinq variables sont respectivement 9.67, 9.83, 10.22, 10.05 et 11. Le tableau centré en colonne X est obtenu en soustrayant à chaque colonne la moyenne correspondante :

	math	scie	fran	lati	d-m
jean	-3.67	-3.83	-5.22	-4.56	-3.00
alin	-1.67	-1.83	-2.22	-2.06	-2.00
anni	-3.67	-2.83	0.78	-0.56	0.00
moni	4.83	4.67	5.28	4.94	-3.00
didi	4.33	4.17	1.78	2.44	-1.00
andr	1.33	0.17	-4.72	-3.06	2.00
pier	-4.17	-2.83	3.78	1.44	-1.00
brig	3.33	2.67	-1.72	-0.56	1.00
evel	-0.67	-0.33	2.28	1.94	7.00

On obtient ainsi un tableau dont la somme de chaque colonne est nulle.

Matrice de variance

$$V = X^T D_p X = \frac{1}{9} X^T X$$

	math	scie	fran	lati	d-m
math	11.39	9.92	2.66	4.82	0.11
scie	9.92	8.94	4.12	5.48	0.06
fran	2.66	4.12	12.06	9.29	0.39
lati	4.82	5.48	9.29	7.91	0.67
d-m	0.11	0.06	0.39	0.67	8.67

Axes principaux d'inertie

La diagonalisation de la matrice de variance fournit les valeurs propres suivantes (rangées par ordre décroissant)

$$\lambda_1 = 28.2533, \quad \lambda_2 = 12.0747, \quad \lambda_3 = 8.6157, \quad \lambda_4 = 0.0217, \quad \lambda_5 = 0.0099.$$

et les vecteurs propres normés ou axes principaux d'inertie suivants

$$\mathbf{u}_1 = \begin{pmatrix} 0.51 \\ 0.51 \\ 0.49 \\ 0.48 \\ 0.03 \end{pmatrix}, \quad \mathbf{u}_2 = \begin{pmatrix} -0.57 \\ -0.37 \\ 0.65 \\ 0.32 \\ 0.11 \end{pmatrix}, \quad \mathbf{u}_3 = \begin{pmatrix} -0.05 \\ -0.01 \\ 0.11 \\ 0.02 \\ -0.99 \end{pmatrix}, \quad \mathbf{u}_4 = \begin{pmatrix} 0.29 \\ -0.55 \\ -0.39 \\ 0.67 \\ -0.03 \end{pmatrix}, \quad \mathbf{u}_5 = \begin{pmatrix} -0.57 \\ 0.55 \\ -0.41 \\ 0.45 \\ -0.01 \end{pmatrix}.$$

Qualité de la représentation

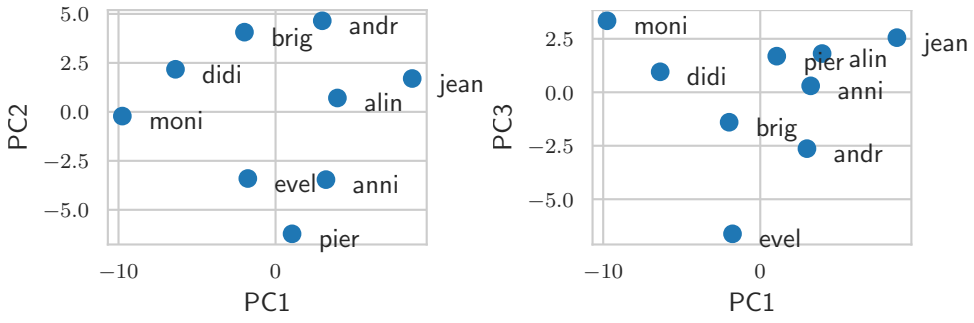
Rappelons que les inerties du nuage projeté sur les 5 axes sont égales aux valeurs propres. L'inertie du nuage est égale à $\mathrm{Tr}(VM) = \mathrm{Tr}(V)$, c'est-à-dire aussi à la somme des valeurs propres, ici 48.975. Les pourcentages d'inertie expliquée par chaque axe sont donc de 57.69, 24.65, 17.59, 0.04 et 0.02. Les pourcentages d'inertie expliquée par les sous-espaces principaux sont 57.69, 82.34, 99.94, 99.98 et 100.00. On peut donc conclure que le nuage initial est pratiquement dans un espace de dimension 3.

Composantes principales

La matrice des composantes principales $C = XMU = XU$ est la suivante :

	PC1	PC2	PC3	PC4	PC5
jean	8.70	1.70	2.55	-0.15	-0.12
alin	3.94	0.71	1.81	-0.09	0.04
anni	3.21	-3.46	0.30	0.17	0.02
moni	-9.76	-0.22	3.34	-0.17	0.10
didi	-6.37	2.17	0.96	0.07	-0.19
andr	2.97	4.65	-2.63	-0.02	0.15
pier	1.05	-6.23	1.69	0.12	0.04
brig	-1.98	4.07	-1.40	0.24	0.01
evel	-1.77	-3.40	-6.62	-0.16	-0.06

Ces composantes principales permettent d'obtenir, par exemple, les plans de représentation 1,2 et 1,3 suivants :



Contributions relatives des axes aux individus

	PC1	PC2	PC3	PC4	PC5
jean	0.89	0.03	0.08	0.00	0.00
alin	0.80	0.03	0.17	0.00	0.00
anni	0.46	0.53	0.00	0.00	0.00
moni	0.89	0.00	0.11	0.00	0.00
didi	0.88	0.10	0.02	0.00	0.00
andr	0.24	0.58	0.19	0.00	0.00
pier	0.03	0.91	0.07	0.00	0.00
brig	0.17	0.74	0.09	0.00	0.00
evel	0.05	0.20	0.75	0.00	0.00

Contributions relatives des individus aux axes

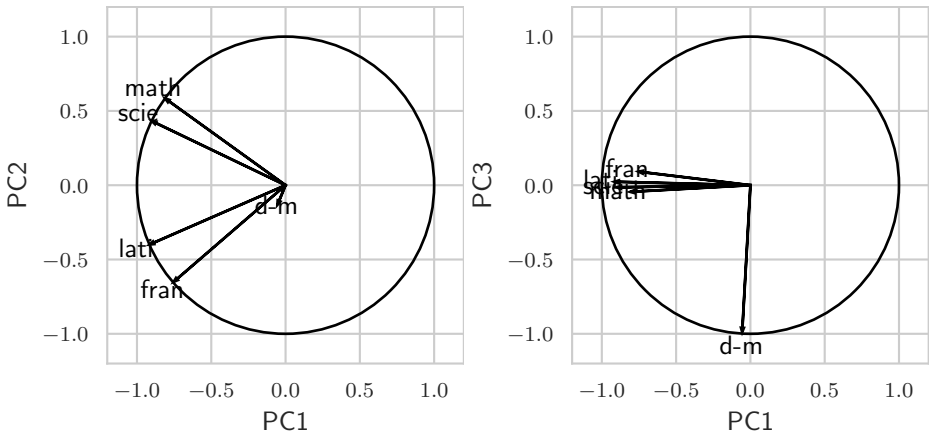
	PC1	PC2	PC3	PC4	PC5
jean	0.30	0.03	0.08	0.11	0.15
alin	0.06	0.00	0.04	0.04	0.02
anni	0.04	0.11	0.00	0.15	0.00
moni	0.37	0.00	0.14	0.15	0.11
didi	0.16	0.04	0.01	0.03	0.40
andr	0.03	0.20	0.09	0.00	0.25
pier	0.00	0.36	0.04	0.07	0.02
brig	0.02	0.15	0.03	0.30	0.00
evel	0.01	0.11	0.56	0.14	0.04

Analyse dans $\mathbb{R}^n$

Calcul des corrélations $\mathrm{Cor}(\alpha, j)$.

	PC1	PC2	PC3	PC4	PC5
math	-0.81	0.58	-0.04	-0.01	0.02
scie	-0.90	0.43	-0.01	0.03	-0.02
fran	-0.75	-0.65	0.09	0.02	0.01
lati	-0.92	-0.40	0.02	-0.04	-0.02
d-m	-0.06	-0.13	-0.99	0.00	0.00

Finalement, ces composantes principales normées associées aux variables permettent d'obtenir, par exemple, les plans de représentation 1,2 et 1,3 :



Chapitre 6

Positionnement multidimensionnel

1 Introduction

Lorsque les données sont fournies sous la forme d'un ensemble d'individus mesurés par un ensemble de variables, l'analyse en composantes principales et les méthodes qui en sont issues comme l'analyse des correspondances et l'analyse des correspondances multiples fournissent une représentation fidèle des données dans des espaces euclidiens de faible dimension permettant, par exemple, de visualiser les données sur un plan.

L'analyse des proximités, encore appelée positionnement multidimensionnel (*multidimensional scaling, MDS*) ou analyse ordinale (en écologie par exemple), a aussi pour objectif d'obtenir une représentation fidèle des données dans des espaces euclidiens de faible dimension, souvent le plan, mais cette fois à partir d'un tableau de proximités entre les individus. Rappelons que les notions de proximité et de tableau de proximités ont été définies dans le chapitre 2.

Historiquement, ces méthodes ont été développées et proposées dans la revue Psychometrika dans les années 1950 par Torgerson et Shepard. Parmi les références portant sur l'analyse des proximités, on peut citer les deux ouvrages récents Borg and Groenen (2005) et Cox and Cox (1994).

De manière générale, dans tout ce chapitre, les résultats seront bien sûr toujours déterminés aux isométries près (translations, rotations, symétries,...).

2 Le problème

Supposons que l'on dispose d'une matrice de dissimilarités $\Delta = (\delta_{ij})$ portant sur n individus. L'objectif du positionnement multidimensionnel est de déterminer une représentation X de dimension p telle que la distance euclidienne associée $D(X)$ soit proche de la dissimilarité initiale Δ.

3 Quelques résultats théoriques

3.1 Matrice de centrage

Dans ce qui suit, on aura besoin de centrer des vecteurs ou des matrices. Pour cela, on introduit la matrice de centrage.

Définition 1. *La matrice de centrage d'ordre n notée Q_n est la matrice carré de taille n suivante :*

$$Q_n = I_n - \frac{1}{n}U_n = \begin{pmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \dots & -\frac{1}{n} & 1-\frac{1}{n} \end{pmatrix}.$$

La matrice Q_n est utilisée pour réaliser le centrage d'un vecteur. Ainsi, si u est un vecteur de taille n alors $Q_n u$ est le vecteur u centré. De même, si A est une matrice de taille $n \times p$, la matrice $Q_n A$ est la matrice A après centrage de toutes les colonnes et la matrice AQ_p est la matrice A après centrage de toutes les lignes. Enfin, la matrice $Q_n A Q_p$ est la matrice A centrée en lignes et en colonnes (on montre au passage que l'ordre du centrage n'est pas important par associativité du produit matriciel). La matrice Q_n est déjà centrée en colonnes (et donc en lignes puisque qu'elle est symétrique), on a donc $Q_n^2 = Q_n$.

Géométriquement, l'opérateur Q_n réalise une projection orthogonale parallèlement au vecteur $(1, \dots, 1)^T$ dans l'espace $\mathbb{R}^n$.

3.2 Bijection fondamentale

Définissons les deux sous-espaces vectoriels de matrices suivants.

$$\mathcal{D}_0 = \{D \in \mathcal{M}_n(\mathbb{R}) \mid D \text{ symétrique et } d_{ii} = 0\},$$

l'espace vectoriel des matrices symétriques de diagonale nulle. Il contient en particulier les matrices de dissimilarité.

$$\mathcal{W} = \{W \in \mathcal{M}_n(\mathbb{R}) \mid W \text{ symétrique et centrée}\},$$

l'espace vectoriel des matrices symétriques centrées (donc centrées aussi en lignes).

On va montrer qu'il existe une bijection entre ces deux ensembles. L'intérêt de cette bijection est que des propriétés difficiles à caractériser dans l'ensemble $\mathcal{D}_0$ correspondent à des propriétés bien connues dans l'ensemble $\mathcal{W}$.

Proposition 2. *Les deux ensembles $\mathcal{D}_0$ et $\mathcal{W}$ sont en bijection et les fonctions suivantes sont réciproques l'une de l'autre*

$$\begin{array}{rccl} \varphi\colon \mathcal{D}_0 & \longrightarrow & \mathcal{W} & \\ D & \longmapsto & \varphi(D) = -\frac{1}{2}Q_n D Q_n & \end{array} \qquad \begin{array}{rccl} \psi\colon \mathcal{W} & \longrightarrow & \mathcal{D}_0 & \\ W & \longmapsto & \psi(W) = \mathbf{h}\mathbb{1}_n^T - 2W + \mathbb{1}_n\mathbf{h}^T & \end{array}$$

où $\mathbf{h}$ est la diagonale de W : $\mathbf{h} = \operatorname{diag}(W)$. On peut ainsi définir la matrice $\psi(W)$ par son terme général : $(\psi(W))_{ij} = w_{ii} - 2w_{ij} + w_{jj}$.

L'opération $Q_n D Q_n$ est appelée double-centrage de la matrice D. La fonction φ est donc simplement un double-centrage suivi de la multiplication par $-1/2$. L'écriture $\mathbf{h}\mathbb{1}_n^T$ dénote la matrice où la colonne $\mathbf{h}$ a été mise n fois en lignes. De même, la matrice $\mathbb{1}_n\mathbf{h}^T$ est la matrice où n vecteurs $\mathbf{h}$ sont mis côte à côte.

$$\mathbf{h}\mathbb{1}_n^T = \begin{pmatrix} \mathbf{h}^T \\ \vdots \\ \mathbf{h}^T \end{pmatrix} \quad \text{et} \quad \mathbb{1}_n\mathbf{h}^T = [\mathbf{h}, \dots, \mathbf{h}].$$

Preuve de la proposition 2. On vérifie sans problème que les fonctions φ et ψ sont bien définies c'est à dire que $-\frac{1}{2}Q_n D Q_n \in \mathcal{W}$ lorsque $D \in \mathcal{D}_0$ et $\mathbf{h}\mathbb{1}_n^T - 2W + \mathbb{1}_n\mathbf{h}^T \in \mathcal{D}_0$ lorsque $W \in \mathcal{W}$.

Les deux fonctions sont également linéaires. Il suffit donc de montrer que la fonction $\varphi\circ\psi$ est égale à l'identité. Soit $W \in \mathcal{W}$, on a successivement

$$\begin{aligned}\varphi \circ \psi(W) &= -\frac{1}{2}Q_n\big(\mathbf{h}\mathbb{1}_n^T - 2W + \mathbb{1}_n\mathbf{h}^T\big)Q_n \\ &= -\frac{1}{2}Q_n\mathbf{h}\mathbb{1}_n^TQ_n + Q_nWQ_n - \frac{1}{2}Q_n\mathbb{1}_n\mathbf{h}^TQ_n\end{aligned}$$

Or, on a $Q_n\mathbb{1} = 0$ et $Q_nW = W$ car W est centré. On a donc

$$\varphi \circ \psi(W) = Q_nWQ_n = W.$$

□

4 Distances euclidiennes

Dans cette section, on s'intéresse à la bijection établie à la section précédente pour les matrices de distance euclidienne ou plutôt pour les matrices de distance euclidienne au carré.

Si D est une matrice de distance, on note alors D^2 la matrice des carrés des entrées de D, à distinguer du produit matriciel de D par elle-même.

Le théorème suivant est intéressant puisqu'il caractérise les distances euclidiennes grâce à la bijection.

Théorème 3. *Soit D une matrice de distance. Alors, D est euclidienne si et seulement si $\varphi(D^2) = -1/2Q_nD^2Q_n$ est semi-définie positive.*

Si cette propriété est vérifiée, on a alors les propriétés supplémentaires suivantes :

(1) Pour toute représentation centrée X de la distance euclidienne D, on a

$$\varphi(D^2) = XX^T,$$

(2) Toute représentation X de la distance euclidienne D est dans un espace de dimension au moins r avec r le rang de la matrice $\varphi\big(D^2\big)$

(3) Une représentation dans un espace de dimension r est donnée par $X_r = V_r\sqrt{\Lambda_r}$ avec V_r les vecteurs propres de $\varphi\big(D^2\big)$ correspondants aux valeurs propres non nulles et Λ_r la matrice diagonale de ces valeurs propres, dans le même ordre.

Preuve. Si D est euclidienne, soit X une représentation centrée de la distance D. La matrice XX^T est donc centrée et symétrique. Montrons que $\varphi(D^2) = XX^T$ ou de manière équivalente, $\psi(XX^T) = D^2$. On a

$$\begin{aligned}\big(\psi(XX^T)\big)_{ij} &= \big(XX^T\big)_{ii} - 2\big(XX^T\big)_{ij} + \big(XX^T\big)_{jj} \\ &= \|\boldsymbol{x}_i\|^2 - 2\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \|\boldsymbol{x}_j\|^2 \\ &= \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2.\end{aligned}$$

On a donc bien $\varphi(D^2) = XX^T$ ce qui au passage démontre (1). Dès lors, $\varphi(D^2)$ est semi-définie positive puisque pour tout $u \in \mathbb{R}^n$

$$\begin{aligned}u^T\varphi(D^2)u &= u^TXX^Tu \\ &= \big(X^Tu\big)^TX^Tu \\ &= \big\|X^Tu\big\|^2 \geq 0.\end{aligned}$$

Inversement, supposons $\varphi\big(D^2\big)$ semi-définie positive. En la diagonalisant, on obtient

$$\varphi\big(D^2\big) = V\Lambda V^T,$$

avec V orthogonale et $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$, les valeurs propres étant positives car $\varphi\big(D^2\big)$ est semi-définie positive.

En notant r le rang de $\varphi\big(D^2\big)$, V_r la matrice des r premières colonnes de V et $\Lambda_r = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_r)$ on a aussi

$$\varphi\big(D^2\big) = V_r \Lambda_r V_r^T,$$

car $\lambda_{r+1} = \cdots = \lambda_n = 0$.

En notant $\sqrt{\Lambda_r} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$, on a

$$\varphi\big(D^2\big) = V_r \sqrt{\Lambda_r}\sqrt{\Lambda_r} V_r^T.$$

En posant $X_r = V_r\sqrt{\Lambda_r}$, on trouve $\varphi\big(D^2\big) = X_r X_r^T$. On a alors

$$D_{ij}^2 = \big(\psi\big(X_r X_r^T\big)\big)_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2,$$

ce qui montre que D^2 est une distance euclidienne, termine l'équivalence et démontre au passage la propriété (3).

La propriété (2) découle du fait que si on considère une représentation quelconque X dans une espace de dimension p que l'on centre, on a $\varphi(D^2) = XX^T$ d'après (1) et

$$\begin{aligned} r &= \mathrm{rang}\,\varphi(D^2) \\ &= \mathrm{rang}\,\big(XX^T\big) \\ &\leq p \end{aligned}$$

car X compte p colonnes. □

Remarquons que comme $\phi(D^2)$ est doublement centrée, $\phi(D^2)\mathbb{1}_n = 0$ donc 0 est valeur propre. Son rang r est donc inférieur ou égal à $n-1$. En d'autre terme, une matrice de distance euclidienne sur n points admet une représentation dans un espace de dimension inférieure à $n-1$.

Exemple 6.1. *Soit D la matrice de distance suivante*

$$D = \begin{pmatrix} 0 & 1 & \sqrt{2} \\ 1 & 0 & 1 \\ \sqrt{2} & 1 & 0 \end{pmatrix}.$$

Pour savoir si la matrice D est une matrice de distance euclidienne, on calcule la matrice W correspondante. Pour cela, on commence par mettre au carré tous les éléments de D, on multiplie par $-1/2$ et on effectue un double–centrage. On obtient

$$W = \frac{1}{9}\begin{pmatrix} 5 & -1 & -4 \\ -1 & 2 & -1 \\ -4 & -1 & 5 \end{pmatrix}.$$

Pour savoir si W est semi-définie positive, on calcule ses valeurs propres qui sont 1, $1/3$ et 0. La distance D est donc bien euclidienne.

Pour trouver une représentation de cette distance euclidienne, on cherche des vecteurs propres c_1 et c_2 associés à $\lambda_1 = 1$ et $\lambda_2 = 1/3$ vérifiant $\|c_i\|^2 = \lambda_i$, $i = 1, 2$. On trouve

$$c_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{pmatrix}, \qquad c_2 = \begin{pmatrix} \frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{3} \\ \frac{\sqrt{2}}{6} \end{pmatrix}$$

D'où la représentation

$$x_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{6} \end{pmatrix}, \quad x_2 = \begin{pmatrix} 0 \\ -\frac{\sqrt{2}}{3} \end{pmatrix}, \quad x_3 = \begin{pmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{6} \end{pmatrix}.$$

5 Analyse factorielle d'un tableau de distances

Cette méthode est historiquement la première technique de positionnement multidimensionnel et a été développée par Torgerson (1952). Elle est aussi connue sous les noms d'analyse du triple (Benzecri (1973)), de codage en composantes principales, de *principal coordinate analysis* ou encore de *classical scaling*.

5.1 $W = -\frac{1}{2}Q_n\Delta^2 Q_n$ est SDP

Nous venons de voir que dans ce cas, il existait une représentation euclidienne exacte de dimension $\leq n-1$. Pour obtenir une représentation de dimension p fixée, il suffit alors d'utiliser l'ACP sur X et de retenir les p premiers axes. Mais comme les composantes principales sont les vecteurs propres ordonnés de norme λ_α de $\frac{1}{n}W$, la matrice des p premiers vecteurs propres fournit une solution au problème.

En pratique, il faudra donc :

1. calculer la matrice $W = \varphi^{-1}(\Delta^2) = -\frac{1}{2}Q_n\Delta^2 Q_n$,
2. diagonaliser la matrice $\frac{1}{n}W$,
3. ordonner les valeurs propres et vecteurs propres et normer les vecteurs propres (au sens de $\frac{1}{n}I$: si les vecteurs propres étaient normés au sens habituel, il suffit de les multiplier par $\sqrt{n}$),
4. calculer les composantes principales $C = V\sqrt{L}$ où L et V sont les matrices associées à ces valeurs propres et vecteurs propres
5. utiliser ces résultats comme pour une ACP classique (pourcentage d'inertie, choix du nombre d'axes, …).

La vérification de l'hypothèse $W = -\frac{1}{2}Q_n\Delta^2 Q_n$ est SDP se fait *a posteriori* ; il faut et il suffit que toutes les valeurs propres soient positives ou nulles.

5.2 $W = -\frac{1}{2}Q_n\Delta^2 Q_n$ n'est pas SDP

Lorsqu'il existe des valeurs propres négatives, plusieurs stratégies peuvent être envisagées :

Application directe de l'AFTD

L'AFTD est utilisée normalement comme s'il existait une représentation euclidienne et seules les composantes principales associées aux valeurs propres positives sont utilisées. Les résultats seront en pratique assez bons si les valeurs propres négatives sont petites (en valeur absolue). Toutefois, la définition du pourcentage d'inertie expliquée par un axe ne convient plus puisque la somme des valeurs propres positives est supérieure à la somme totale des valeurs propres. Généralement, la somme des valeurs propres est remplacée par la somme des valeurs absolues des valeurs propres.

Transformation de la dissimilarité en distance

Il existe différents moyens pour ce faire. On peut par exemple additionner une certaine constante à la dissimilarité initiale, afin de la transformer en une distance. On peut alors appliquer l'AFTD sur cette distance. En pratique, cette méthode ne donne pas toujours de très bons résultats.

6 Qualité de l'ajustement

Pour évaluer la qualité du positionnement multidimensionnel, on dispose de plusieurs méthodes. Comme le positionnement multidimensionnel dérive d'une ACP, on peut s'intéresser aux valeurs propres et aux inerties expliquées correspondantes. La dimension de la représentation peut être alors choisie avec la méthode du coude.

6.1 Méthode du coude

Il s'agit de regarder la décroissance des valeurs propres et de repérer visuellement un lieu de rupture dans la décroissance. Contrairement à l'ACP, les valeurs propres peuvent être négatives dans le cas où la distance n'est pas euclidienne.

6.2 Diagramme de Shepard

Le diagramme de Shepard consiste à confronter les distances réelles aux distances issues de l'AFTD. Plus l'AFTD est de bonne qualité, plus le nuage de points se rapproche de la droite $y = x$.

7 Exemple d'AFTD

Nous allons appliquer l'AFTD aux données d'Ekman portant sur la couleur (voir Ekman (1954)). Il s'agit d'un tableau de similarité entre 14 couleurs.

	1434	1445	1465	1472	1490	1504	1537	1555	1584	1600	1610	1628	1651	1674
1434	1.00	0.86	0.42	0.42	0.18	0.06	0.07	0.04	0.02	0.07	0.09	0.12	0.13	0.16
1445	0.86	1.00	0.50	0.44	0.22	0.09	0.07	0.07	0.02	0.04	0.07	0.11	0.13	0.14
1465	0.42	0.50	1.00	0.81	0.47	0.17	0.10	0.08	0.02	0.01	0.02	0.01	0.05	0.03
1472	0.42	0.44	0.81	1.00	0.54	0.25	0.10	0.09	0.02	0.01	0.00	0.01	0.02	0.04
1490	0.18	0.22	0.47	0.54	1.00	0.61	0.31	0.26	0.07	0.02	0.02	0.01	0.02	0.00
1504	0.06	0.09	0.17	0.25	0.61	1.00	0.62	0.45	0.14	0.08	0.02	0.02	0.02	0.01
1537	0.07	0.07	0.10	0.10	0.31	0.62	1.00	0.73	0.22	0.14	0.05	0.02	0.02	0.00
1555	0.04	0.07	0.08	0.09	0.26	0.45	0.73	1.00	0.33	0.19	0.04	0.03	0.02	0.02
1584	0.02	0.02	0.02	0.02	0.07	0.14	0.22	0.33	1.00	0.58	0.37	0.27	0.20	0.23
1600	0.07	0.04	0.01	0.01	0.02	0.08	0.14	0.19	0.58	1.00	0.74	0.50	0.41	0.28
1610	0.09	0.07	0.02	0.00	0.02	0.02	0.05	0.04	0.37	0.74	1.00	0.76	0.62	0.55
1628	0.12	0.11	0.01	0.01	0.01	0.02	0.02	0.03	0.27	0.50	0.76	1.00	0.85	0.68
1651	0.13	0.13	0.05	0.02	0.02	0.02	0.02	0.02	0.20	0.41	0.62	0.85	1.00	0.76
1674	0.16	0.14	0.03	0.04	0.00	0.01	0.00	0.02	0.23	0.28	0.55	0.68	0.76	1.00

Comme il s'agit d'un tableau de similarité dont les valeurs sont entre 0 et 1, la première chose à faire est de le transformer en un tableau de dissimilarités en complémentant à 1. Les valeurs propres obtenues par l'AFTD sont tracées à la figure 6.1. On peut remarquer qu'il y a des valeurs propres négatives (la matrice de dissimilarité initiale n'est pas euclidienne) mais qu'elles sont très petites et ne sont pas gênantes.

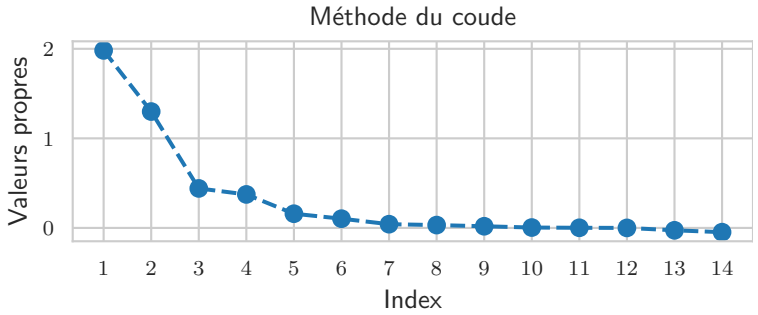
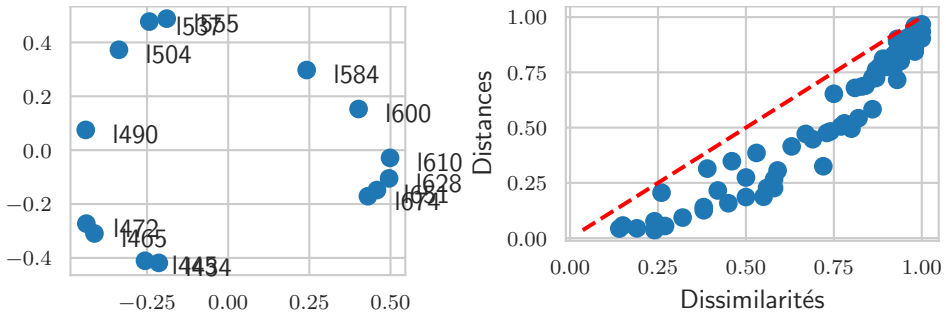


FIGURE 6.1 – Valeurs propres issues de l'AFTD sur le jeu de données Ekman

On en déduit la représentation dans le premier plan factoriel à la figure 6.2a qui fournit une bonne représentation des données comme le montre le diagramme de Shepard correspondant à la figure 6.2b.



(a) Données d'Ekman dans le premier plan factoriel

(b) Diagramme de Shepard pour $k = 2$

FIGURE 6.2 – Représentation et diagnostic sur le jeu de données Ekman

8 Méthodes non linéaires

Il est possible de montrer que la solution obtenue par l'AFTD minimise le critère $\sum_{i,i'}(\delta_{ii'}^2 - d_{ii'}^2)$ sous la contrainte que la représentation est de dimension p fixée et qu'en plus $d_{ii'} \leq \delta_{ii'}$ pour tous les couples d'individus i, i'. Cette méthode revient à projeter dans un espace de faible dimension une représentation parfaite dans un espace de grande dimension de la distance initiale et donc finalement à effectuer une transformation linéaire des données initiales. Les méthodes développées dans la suite n'imposeront plus que la représentation obtenue soit une projection linéaire. L'objectif de ces méthodes sera donc de trouver une représentation euclidienne X dans un espace de dimension fixée k telle que la distance euclidienne D associée minimise une fonction d'écart entre Δ et D appelée *Stress*.

8.1 Fonctions Stress

Plusieurs fonctions ont été proposées :

$$\text{Stress}_1(X) = \frac{\sum_{i<i'}(\delta_{ii'} - d_{ii'})^2}{\sum_{i<i'} d_{ii'}^2}$$

$$\text{Stress}_2(X) = \frac{\sum_{i<i'} w_{ii'}(\delta_{ii'} - d_{ii'})^2}{\sum_{i<i'} w_{ii'} d_{ii'}^2}$$

où les $w_{ii'}$ sont des pondérations données *a priori* (ces pondérations permettent de prendre en compte, par exemple, la présence de données manquantes).

$$\text{Stress}_3(X) = \frac{1}{\sum_{i<i'} \delta_{ii'}} \sum_{i<i'} \frac{(\delta_{ii'} - d_{ii'})^2}{\delta_{ii'}}$$

Tous ces critères sont normalisés de manière à être invariants pour des rotations, translations et changements d'échelles. Remarquons que le dernier critère prend en compte de manière plus importante les erreurs commises sur les petites distances.

8.2 Optimisation

Il n'existe pas d'algorithme permettant de résoudre en toute généralité ce type de problème et le plus souvent, les méthodes proposées sont des méthodes d'optimisation itératives qui font simplement décroître le critère et conduisent donc à des optima locaux du critère. On peut citer les méthodes suivantes :

— méthodes de gradient ;

— méthode SMACOF (« Scaling by MAjorizing a COmplicated Function », la plus efficace à ce jour) ;
— méthode de Newton.

8.3 Projection de Sammon

La projection de Sammon, très utilisée dans le monde de la reconnaissance de formes, utilise le critère Stress_3 et la méthode de Newton. En R, la fonction `sammon` est disponible dans le module `MASS`.

8.4 Remarques

— Le choix du nombre de dimensions se fait généralement, comme pour l'ACP, en étudiant la décroissance du critère en fonction de la dimension (méthode du coude). Toutefois, contrairement à l'AFTD, les calculs doivent être recommencés pour chaque dimension et les solutions ne sont pas emboîtées.
— Cette approche ne pose le problème des valeurs propres négatives comme pour l'AFTD ; quelque soit la dissimilarité initiale, une solution est obtenue. Toutefois, la méthode ne garantit pas l'optimum global et donc l'unicité de la solution. Généralement les logiciels prennent comme point de départ les résultats obtenus par l'AFTD.
— En dehors du critère minimisé, un certain nombre d'outils permettent d'analyser les résultats. On peut citer, par exemple, le graphique représentant les couples δ_{ij}, d_{ij}.

9 Méthodes non métriques ou ordinales

9.1 Généralisation

L'approche précédente peut être étendue en relâchant les contraintes du problème. L'idée sous-jacente est qu'en relâchant le lien entre la dissimilarité et la distance obtenue, le résultat soit plus fidèle. Pour ceci, une fonction supplémentaire f est introduite dans le critère de la façon suivante :

$$\text{Stress}(X, f) = \frac{\sum_{i<i'}(f(\delta_{ii'}) - d_{ii'})^2}{\sum_{i<i'} d_{ii'}^2}.$$

L'objectif est alors de déterminer le couple (X, f) minimisant ce critère. Plusieurs situations ont été envisagées ; par exemple

— f est une fonction linéaire $f(d_{ii'}) = \alpha d_{ii'} + \beta$
— f est une fonction exponentielle $f(d_{ii'}) = e^{\alpha d_{ii'} + \beta}$
— f est simplement une fonction monotone croissante : le critère ne prend en compte que l'ordre induit sur tous les couples d'individus par la dissimilarité initiale.

La solution du problème est obtenue par optimisation alternée :

— pour f fixée, on cherche la meilleure représentation X ; pour cela, il suffit d'appliquer l'une des méthodes précédentes à la dissimilarité $f(\Delta)$;
— pour X fixée, on cherche la meilleure fonction f ; il s'agit alors d'un problème de régression.

9.2 Projection de Kruskal

Dans cette méthode, développée par Shepard et Kruskal et connue sous le nom de *Non metric multidimensional scaling*, la fonction f est simplement monotone croissante et

l'algorithme de régression est un algorithme original appelé régression isotonique. En `R`, la fonction correspondante `isoMDS` est disponible dans le module `MASS`.

Comme pour les méthodes précédentes, des outils d'analyse, comme le diagramme de Shepard, ont été développés.

10 Quelques remarques

10.1 Dissimilarités initiales

La dissimilarité initiale Δ peut recouvrir de nombreuses situations. En particulier, ces méthodes peuvent être utilisées pour étudier les liens existant entre les variables, par exemple en partant d'une distance entre variables définie à partir des corrélations.

10.2 Autres méthodes

On peut citer quelques méthodes voisines : par exemple, l'analyse procrustéenne permet de comparer deux tableaux de dissimilarités et si il y a plus de deux tableaux de dissimilarités, les méthodes de dépliage (*unfolding method*) permettent de comparer ces différents tableaux et la méthode Indscal (*Individual differences*) permet de représenter simultanément les tableaux et les individus sur lesquels portent ces dissimilarités.

Chapitre 7

La classification automatique

1 Introduction

Comme toutes les méthodes de l'Analyse des Données, la *Classification Automatique* a pour but d'obtenir une représentation simplifiée des données initiales. Il s'agit donc, comme l'analyse en composantes principales, d'une méthode de réduction des données. La classification, à ne pas confondre avec le classement, est l'organisation d'un ensemble en *classes homogènes* ou *classes naturelles*. La classification est la définition de classes alors que le classement est le rangement dans des classes déjà existantes. Il s'agit d'une démarche très courante. Par exemple, en statistique, cela permet d'identifier plusieurs populations dans une population initiale hétérogène et ainsi de faciliter une étude statistique ultérieure ; en politique, la classification en *droite* et *gauche* permet de mieux situer les hommes politiques ; en science naturelle, la classification du règne animal et du règne végétal proposée pour la première fois par Linné (naturaliste suédois du 18^{e} siècle) est l'une des classifications les plus connues ; et, de manière plus générale, le fait de nommer des objets est une forme de classification.

La terminologie peut dépendre du domaine : en science naturelle, la *systématique* ou *taxinomie* encore appelée *taxonomie* se définit comme la science de la classification des formes vivantes ; en médecine, la *nosologie* est la classification des maladies ; en reconnaissance des formes, la classification automatique est connue sous le nom de *classification non supervisée* ou *classification sans professeur* ; enfin en marketing, on parle plutôt de *typologie*.

La classification automatique, encore appelée *clustering* ou *taxonomie numérique*, objet de ce chapitre, recouvre l'ensemble des méthodes permettant la construction *automatique* de telles classifications. Une définition formelle de la classification, qui puisse servir de base à un processus automatisé, amène à se poser les questions suivantes : Comment les objets à classer sont-ils définis ? Comment définir la notion de ressemblance entre objets ? Qu'est-ce qu'une classe ? Comment sont structurées les classes ? Comment juger une classification par rapport à une autre ?

Pour effectuer cette classification, deux démarches sont généralement utilisées :

— On regroupe en classe les objets qui partagent certaines caractéristiques. Considérons le nombre de doigts d'un être vivant et comparons le singe et l'homme : sur ce critère de comparaison (et sur bien d'autres) les deux espèces seront jugés semblables. Ce genre de démarche aboutit à une classification *monothétique*, base de l'approche aristotélicienne (Sutcliffe, 1994). Tous les objets d'une même classe partagent alors un certain nombre de caractéristiques (par exemple : « Tous les hommes sont mortels »).

— On peut aussi regrouper en classe les objets qui posséderont des caractéristiques « proches ». Cette démarche est dite *polythétique*. Par exemple, une espèce polythétique est une espèce définie par un certain nombre de critères, dont aucun

n'est nécessaire ou suffisant par lui-même. Chaque individu de l'espèce doit posséder un certain nombre de caractéristiques mais aucune de celles-ci ne doivent être communs à chacun des individus de l'espèce. Généralement, on utilise pour cela la notion de mesure de proximité qui peut être une distance, une dissimilarité ou une similarité. C'est cette approche qui sera étudiée dans ce chapitre.

Terminons cette introduction par deux remarques : lorsque les données se présentent sous la forme d'un tableau individus-variables, la classification, souvent effectuée sur l'ensemble des individus, peut sans difficulté être étendue à l'ensemble des variables ; enfin, certains problèmes sans rapport apparent avec l'analyse de données peuvent se formaliser comme des problèmes de classification automatique. On peut citer, par exemple, la localisation des centres en recherche opérationnelle et la segmentation en traitement d'images.

2 Structures de Classification

Les structures de classification peuvent être variées : *partitions*, *suite de partitions emboîtées* ou *hiérarchie*, *classes empiétantes* ou *recouvrement*, *classes de fortes densités, partitions floues.*

Dans toute cette partie, on cherche à classifier un ensemble fini Ω de cardinal n, $\Omega = \{\omega_1, \ldots, \omega_n\}$.

2.1 Partition

Définition 1. *Soit Ω un ensemble fini, on appelle* partition *de Ω un ensemble $P = (P_1, P_2, \ldots, P_g)$ de parties de Ω telles que*

(1) $\forall k \neq \ell$, $P_k \cap P_\ell = \varnothing$,

(2) $\bigcup_{k=1}^{g} P_k = \Omega$.

D'après la propriété (1), on a nécessairement $P_k \neq \varnothing$ pour tout $k = 1, \ldots, g$: les éléments d'une partition sont non vides.

Dans un ensemble Ω partitionné en g classes, chaque élément de l'ensemble appartient à une classe et une seule. Une manière pratique de décrire cette partition P consiste à lui associer la matrice de classification suivante :

$$\mathbf{z} = \begin{pmatrix} z_{11} & \cdots & z_{1g} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{ng} \end{pmatrix},$$

où $z_{ik} = 1$ si $i \in P_k$ et 0 sinon. Remarquons que la somme de la i^{e} ligne est égale à 1 (un élément appartient à une seule classe) et la somme des valeurs de la k^{e} colonne vaut n_k le nombre d'éléments de la classe P_k.

Partition floue La notion de partition repose sur une conception ensembliste classique. Considérant les travaux de Zadeh (1965) sur les ensembles flous, une définition du concept de partition floue semble « naturelle ». La classification floue, développée au début des années 1970 (Ruspini, 1969), généralise l'approche classique en classification en élargissant la notion d'appartenance à une classe. Dans le cadre de la conception ensembliste classique, un individu $\boldsymbol{x}_i$ appartient ou n'appartient pas à un ensemble donné P_k. Dans la théorie des sous-ensembles flous, un individu peut appartenir à plusieurs classes avec différents degrés d'appartenance. En classification, cela se traduit par le relâchement de la contrainte de binarité sur les coefficients d'appartenance c_{ik}. Une partition floue est définie par une matrice de classification floue $\mathbf{c} = \{c_{ik}\}$ vérifiant les conditions suivantes :

1. $\forall i, k$, $c_{ik} \in [0, 1]$,

2. $\forall k, \sum_i c_{ik} > 0$,
3. $\forall i, \sum_k c_{ik} = 1$.

La seconde condition traduit le fait qu'aucune classe ne doit être vide et la troisième exprime le concept d'appartenance totale.

2.2 La hiérarchie indicée

Définition 2. *Soit Ω un ensemble fini. On appelle* hiérarchie *sur Ω un ensemble H de parties non vides de Ω telles que*

(1) $\Omega \in H$,

(2) $\forall \boldsymbol{x} \in \Omega, \quad \{\boldsymbol{x}\} \in H$,

(3) $\forall h, h' \in H, \quad h \cap h' = \varnothing$ *ou* $h \subset h'$ *ou* $h' \subset h$.

Une hiérarchie contient donc tous les singletons ainsi que l'ensemble tout entier Ω. La propriété (3) montre que si l'intersection de deux parties est non vide alors l'une est inclue dans l'autre.

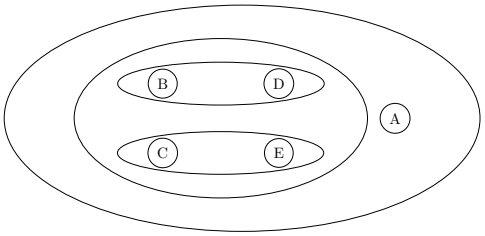
Exemple 7.1. *On suppose $\Omega = \{A, B, C, D, E\}$, les ensembles suivants sont des hiérarchie sur Ω*

1. $\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B, C, D, E\}\}$,
2. $\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{B, D\}, \{C, E\}, \{B, C, D, E\}, \{A, B, C, D, E\}\}$.

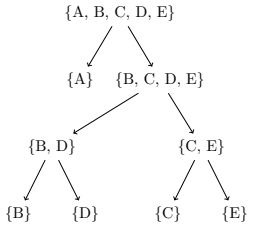
Les ensembles suivants n'en sont pas

1. $\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{A, B, C\}, \{C, D, E\}, \{A, B, C, D, E\}\}$,
2. $\{\{A\}, \{B, C\}, \{D, E\}, \{A, B, C\}, \{A, B, C, D, E\}\}$

Représentations graphiques d'une hiérarchie Une hiérarchie peut être représentée graphiquement de manière ensembliste (figure 7.1a) ou à l'aide d'une structure d'arbre (figure 7.1b).



(a) Représentation ensembliste



(b) Représentation en arbre

FIGURE 7.1 – Représentations graphiques d'une hiérarchie

Ces représentations sont rarement utilisées. Plus souvent, on préfère adjoindre un indice à la hiérarchie pour obtenir une représentation plus lisible.

Définition 3. *On appelle* indice *sur une hiérarchie H une fonction i de H dans $\mathbb{R}^+$ vérifiant les propriétés :*

(1) $h \subset h'$ *et* $h \neq h' \Rightarrow i(h) < i(h')$,

(2) $\forall \omega \in \Omega, \quad i(\{\omega\}) = 0$.

Le couple (H, i) est alors appelé hiérarchie indicée.

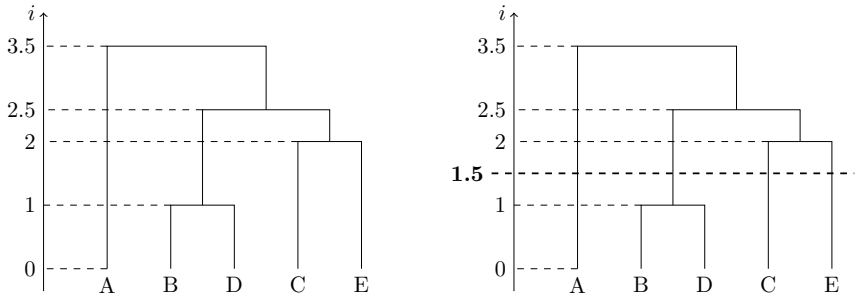
La propriété (1) impose à l'indice d'être une fonction strictement croissante. La propriété (2) impose à l'indice d'être nul sur les singletons (et sur eux seulement d'après la croissance stricte).

Exemples 7.2.

1. *La fonction i qui associe à un élément de H son cardinal ôté de* 1 *est un indice sur H.*
2. *On peut associer aux classes* $\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{B, D\}, \{C, E\}, \{B, C, D, E\}, \{A, B, C, D, E\}$ *de la hiérarchie précédente les valeurs* 0, 0, 0, 0, 0, 1, 2, 2.5, 3.5.

Dendrogramme

En utilisant un indice, il est possible d'obtenir une représentation graphique, appelée *dendrogramme*, ajoutant à la structure d'arbre le niveau de regroupement (voir figure 7.2a).



(a) Dendrogramme de la hiérarchie 7.1 avec l'indice i de l'exemple 2

(b) Section du dendrogramme à l'indice 1.5

FIGURE 7.2 – Dendrogramme et section

2.3 Partition et hiérarchie

Soit $P = (P_1, P_2, \ldots, P_g)$ une partition de Ω. L'ensemble H formé des classes P_k de P, des singletons de Ω et de l'ensemble Ω lui-même,

$$H = \{\{\omega_1\}, \ldots, \{\omega_n\}, P_1 \ldots, P_g, \Omega\},$$

forme une hiérarchie.

Remarquons qu'inversement, il est possible d'associer à chaque niveau d'une hiérarchie indicée une partition. Formellement, pour une valeur d'indice v, on associe la partition constituée des éléments maximaux de l'ensemble

$$H_{\leq v} = \{h \in H \mid i(h) \leq v\}.$$

La partition correspondant à l'indice maximum est la partition la plus grossière $P = \{\Omega\}$ et la partition correspondant à l'indice 0 est la partition la plus fine

$$P = \{\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_n\}\}.$$

Plus généralement, lorsque l'indice augmente, la partition devient de plus en plus grossière. Une hiérarchie indicée correspond donc à un ensemble de partitions emboîtées.

Visuellement la partition correspondant à un indice se déduit facilement lorsqu'on dispose du dendrogramme de la hiérarchie indicée.

Exemple 7.3. *La partition correspondant à l'indice* 1.5 *sur le dendrogramme de la figure 7.2b se voit en traçant la section de niveau* 1.5*. La partition résultante est donc*

$$\{\{A\}, \{B, D\}, \{C\}, \{E\}\}.$$

2.4 Aspects combinatoires

Le nombre de hiérarchies et de partitions qu'il est possible de définir sur un ensemble Ω devient vite énorme lorsque le cardinal de Ω augmente. Par exemple, le nombre de partitions d'un ensemble de n éléments en g classes est donné par la formule suivante :

$$S(n,g) = \frac{1}{g!}\sum_{k=0}^{g}(-1)^{k-1}C_g^k k^n.$$

Pour g fixé, lorsque n devient grand on a $S(n,g) \sim \frac{g^n}{g!}$.

Pour les premières valeurs, les nombres exacts sont les suivants :

		g							
		1	2	3	4	5	6	7	8
n	1	1							
	2	1	1						
	3	1	3	1					
	4	1	7	6	1				
	5	1	15	25	10	1			
	6	1	31	90	65	15	1		
	7	1	63	301	350	140	21	1	
	8	1	127	966	1701	1050	266	28	1

et on a, par exemple, $S(100,5) \approx 10^{67}$.

De la même manière, le nombre de hiérarchie sur un ensemble de cardinal n est donné par

$$\frac{(2n-2)!}{2^{n-1}(n-1)!} \sim \frac{2^{n-1}(n-2)!\sqrt{n-1}}{\sqrt{\pi}}.$$

n	1	2	3	4	5	6	7	8	9
Hiérarchies	1	1	3	15	105	945	10395	135135	2027025

3 Liens avec la notion d'ultramétrique

3.1 Recherche de partitions associées à une mesure de dissimilarité

Disposant d'une mesure de dissimilarité d sur l'ensemble Ω, on peut associer à toute valeur réelle $\alpha \geq 0$ la relation binaire de voisinage V_α sur Ω :

$$\boldsymbol{x}\,\mathrm{V}_\alpha\,\boldsymbol{y} \quad \text{si et seulement si} \quad d(\boldsymbol{x},\boldsymbol{y}) \leq \alpha.$$

Problème Peut-on trouver une partition de Ω qui est telle que tous les éléments d'une classe soient voisins et les éléments classés séparément ne soient pas voisins ?

Pour cela, il faut et il suffit que la relation V_α soit une relation d'équivalence. Les classes de la partition sont alors les classes d'équivalence de la relation. La fonction d étant une mesure de dissimilarité, la relation est réflexive et symétrique. Il faut et il suffit donc que la transitivité soit vérifiée, c'est-à-dire que

$$\forall\alpha \geq 0, \quad \boldsymbol{x}\,\mathrm{V}_\alpha\,\boldsymbol{y} \quad \text{et} \quad \boldsymbol{y}\,\mathrm{V}_\alpha\,\boldsymbol{z} \qquad \Rightarrow \qquad \boldsymbol{x}\,\mathrm{V}_\alpha\,\boldsymbol{z},$$

ce qui donne

$$\forall\alpha \geq 0, \quad d(\boldsymbol{x},\boldsymbol{y}) \leq \alpha \quad \text{et} \quad d(\boldsymbol{y},\boldsymbol{z}) \leq \alpha \qquad \Rightarrow \qquad d(\boldsymbol{x},\boldsymbol{z}) \leq \alpha. \tag{7.1}$$

Cette propriété n'est généralement pas vraie pour une mesure de dissimilarité mais nous allons montrer maintenant que cette propriété est équivalente à l'inégalité ultramétrique :

— Si d est une ultramétrique alors il est clair que l'équation 7.1 est vraie.

— Réciproquement, si d vérifie 7.1, pour tout triplet $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ quelconque de Ω, on obtient, en posant $\alpha = \max(d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{z}))$,

$$d(\boldsymbol{x}, \boldsymbol{y}) \leq \alpha \quad \text{et} \quad d(\boldsymbol{y}, \boldsymbol{z}) \leq \alpha \qquad \text{et donc} \qquad d(\boldsymbol{x}, \boldsymbol{z}) \leq \alpha,$$

qui s'écrit

$$d(\boldsymbol{x}, \boldsymbol{z}) \leq \max(d(\boldsymbol{x}, \boldsymbol{y}), d(\boldsymbol{y}, \boldsymbol{z})).$$

La relation 7.1 entraîne donc bien l'inégalité ultramétrique et l'équivalence est montrée.

3.2 Ultramétrique associée à une hiérarchie indicée : fonction φ

(H, i) étant une hiérarchie indicée sur Ω, on peut lui associer la mesure de dissimilarité

$$\delta : \Omega \times \Omega \to \mathbb{R}^+,$$

de la façon suivante :

$$\forall (\boldsymbol{x}, \boldsymbol{y}) \in \Omega^2, \quad \delta(\boldsymbol{x}, \boldsymbol{y}) = \inf \{ i(h), \quad h \in H \text{ et } \{\boldsymbol{x}, \boldsymbol{y}\} \subset h \}.$$

Remarquons que cette définition a bien un sens car l'ensemble $\{h \in H, \{\boldsymbol{x}, \boldsymbol{y}\} \subset h\}$ n'est pas vide puisqu'il contient au moins Ω.

Cette définition signifie que $\delta(\boldsymbol{x}, \boldsymbol{y})$ est égal au plus petit indice de toutes les classes de H contenant $\boldsymbol{x}$ et $\boldsymbol{y}$. La fonction i étant par définition croissante avec la relation d'inclusion, c'est-à-dire

$$h_1 \subset h_2 \Rightarrow i(h_1) \leq i(h_2).$$

$\delta(\boldsymbol{x}, \boldsymbol{y})$ s'interprète aussi comme l'indice de la plus petite classe (au sens de l'inclusion) de H contenant $\boldsymbol{x}$ et $\boldsymbol{y}$. On peut alors montrer la propriété suivante :

Proposition 4. *$\delta = \varphi(H, i)$ est une ultramétrique sur Ω.*

3.3 Hiérarchie indicée associée à une ultramétrique : fonction ψ

On considère les relations V_α sur Ω définies comme précédemment, mais cette fois à partir de l'ultramétrique δ. Nous savons alors que ces relations V_α sont pour tout $\alpha \geq 0$ des relations d'équivalence.

Construction d'une hiérarchie à partir de l'ultramétrique δ

D_δ étant l'ensemble des valeurs prises par l'ultramétrique δ sur Ω, on définit l'ensemble H comme l'ensemble de toutes les classes d'équivalence des relations V_α lorsque α parcourt D_δ. On peut alors montrer la proposition suivante :

Proposition 5. *H est une hiérarchie sur Ω.*

On définit alors la fonction i sur l'ensemble H :

$$\forall h \in H, \quad i(h) = \max_{\boldsymbol{x}, \boldsymbol{y} \in h} \delta(\boldsymbol{x}, \boldsymbol{y}). \qquad \text{(diamètre)}$$

La proposition suivante peut alors être facilement montrée.

Proposition 6. *La fonction i définit un indice sur la hiérarchie H.*

3.4 Équivalence entre hiérarchie indicée et ultramétrique

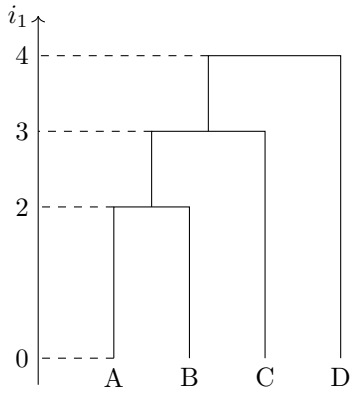
Proposition 7. *Les fonctions φ et ψ sont réciproques. C'est-à-dire :*

$$\psi \circ \varphi(H, i) = (H, i) \qquad et \qquad \varphi \circ \psi(\delta) = \delta.$$

Il y a donc équivalence entre la notion de hiérarchie indicée et d'ultramétrique.

3.5 Exemples

On part de la hiérarchie (H_1, i_1) dont le dendrogramme est le suivant :



La distance ultramétrique $\delta_1 = \varphi(H_1, i_1)$ obtenue est alors la suivante

	A	B	C
B	2		
C	3	3	
D	4	4	4

On peut maintenant appliquer la fonction ψ à l'ultramétrique δ_1 pour obtenir une hiérarchie indicée (H_2, i_2). On a $D_\delta = \{0, 2, 3, 4\}$. Les classes d'équivalence des 4 relations V_α sont :

$$\begin{array}{ll} V_0 : & \{\{A\}, \{B\}, \{C\}, \{D\}\} \\ V_2 : & \{\{A, B\}, \{C\}, \{D\}\} \\ V_3 : & \{\{A, B, C\}, \{D\}\} \\ V_4 : & \{\{A, B, C, D\}\}. \end{array}$$

La hiérarchie H_2 est donc

$$\{\{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{A, B, C\}, \{A, B, C, D\}\},$$

et les indices associés aux parties de cette hiérarchie sont respectivement

$$0, 0, 0, 0, 2, 3, 4.$$

On a bien retrouvé la hiérarchie indicée (H_1, i_1) initiale.

4 Objectifs de la classification

4.1 Difficultés de caractériser les objectifs

Rappelons que l'objectif de la classification automatique est l'organisation en classes homogènes des éléments d'un ensemble Ω. Pour définir cette notion de classes homogènes,

on utilise le plus souvent une mesure de similarité (ou de dissimilarité) sur Ω. Par exemple, si d est une mesure de dissimilarité sur Ω, on peut caractériser cette homogénéité en imposant aux classes de la partition recherchée de vérifier la propriété suivante :

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \text{même classe et} \forall \boldsymbol{z}, \boldsymbol{t} \in \text{classes différentes} \Rightarrow d(\boldsymbol{x}, \boldsymbol{y}) < d(\boldsymbol{z}, \boldsymbol{t}).$$

Cette propriété signifie simplement que l'on cherche à obtenir des classes telles que deux points d'une même classe se ressemblent plus que deux points de classes différentes.

En pratique, cet objectif est inutilisable. Par exemple sur la figure 7.3, alors qu'on « distingue clairement » deux classes, la distance entre les deux points 1 et 3 situés dans une même classe est supérieure à la distance entre les deux points 1 et 2 pourtant classés séparément.

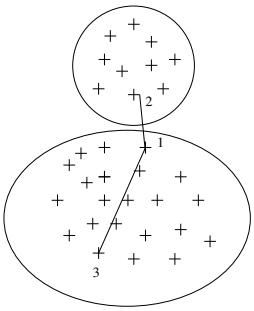


FIGURE 7.3

Plusieurs démarches sont alors utilisées pour remplacer cet objectif trop difficile à atteindre.

4.2 Démarche numérique

Partition

On remplace cette condition trop exigeante par une fonction numérique qui mesurera la qualité d'homogénéité d'une partition. Cette fonction est appelée généralement *critère*. Le problème peut paraître alors très simple. En effet, par exemple, dans le cas de la recherche d'une partition, il suffit de chercher parmi l'ensemble fini de toutes les partitions celle qui optimise le critère numérique. Malheureusement, le nombre de ces partitions étant très grand, leur énumération est impossible dans un temps raisonnable (explosion combinatoire). On utilise alors des heuristiques qui donnent, non pas la meilleure solution, mais une « bonne solution », c'est-à-dire une solution proche de la solution optimale. On parle alors d'optimisation locale. Lorsqu'il existe une structure d'ordre sur l'ensemble Ω et que celle-ci doit être respectée par la partition, il existe un algorithme de programmation dynamique, appelé algorithme de Fisher, qui fournit la solution optimale.

Hiérarchie

Dans le cas d'une hiérarchie, on cherchera à obtenir des classes d'autant plus homogènes qu'elles sont situées dans le bas de la hiérarchie. La définition d'un critère est moins facile. Nous verrons qu'il est possible de le faire en utilisant la notion d'ultramétrique (ultramétrique optimale).

Exemple de critère : inertie intra-classe

Ce critère peut être utilisé lorsque l'ensemble Ω à classifier correspond à un ensemble de n individus mesurés par p variables quantitatives. Il est alors possible, comme pour

l'ACP, de lui associer un nuage de points dans $\mathbb{R}^p$ muni des pondérations $\frac{1}{n}$ et de la distance euclidienne. La matrice de variance peut alors s'écrire

$$V = \frac{1}{n}(X - \mathbb{1}_n\overline{\boldsymbol{x}})^T(X - \mathbb{1}_n\overline{\boldsymbol{x}}) = \frac{1}{n}\sum_i(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T,$$

et l'inertie $\mathcal{I} = \frac{1}{n}\sum_i d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}})$ vérifie $\mathcal{I} = \mathrm{Tr}(V)$.

Si $P = (P_1, \ldots, P_g)$ est une partition de Ω en g classes, X_k la matrice X réduite aux lignes correspondant à la classe k et $\overline{\boldsymbol{x}}_k$ le centre de gravité de la classe k, on peut définir la matrice de variance intra-classe

$$V_W = \frac{1}{n}\sum_k n_k V_k,$$

où V_k est la matrice de variance de chaque classe ($V_k = \frac{1}{n_k}(X_k - \mathbb{1}_n\overline{\boldsymbol{x}}_k)^T(X_k - \mathbb{1}_n\overline{\boldsymbol{x}}_k)$) et l'inertie intra-classe

$$\mathcal{I}_W = \sum_k \mathcal{I}(P_k),$$

où $\mathcal{I}(P_k) = \frac{1}{n}\sum_{i\in P_k} d^2(\boldsymbol{x}_i, \overline{\boldsymbol{x}}_k)$ est l'inertie de la classe k. On peut alors montrer la relation

$$\mathcal{I}_W = \mathrm{Tr}(V_W).$$

Il est possible alors d'utiliser l'inertie intra-classe comme critère de classification : une partition sera d'autant plus homogène que l'inertie intra-classe sera proche de 0 ; en particulier, ce critère sera nul si tous les points de chaque classe sont concentrés en un même point.

4.3 Démarche algorithmique

Il s'agit cette fois de définir directement un algorithme qui construit des classes homogènes en tenant compte de la mesure de similarité. Il est relativement facile de proposer de tels algorithmes, le problème est de pouvoir vérifier que les résultats fournis sont intéressants et répondent au problème posé.

En réalité, cette démarche rejoint assez souvent la précédente. De nombreux algorithmes proposés sans référence à un critère et donnant de bons résultats optimisent un critère numérique. C'est le cas pour l'algorithme des centres mobiles qui sera décrit dans le paragraphe suivant.

5 La classification ascendante hiérarchique (CAH)

L'objectif est de construire une hiérarchie indicée d'un ensemble Ω sur lequel on connaît une mesure de dissimilarité d telle que les points les plus proches soient regroupés dans les classes de plus petit indice. Il existe essentiellement deux approches :

— la *classification descendante* : on divise l'ensemble Ω en classes, puis on recommence sur chacune de ces classes et ainsi de suite jusqu'à ce que les classes soient réduites à des singletons. Par exemple, on peut découper les classes par dichotomies successives, chacune de ces dichotomies étant définies par la vérification ou non d'une propriété. Dans le cas de classification animale, on sépare à une certaine étape, par exemple, ceux qui ont un squelette et ceux qui n'en ont pas.

— la *classification ascendante* : cette fois on part de la partition de Ω où chaque classe est un singleton. On procède alors par fusion successive des classes qui se « ressemblent » jusqu'à obtenir une seule classe, c'est-à-dire l'ensemble Ω lui-même. C'est cette procédure, beaucoup plus utilisée que la précédente, que nous étudions dans ce paragraphe.

5.1 L'algorithme

Critère d'agrégation

Ω étant l'ensemble à classifier et d une mesure de dissimilarité sur cet ensemble Ω, on définit, à partir de d, une « distance » D entre les parties de Ω. Cette distance D, souvent appelée *critère d'agrégation* est en réalité une mesure de dissimilarité qui ne vérifie pas nécessairement toutes les propriétés d'une distance sur l'ensemble des parties de Ω.

Construction de la hiérarchie

L'algorithme est alors le suivant :

1. Initialisation : partition des singletons et calcul des distances entre classes.
2. Tant que le nombre de classes est > 1
   - regroupement des 2 classes les plus proches au sens de D,
   - calcul des distances entre la nouvelle classe et les anciennes classes non regroupées.

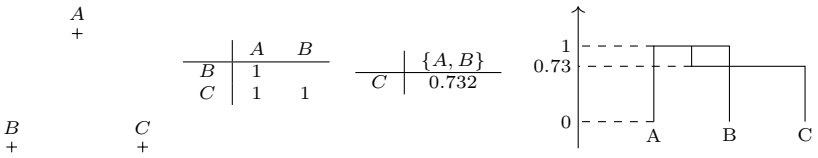
Il est facile de montrer que l'ensemble des classes définies au cours de cet algorithme forme une hiérarchie sur Ω.

Construction de l'indice

Après avoir défini une hiérarchie, il est nécessaire de lui associer un indice. Pour les classes du bas de la hiérarchie, c'est-à-dire les singletons, cet indice est nécessairement la valeur 0. Pour les autres classes, cet indice est généralement défini en associant à chacune des classes construites au cours de l'algorithme la distance D qui séparaient les deux classes fusionnées pour former cette nouvelle classe. Pour que cette définition conduise bien à un indice, il est nécessaire que les indices obtenus soient *strictement croissants* avec le niveau de la hiérarchie.

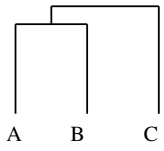
Plusieurs difficultés peuvent apparaître :

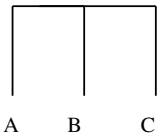
Inversion Pour certain critère d'agrégation, l'indice ainsi défini n'est pas nécessairement croissant. On parle alors d'inversion. Par exemple, si les données sont formées par trois points du plan situés au sommet d'un triangle équilatéral de côté 1 et si on prend comme distance D entre classes la distance entre les centres de gravité, on obtient une inversion.



Avec les critères d'agrégation étudiés dans ce chapitre, il est possible de montrer que l'inversion est impossible.

Croissante non stricte Lorsqu'il y a égalité de l'indice pour plusieurs niveaux emboîtés, il suffit de « filtrer » la hiérarchie, c'est-à-dire conserver une seule classe qui regroupe toutes les classes emboîtées ayant le même indice. En reprenant l'exemple du triangle équilatéral et en considérant cette fois le critère d'agrégation du lien maximum, la classe $\{A, B\}$ a le même indice que la classe $\{A, B, C\}$. Elle peut donc être supprimée. Ce problème peut se produire avec les critères d'agrégation que nous allons étudier et les algorithmes de mise en place de ces critères nécessiteront donc de prévoir cette opération de filtrage.





5.2 Les critères d'agrégation

Il existe de nombreux critères d'agrégation, mais les plus utilisés sont les suivants :

— critère du lien minimum (ou saut minimum ou *single linkage*)

$$D_{\min}(A,B) = \min\{d(i,i'),\ i \in A \text{ et } i' \in B\},$$

— critère du lien maximum (ou saut maximum ou *maximum linkage*)

$$D_{\max}(A,B) = \max\{d(i,i'),\ i \in A \text{ et } i' \in B\},$$

— critère de la distance moyenne (ou UPGMA [1])

$$D_{\text{moy}}(A,B) = \frac{1}{n_A \cdot n_B} \sum_{\substack{i \in A \\ i' \in B}} d(i,i').$$

où n_E représente le cardinal de l'ensemble E. Remarquons que les hiérarchies fournies par les deux premiers critères ne dépendent que des valeurs extrêmes des distances entre éléments de A et B et sont donc sensibles à la présence de données aberrantes. Le critère du lien minimum tend à favoriser les classes oblongues (effet de chaîne) alors que le lien maximum favorisent les classes compactes.

5.3 Formule de récurrence de Lance et Williams

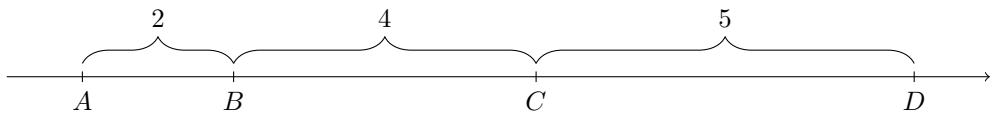
Pour les trois critères d'agrégation précédents, il existe des relations de simplification (Lance and Williams, 1967) du calcul des distances entre classes essentielles pour la mise en place pratique de l'algorithme de CAH qui, sans cette relation, serait prohibitive en temps de calcul. Ces relations appelées généralement formules de récurrence de Lance et Williams, sont les suivantes.

Proposition 8. *Pour les trois critères d'agrégation du saut minimum, du saut maximum et de la moyenne, on peut calculer la distance entre les deux classes A et $B \cup C$ uniquement à partir des distances entre A et B et entre A et C :*

$$\begin{aligned}
D_{\min}(A, B \cup C) &= \min\{D_{\min}(A,B), D_{\min}(A,C)\},\\
D_{\max}(A, B \cup C) &= \max\{D_{\max}(A,B), D_{\max}(A,C)\},\\
D_{\text{moy}}(A, B \cup C) &= \frac{n_B \cdot D_{\text{moy}}(A,B) + n_C \cdot D_{\text{moy}}(A,C)}{n_B + n_C}.
\end{aligned}$$

5.4 Un exemple

On considère 4 points alignés séparés par les distances 2, 4 et 5 :



1. Unweighted Pair Group Method with Arithmetic mean

On prend comme mesure de dissimilarité entre ces points la distance euclidienne habituelle et on effectue la CAH suivant les trois critères d'agrégation $D_{\min}$, $D_{\max}$ et D_{moy} (Figure 7.4). Remarquons que dans le dernier cas, on peut obtenir deux solutions différentes suivant que l'on choisit de regrouper les classes $\{A, B\}$ et $\{C\}$ ou les classes $\{C\}$ et $\{D\}$.

5.5 Méthode de Ward

Lorsque l'ensemble Ω à classifier correspond à un nuage de points muni des pondérations $\frac{1}{n}$ dans $\mathbb{R}^p$ muni de la distance euclidienne, le critère d'agrégation le plus utilisé dans cette situation est alors :

$$D(A, B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B),$$

où g_E représente le centre de gravité de l'ensemble E.

L'algorithme de CAH que l'on obtient est souvent connu sous le nom de méthode de Ward (1963).

Il existe aussi dans ce cas une formule de récurrence :

$$D(A, B \cup C) = \frac{(n_A + n_B) \times D(A, B) + (n_A + n_C) \times D(A, C) - n_A \times D(B, C)}{n_A + n_B + n_C}.$$

5.6 Propriétés d'optimalité

Nous avons vu que la notion de hiérarchie indicée est équivalente à la notion d'ultramétrique. La CAH transforme donc une mesure de dissimilarité d initiale en une nouvelle mesure de dissimilarité δ qui possède la propriété d'être ultramétrique. La classification hiérarchique pourrait alors être posée en ces termes : *trouver l'ultramétrique δ la plus proche de d.*

Il reste à munir l'espace des mesures de dissimilarité sur Ω d'une distance. On pourra utiliser, par exemple :

$$\Delta(d, \delta) = \sum_{i,i' \in \Omega} (d(i, i') - \delta(i, i'))^2,$$

ou encore

$$\Delta(d, \delta) = \sum_{i,i' \in \Omega} |d(i, i') - d(i, i')|.$$

Il s'agit malheureusement d'un problème difficile et nous allons maintenant étudier les propriétés d'optimalité des différents algorithmes décrits précédemment.

Hiérarchie du saut minimum

Soit U l'ensemble de toutes les ultramétriques inférieures à la mesure de dissimilarité initiale.

$$\delta \in U \Leftrightarrow \forall i, i' \in \Omega \qquad \delta(i, i') \leq d(i, i').$$

Soit δ_m l'enveloppe supérieure de U. C'est-à-dire la fonction de $\Omega \times \Omega$ dans $\mathbb{R}$ vérifiant :

$$\forall i, i' \in \Omega \quad \delta_m(i, i') = \sup\{\delta(i, i'), \delta \in U\}.$$

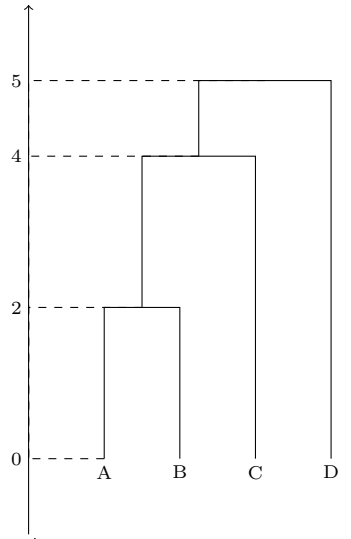
On peut montrer que δ_m est encore une ultramétrique. On l'appelle ultramétrique sous-dominante.

Proposition 9. *Quelque soit Δ la distance entre deux mesures de dissimilarité, l'ultramétrique sous-dominante est l'ultramétrique la plus proche, au sens de Δ, d'une mesure de dissimilarité d parmi toutes les ultramétriques inférieures à d.*

	A	B	C	D
A	0			
B	**2**	0		
C	6	4	0	
D	11	9	5	0

	$\{A,B\}$	C	D
$\{A,B\}$	0		
C	**4**	0	
D	9	5	0

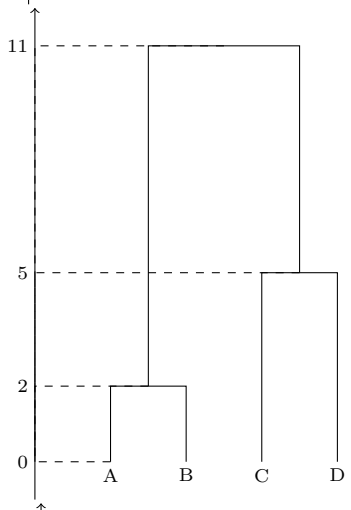
	$\{A,B,C\}$	D
$\{A,B,C\}$	0	
D	**5**	0



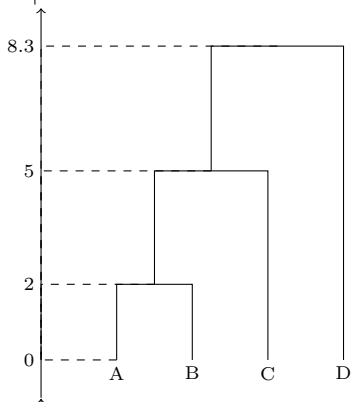
	A	B	C	D
A	0			
B	**2**	0		
C	6	4	0	
D	11	9	5	0

	$\{A,B\}$	C	D
$\{A,B\}$	0		
C	6	0	
D	11	**5**	0

	$\{A,B\}$	$\{C,D\}$
$\{A,B\}$	0	
$\{C,D\}$	**11**	0



	$\{A,B,C\}$	D
$\{A,B,C\}$	0	
D	**8.3**	0



	A	B	C	D
A	0			
B	**2**	0		
C	6	4	0	
D	11	9	5	0

	$\{A,B\}$	C	D
$\{A,B\}$	0		
C	**5**	0	
D	10	**5**	0

	$\{A,B\}$	$\{C,D\}$
$\{A,B\}$	0	
$\{C,D\}$	**7.5**	0

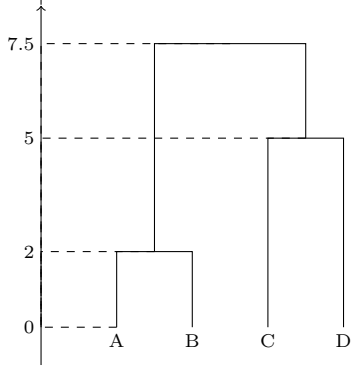


FIGURE 7.4 – CAH avec les critères $D_{\min}$ (haut), $D_{\max}$ (milieu) et D_{moy} (bas).

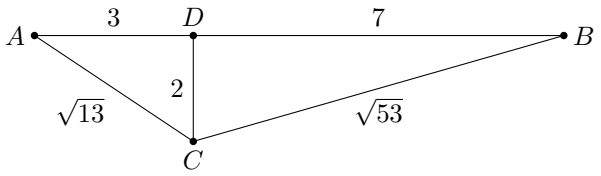
Proposition 10. *L'ultramétrique associée à la hiérarchie indicée obtenue par la CAH avec le critère du saut minimum est l'ultramétrique sous-dominante.*

Cette propriété entraîne le corollaire suivant :

Corollaire 11. *La hiérarchie indicée fournie par la CAH avec le critère du saut minimum est unique.*

Un autre propriété de cette classification hiérarchique est son lien avec la recherche de l'arbre de longueur minimum, problème bien connu en théorie des graphes. On considère le graphe complet défini sur Ω. Chaque arête (a, b) de ce graphe est valuée par la distance $d(a, b)$. On peut montrer que la recherche de l'arbre de longueur minimum de ce graphe est équivalente à la recherche de l'ultramétrique sous-dominante. Pour trouver la hiérarchie du saut minimum, il est possible d'utiliser les algorithmes qui ont été développés pour la recherche de cet arbre de longueur minimum, en particulier les algorithmes de Prim (1957) et de Kruskal. On peut mettre en évidence sur petit exemple :

— les données :

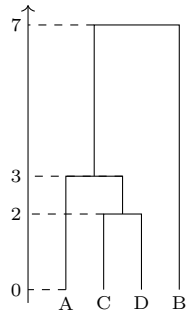


— construction de l'ultramétrique sous-dominante :

	A	B	C
B	10		
C	$\sqrt{13}$	$\sqrt{53}$	
D	3	7	**2**

	A	B
B	10	
$\{C, D\}$	**3**	7

	$\{A, C, D\}$
B	**7**



	A	B	C
B	7		
C	3	7	
D	3	7	2

En ne retenant du graphe complet initial que les 3 arêtes ayant participé à l'algorithme, c'est-à-dire l'arête CD de longueur 2, l'arête AD de longueur 3 et l'arête DB de longueur 7, on obtient l'arbre de longueur minimum.

Ce lien avec l'arbre de longueur minimum permet aussi de mettre en évidence un défaut de ce critère appelé « effet de chaîne ». En effet, deux points situés loin l'un de l'autre peuvent être regroupés ensemble assez tôt dans la hiérarchie s'il existe une chaîne de points les reliant.

Hiérarchie du saut maximum

Cette fois, l'ultramétrique est supérieure à la dissimilarité d. Malheureusement, les propriétés de l'ultramétrique fournie par la CAH ne sont pas aussi intéressantes que celles de l'ultramétrique sous-dominante. En particulier, il n'y a pas nécessairement unicité. Par exemple, on pourra obtenir des résultats différents si on change l'ordre des éléments de Ω.

Remarquons que l'on peut construire de façon parallèle à l'ultramétrique sous-dominante, qui a été définie comme l'enveloppe supérieure des ultramétriques inférieures, l'enveloppe inférieure des ultramétriques supérieures à d. Malheureusement cette enveloppe n'est pas nécessairement une ultramétrique. L'exemple de la figure 7.5 en est un contre-exemple.

d	a	b	c
a	0		
b	1	0	
c	2	1	0

δ_1	a	b	c
a	0		
b	1	0	
c	2	2	0

δ_2	a	b	c
a	0		
b	2	0	
c	2	1	0

FIGURE 7.5 – Distance d et ultramétriques δ_1 et δ_2

On peut vérifier que δ_1 et δ_2 sont deux ultramétriques supérieures à la distance d définie sur les 3 points a, b, c et que l'enveloppe inférieure de ces deux ultramétriques est tout simplement d. Par conséquent, l'enveloppe inférieure de toutes les ultramétriques supérieures à d est nécessairement d qui n'est pas ultramétrique.

Hiérarchie de la moyenne

Elle ne vérifie aucun problème d'optimalité, mais l'expérience a montré qu'elle s'approche de l'ultramétrique minimisant

$$\sum_{i,i' \in \Omega} (d(i,i') - \delta(i,i'))^2.$$

Méthode de Ward

Soit $P = (P_1, \ldots, P_g)$ une partition et P' la partition obtenue à partir de P en fusionnant les classes P_k et P_ℓ. On peut alors montrer le résultat suivant :

$$I_W(P') - I_W(P) = \frac{n_k n_\ell}{n_k + n_\ell} d^2(\overline{\boldsymbol{x}}_k, \overline{\boldsymbol{x}}_\ell).$$

La fusion de deux classes augmentent nécessairement le critère d'inertie intra-classe.

Il est alors possible de proposer l'algorithme de classification ascendante hiérarchique qui fusionne à chaque étape les deux classes augmentant le moins possible le critère d'inertie, c'est-à-dire minimisant l'expression :

$$D(A,B) = \frac{n_k n_\ell}{n_k + n_\ell} d^2(\overline{\boldsymbol{x}}_k, \overline{\boldsymbol{x}}_\ell).$$

On retrouve ainsi tout simplement la méthode de Ward. Cette méthode possède donc une propriété d'optimisation locale : à chaque étape de l'algorithme, on cherche à minimiser le critère d'inertie intra-classe. Toutefois, cet algorithme ne possède aucune propriété globale d'optimisation.

5.7 Utilisation des méthodes

La première difficulté est le choix de la mesure de dissimilarité sur Ω et du critère d'agrégation. Généralement, lorsqu'on dispose de variables quantitatives, le critère conseillé est

FIGURE 7.6 – Données de 10 points dans le plan

le critère d'inertie. Les résultats sont alors utilisables conjointement à ceux de l'ACP. Ensuite, il est souvent nécessaire de disposer d'outils d'aide à l'interprétation et d'outils permettant de diminuer le nombre de niveaux de hiérarchie. Il est d'autre part conseillé d'utiliser conjointement d'autres méthodes d'analyse des données comme l'ACP. Signalons enfin que les problèmes posés par la complexité des algorithmes de CAH en taille et en temps sont résolus en pratique par l'utilisation d'algorithmes plus efficaces comme l'algorithme des voisins réciproques (De Rham, 1980).

6 Recherche de partitions

Ce dernier paragraphe est consacré aux méthodes de partitionnement généralement connus sous le nom de méthode de classification non hiérarchique (*clustering*) et nous commençons par la plus utilisée, la méthode des centres mobiles.

6.1 La méthode des centres mobiles

La méthode des centres mobiles, encore connue sous le nom de méthode de réallocation-centrage ou des *k-means* (MacQueen, 1967) est la méthode de référence lorsque l'ensemble à classifier est mesuré par p variables continues. Dans tout ce paragraphe, l'ensemble Ω à classifier correspond donc à un ensemble de n individus mesurés par p variables quantitatives. Comme nous l'avons vu pour l'ACP, il est alors possible de lui associer un nuage de points dans $\mathbb{R}^p$ muni des pondérations $\frac{1}{n}$ et de la distance euclidienne.

Définition de l'algorithme

On fixe *a priori* le nombre $K \leq n$ de classes que l'on veut identifier. L'algorithme des centres mobiles peut se définir alors de la manière suivante.

1. Tirage au hasard de K points de Ω qui forment les centres initiaux des K classes.
2. Tant que non convergence
   (a) construction de la partition suivante en affectant chaque point de Ω à la classe dont il est le plus près du centre (en cas d'égalité, l'affectation se fait à la classe de plus petit indice).
   (b) les centres de gravité de la partition qui vient d'être calculée deviennent les nouveaux centres.

Si $L = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$ représente un K-uplet de $\mathbb{R}^p$ et $P = (P_1, \ldots, P_K)$ une partition de Ω en K classes, la suite construite par l'algorithme peut être notée sous la forme :

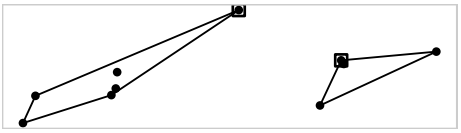
$$L^0 \to P^1 \to L^1 \to P^2 \to L^2 \to \ldots \to P^n \to L^n \to \ldots$$

Exemple

Les données sont constituées d'un ensemble Ω de 10 points du plan décrit à la figure 7.6.

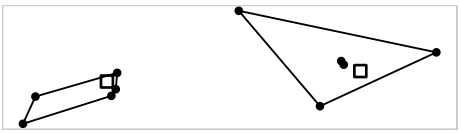
L'algorithme des centres mobiles peut alors se résumer à la suite d'étapes décrites à la figure 7.7. La poursuite de cet algorithme ne changera plus les résultats : l'algorithme a

(a) Étape L^0 : Choix de 2 représentants au hasard



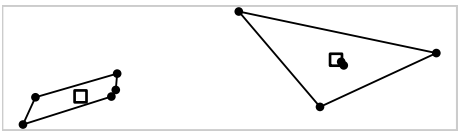
(b) Étape P^1 : Affectation de chaque point au centre le plus proche

(c) Étape L^1 : Calcul des centres de gravité



(d) Étape P^2 : Affectation de chaque point au centre le plus proche

(e) Étape L^2 : Calcul des centres de gravité



(f) Étape P^3 : Affectation de chaque point au centre le plus proche

FIGURE 7.7 – Algorithme des centres-mobiles appliqué aux données de la figure 7.6

convergé. Remarquons que la classification obtenue correspond effectivement à la structure en deux classes observables visuellement. Nous allons maintenant définir et étudier les propriétés de cet algorithme.

Le critère

La qualité d'un couple partition-centres est mesurée par la somme des inerties des classes par rapport à leur centre :

$$C(P, L) = \sum_{k=1}^{K} \mathcal{I}(P_k, \boldsymbol{\lambda}_k) = \frac{1}{n} \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in P_k} d^2(\boldsymbol{x}, \boldsymbol{\lambda}_k),$$

où $P = (P_1, P_2, \ldots, P_K)$ et $L = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$.

Convergence

On peut montrer qu'à chacune des deux étapes de l'algorithme, on améliore le critère C. Plus précisément, on a les relations suivantes.

Proposition 12. *Le calcul d'une nouvelle partition autour des centres fait décroitre le critère :*

$$C(P^{n+1}, L^n) \leq C(P^n, L^n) \tag{7.2}$$

Le calcul des centres étant donnée une partition fait décroitre le critère :

$$C(P^{n+1}, L^{n+1}) \leq C(P^{n+1}, L^n) \tag{7.3}$$

Preuve. Le critère $C(P, L)$ peut s'écrire :

$$C(P, L) = \frac{1}{n} \sum_{\boldsymbol{x} \in \Omega} d^2(\boldsymbol{x}, \boldsymbol{\lambda}_{k(\boldsymbol{x})}),$$

où $k(\boldsymbol{x})$ est le numéro de la classe à laquelle appartient $\boldsymbol{x}$ dans la partition P.

Lorsque l'on compare les expressions $C(P^{n+1}, L^n)$ et $C(P^n, L^n)$, les centres des classes ne bougent pas et comme P^{n+1} est construit en associant chaque point de Ω au meilleur centre, la relation 7.2 est vraie.

Le critère $C(P, L)$ s'écrit aussi :

$$C(P, L) = \sum_{k=1}^{K} \mathcal{I}(P_k, \boldsymbol{\lambda}_k).$$

Par définition de l'algorithme des centres mobiles, L^{n+1} est formée des K centres de gravité des classes de P^{n+1}. Or, la propriété d'optimalité du centre de gravité (voir théorème de Huygens) entraîne l'inégalité

$$\mathcal{I}(P_k^{n+1}, \boldsymbol{\lambda}_k^{n+1}) \leq \mathcal{I}(P_k^{n+1}, \boldsymbol{\lambda}_k^n).$$

L'inéquation 7.3 est donc démontrée. □

Corollaire 13. *La suite numérique $C(P^n, L^n)$ est une suite stationnaire.*

Preuve. Les deux inégalités 7.2 et 7.3 entraînent la décroissante de la suite $C(P^n, L^n)$. Le nombre de partitions en K classes d'un ensemble fini est fini. En outre, l'ensemble contenant les éléments L^n, formés par construction de centres de classes d'un ensemble fini est aussi fini. Par conséquent, la suite $C(P^n, L^n)$ est une suite décroissante qui ne peut prendre qu'un ensemble fini de valeurs. Elle est donc stationnaire. □

Proposition 14. *La suite (P^n, L^n) est une suite stationnaire.*

Remarquons tout d'abord que la stationnarité de $C(P^n, L^n)$ n'entraîne pas forcément la stationnarité de (P^n, L^n). En effet, il serait tout à fait possible d'avoir une suite de partitions et de centres ayant la forme suivante :

$$\ldots, P, L, P', L', P, L, \ldots, P, L, P', L', \ldots$$

avec

$$P \neq P', \quad L \neq L' \quad \text{et} \quad C(P, L) = C(P', L) = C(P, L').$$

Preuve. La suite $C(P^n, L^n)$ est stationnaire. Il existe donc un rang N tel que pour tout $n \geq N$

$$C(P^n, L^n) = C(P^{n+1}, L^{n+1}). \tag{7.4}$$

Or, d'après les relations 7.2 et 7.3, on a

$$C(P^n, L^n) \geq C(P^{n+1}, L^n) \geq C(P^{n+1}, L^{n+1}).$$

Les deux termes extrêmes de l'inégalité sont égaux d'après 7.4, les inégalités sont donc des égalités et on trouve

$$C(P^n, L^n) = C(P^{n+1}, L^n) = C(P^{n+1}, L^{n+1}).$$

Sachant que pour tout k on a nécessairement $\mathcal{I}(P_k^{n+1}, \boldsymbol{\lambda}_k^n) \geq \mathcal{I}(P_k^{n+1}, \boldsymbol{\lambda}_k^{n+1})$ (propriété du centre de gravité), il découle de l'égalité précédente que $\mathcal{I}(P_k^{n+1}, \boldsymbol{\lambda}_k^n) = \mathcal{I}(P_k^{n+1}, \boldsymbol{\lambda}_k^{n+1})$ pour tout k ; et comme le centre de gravité est l'unique point de $\mathbb{R}^p$ minimisant l'inertie de P_k^{n+1}, on obtient $\boldsymbol{\lambda}_k^{n+1} = \boldsymbol{\lambda}_k^n$ et donc $L^{n+1} = L^n$.

Comme par construction, P^n est définie de manière unique à partir de L^n, l'égalité $L^{n+1} = L^n$ entraîne aussi l'égalité $P^{n+1} = P^n$. On a donc bien démontré que la suite (P^n, L^n) est stationnaire. □

Remarques

Finalement, si notre objectif initial avait été de trouver le couple (P, L) minimisant le critère C, l'algorithme des centres mobiles ne fournit pas nécessairement le meilleur résultat, mais simplement une suite de couples dont la valeur du critère va en décroissant. On parle alors d'« optimisation locale ».

Plus précisément, l'algorithme des centres mobiles est un algorithme d'optimisation alternée. En effet, il est facile de montrer que les deux étapes de l'algorithme des centres mobiles vérifie les deux définitions suivantes :

— recherche de la partition : minimisation de $C(P, L)$ avec L fixé ;

— recherche des centres : minimisation de $C(P, L)$ avec P fixée.

En pratique, la convergence est atteinte très vite (souvent moins de 10 itérations même avec des données de taille importante).

Lien avec le critère d'inertie intra-classe

Puisque L^n est fonction de P^n, il est possible d'exprimer le critère $C(P^n, L^n)$ uniquement en fonction de P^n :

$$C(P^n, L^n) = \sum_{k=1}^{K} \mathcal{I}(P_k^n, \boldsymbol{\lambda}_k^n) = \sum_{k=1}^{K} \mathcal{I}(P_k^n),$$

puisque $\boldsymbol{\lambda}_k^n$ est le centre de gravité de la classe P_k^n. Et en conséquence

$$C(P^n, L^n) = \mathcal{I}_W(P^n).$$

Finalement, l'algorithme des centres mobiles défini de manière algorithmique se révèle être un algorithme dont l'objectif est la recherche de la partition en K classes minimisant le critère d'inertie intra-classe.

La méthode des centres mobiles et la méthode de Ward optimisent toutes deux, à leur façon, le critère d'inertie intra-classe. Cette situation conduit à proposer des stratégies utilisant les deux approches, par exemple,

— appliquer les centres mobiles pour regrouper l'ensemble initial en une cinquantaine de classes ;

— appliquer la méthode de Ward en partant de ces classes ;

— rechercher quelques « bons » niveaux de la hiérarchie ;

— éventuellement, appliquer de nouveau la méthode des centres mobiles sur les partitions obtenues pour améliorer encore leur critère.

Variantes de la méthode des centres mobiles

Parmi les nombreuses variantes, on peut citer deux :

— La méthode séquentielle (MacQueen, 1967), qui remet à jour les centres dès qu'un point change de classe :

1. Les K prototypes sont tirés au hasard parmi les n points.
2. A l'itération q, un individu $\boldsymbol{x}_i$ est choisi au hasard.
   — Détermination du prototype le plus proche de $\boldsymbol{x}_i$:

$$\boldsymbol{\lambda}_k^q = \arg\min_{\boldsymbol{\lambda}_j^q} \|\boldsymbol{x}_i - \boldsymbol{\lambda}_j^q\|.$$

   L'individu est affecté à la classe k.

— Modification du prototype $\boldsymbol{\lambda}_k^q$:

$$\boldsymbol{\lambda}_k^{q+1} = \frac{\boldsymbol{x}_i + n_k^q \cdot \boldsymbol{\lambda}_j^q}{n_k^q + 1},$$

et

$$n_k^{q+1} = n_k^q + 1,$$

où n_k^q représente l'effectif de la classe k à l'itération q.

Ce type d'algorithmes séquentiels (encore appelés adaptatifs) est particulièrement adéquat lorsque toutes les données à classer ne sont pas disponibles à l'avance. Les paramètres définissant les classes peuvent alors être ajustés à l'apparition de chaque nouvelle donnée sans trop de calculs.

— En autorisant la fusion et la division de classes, la méthode Isodata (Ball and Hall, 1967) évite de fixer le nombre de classes. Signalons, toutefois, que cet algorithme nécessite la donnée de plusieurs paramètres numériques difficiles à régler ce qui ne fait pas réellement avancer le problème.

6.2 Généralisation : la méthode des nuées dynamiques

L'idée de base consiste à remplacer les *centres* $\boldsymbol{\lambda}_k$ qui étaient des éléments de $\mathbb{R}^p$ jouant le rôle de *représentant* ou encore de *noyau* de la classe par des éléments de nature très diverse adaptés au problème que l'on cherche à résoudre.

Formalisation

On notera

— $\mathbb{L}$ l'ensemble des noyaux,

— $D : \Omega \times \mathbb{L} \to \mathbb{R}^+$, une mesure de ressemblance entre éléments de Ω et de $\mathbb{L}$.

L'objectif est alors de trouver la partition en K classes (K fixé a priori) de Ω minimisant le critère

$$C(P, L) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in P_k} D(\boldsymbol{x}, \boldsymbol{\lambda}_k),$$

où $P = (P_1, \ldots, P_K)$ et $L = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$ avec $\boldsymbol{\lambda}_k \in \mathbb{L}$.

Pour ceci, on utilise l'algorithme suivant.

Algorithme

Il s'agit, comme pour les centres mobiles, d'un algorithme d'optimisation alternée qui définit la suite

$$L^0 \to P^1 \to L^1 \to P^2 \to L^2 \to \ldots \to P^n \to L^n \to \ldots$$

à partir d'un élément L^0 initial quelconque et à l'aide des deux étapes suivantes :

1. P^{n+1} est obtenue en minimisant $C(\cdot, L^n)$
2. L^{n+1} est obtenue en minimisant $C(P^{n+1}, \cdot)$.

Les conditions d'existence de cet algorithme portent uniquement sur la seconde étape. En effet la première, simple à construire est strictement la même que dans le cas des centres mobiles. Par contre, la seconde étape dépend des situations particulières.

Convergence

Dans tous les cas, on peut montrer que la suite des critères est stationnaire. Quant à la stationnarité de la suite (P^n, L^n), cela dépendra de l'étape (2). Si, comme dans le cas des centres mobiles, il y a unicité, alors on obtient les mêmes résultats.

Voici quelques exemples d'application de cette méthode.

Centres mobiles

Si Ω est inclus dans $\mathbb{R}^p$, $\mathbb{L}$ est l'espace $\mathbb{R}^p$ et $D(\mathbf{x}, \boldsymbol{\lambda}) = d^2(\boldsymbol{x}, \boldsymbol{\lambda})$ où d est la distance euclidienne, on retrouve alors simplement la méthode des centres mobiles.

Tableau de dissimilarités

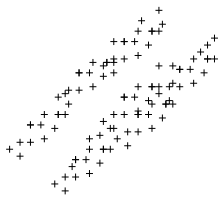
On suppose cette fois que l'on ne connaît sur Ω qu'une mesure de dissimilarité d. On peut alors proposer la situation suivante : $\mathbb{L} = \Omega$ et $D(\boldsymbol{x}, \boldsymbol{\lambda}) = d(\boldsymbol{x}, \boldsymbol{\lambda})$.

Cela permet de proposer une méthode de classification adaptée à la seule donnée d'un tableau de distance. Remarquons que par analogie avec le critère d'inertie, il est souvent préférable de prendre la distance au carré.

Distances adaptatives

$\Omega \subset \mathbb{R}^p$, $\mathbb{L} = \mathbb{R}^p \times D$ où D est l'ensemble de distances quadratiques définies sur $\mathbb{R}^p$ et $D(\boldsymbol{x}, (\mathbf{a}, d)) = d(\boldsymbol{x}, \mathbf{a})$.

Dans cette méthode, on associe à chaque classe comme noyau un centre et une distance, ce qui permet de prendre en compte la forme de la classe et de pouvoir traiter, par exemple, les données suivantes :



Il existe de nombreux autres exemples parmi lesquels on peut citer des centres qui peuvent être des lois de probabilité, des axes factoriels, …

6.3 Mise en œuvre

Choix du critère

La première étape, sans doute la plus délicate est la traduction du problème initial de classification en un problème d'optimisation de critère. Généralement, ceci est réalisé à l'aide d'une mesure de similarité ou de dissimilarité. Comme nous l'avons vu dans le paragraphe précédent, la méthode des nuées dynamiques se révèle être une bonne approche pour proposer de tels critères.

Choix d'un algorithme d'optimisation

Ayant choisi un critère, il faut disposer d'un algorithme d'optimisation. La première solution à laquelle on peut penser est l'énumération de toutes les partitions. Malheureusement le nombre de partitions devient vite extrêmement grand et rend cette solution impraticable.

Le plus souvent, il est impossible de trouver un algorithme fournissant un optimum global. On utilise alors un algorithme d'optimisation locale, par exemple les centres mobiles ou, plus généralement, la méthode des nuées dynamiques. Il existe aussi l'« algorithme d'échange » et l'« algorithme des transferts », qui peuvent s'appliquer à n'importe quel critère : à partir d'une partition initiale, le critère est amélioré en transférant un point d'une classe à une autre, l'algorithme s'arrêtant lorsqu'aucun transfert ne peut améliorer le critère.

Remarquons qu'il existe quelques situations pour lesquelles on dispose d'algorithme efficace permettant de trouver l'optimum global. C'est le cas lorsqu'il y a une contrainte d'ordre sur les partitions. Cette contrainte peut être implicite (par exemple avec le critère de l'inertie sur des données dans $\mathbb{R}$) ou explicite (contrainte imposée par l'utilisateur). On peut alors utiliser un algorithme de programmation dynamique, par exemple l'algorithme de Fisher qui fournit alors l'optimum global.

Exploitation des optima locaux

Sachant que suivant les points de départ choisis, les résultats seront différents, il reste à exploiter ces différents résultats. Plusieurs solutions ont été proposées : On fait différents essais de l'algorithme en tirant au hasard plusieurs initialisations. Plusieurs stratégies sont alors possibles. Soit retenir la meilleure partition, c'est-à-dire celle qui optimise le critère, soit utiliser l'ensemble des résultats pour en déduire les groupes stables (« méthode des formes fortes ») ; On sélectionne une « bonne » initialisation à l'aide d'informations supplémentaires ou à l'aide d'une procédure automatique (points les plus éloignés les uns des autres, zones de forte densité…). Il faut toutefois faire un compromis entre le temps nécessaire à la recherche de la configuration initiale et celui nécessaire à l'algorithme proprement dit ; Il est aussi possible d'utiliser un certain nombre de méthodes stochastiques comme le recuit simulé qui, sans garantir l'optimum global, possèdent des propriétés de convergence asymptotique.

Nombre de classes

En général, le critère n'est pas indépendant du nombre de classes. Par exemple, la partition en n classes où chaque point forme une classe a un critère d'inertie intra-classe nul et est donc, de ce point de vue, la partition optimale ce qui est sans intérêt. Il est donc nécessaire de fixer *a priori* le nombre de classes. Si ce nombre de classes n'est pas connu, plusieurs solutions permettant de résoudre ce problème très difficile sont utilisées. Par exemple, on recherche la meilleure partition pour plusieurs nombres de classes et on étudie la décroissance du critère en fonction du nombre de classes pour sélectionner le nombre de classes (« méthode du coude »). Une autre procédure consiste à pénaliser le critère de classification par une fonction dépendant du nombre de classes rendant ainsi le critère « indépendant » de ce nombre de classes. Il est aussi possible d'ajouter des contraintes supplémentaires portant, par exemple sur le nombre d'individus par classe ou sur le volume d'une classe. C'est l'option retenue par la méthode Isodata. D'autres approches enfin utilisent les tests statistiques.

7 Comparaison de partitions

Pour comparer les résultats de différentes classifications, il faut pouvoir comparer des partitions entre elles. Lorsqu'il y a le même nombre de classes dans les deux partitions et que la correspondance entre les deux est bien identifiée, la proximité est facile à calculer. En revanche, dans les cas moins simples où le nombre de classes diffèrent ou qu'il n'y a pas de bijection claire entre les classes des deux partitions, on utilise généralement l'indice de Rand qui repose sur du dénombrement de paires d'éléments.

Soit Ω de cardinal n et P, Q deux partitions de Ω. On introduit les quantités suivantes :

— a : nombre de paires d'éléments qui font partie d'une même classe dans une partition comme dans l'autre

— d : nombre de paires d'éléments qui sont dans des classes distintes dans une partition comme dans l'autre

— b : nombre de paires d'éléments qui font partie d'une même classe dans la partition P mais qui sont dans des classes distinctes dans la partition Q

— c : nombre de paires d'éléments qui font partie d'une même classe dans la partition Q mais qui sont dans des classes distinctes dans la partition P

On a ainsi $a + b + c + d = \frac{n(n-1)}{2}$.

Définition 15 (Indice de Rand)**.** *L'indice de Rand (Rand index) de deux partitions P et Q, noté* $\mathrm{RI}(P, Q)$ *est défini par*

$$\mathrm{RI}(P, Q) = \frac{a + d}{a + b + c + d} = 2\frac{a + d}{n(n-1)}.$$

Il s'agit de la proportion de paires classées de la même manière dans P et Q sur le nombre de paires totales. L'indice de Rand est donc compris entre 0 et 1 et il vaut 1 ssi $P = Q$.

Exemple 7.4. *Soit* $\Omega = \{A, B, C, D\}$ *et*

$$P = \{\{A, B\}, \{C, D\}\},$$
$$Q = \{\{A\}, \{B, C, D\}\},$$

deux partitions de Ω*. Pour calculer l'indice de Rand entre P et Q, on énumère d'abord les 6 paires d'éléments distincts :*

$$\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}.$$

On trouve $a = 1$ car seule la paire $\{C, D\}$ appartient à un même ensemble dans les deux partitions et $d = 2$ car les paires $\{A, C\}$ et $\{A, D\}$ appartiennent à deux ensembles distincts dans les deux partitions. On a donc

$$\mathrm{RI}(P, Q) = 2 \cdot \frac{3}{4 \cdot 3} = \frac{1}{2}.$$

Les inconvénients majeurs de l'indice de Rand est d'une part qu'il n'est pas nul en moyenne pour deux partitions prises « au hasard » et d'autre part, lorsque le nombre de partitions augmente, l'indice tend asymptotiquement vers 1 même si les partitions sont très différentes.

Exemple 7.5. *Soit* $\Omega = \{1, \ldots, 2n\}$ *et*

$$P_n = \{\{1, 2\}, \{3, 4\}, \ldots, \{2n-1, 2n\}\},$$
$$Q_n = \{\{2, 3\}, \{4, 5\}, \ldots, \{2n-2, 2n-1\}, \{1, 2n\}\},$$

deux partitions de Ω*. On trouve $a = 0$, $d = \frac{2n(2n-1)}{2} - 2n$ donc*

$$\mathrm{RI}(P_n, Q_n) = \frac{\frac{2n(2n-3)}{2}}{\frac{2n(2n-1)}{2}} \xrightarrow{n \to +\infty} 1.$$

L'indice de Rand ajusté permet de remédier à ce problème. Il s'agit d'un recalage de l'indice de Rand, motivé par des considérations statistiques.

Définition 16 (Indice de Rand ajusté)**.** *L'indice de Rand ajusté (ajusted Rand index) de deux partitions P et Q, noté* $\mathrm{ARI}(P, Q)$ *est défini par*

$$\mathrm{ARI}(P, Q) = \frac{\binom{n}{2}(a + d) - ((a + b)(a + c) + (c + d)(b + d))}{\binom{n}{2}^2 - ((a + b)(a + c) + (c + d)(b + d))}.$$

Si on supprime le terme de droite au numérateur et au dénominateur, on retrouve l'indice de Rand classique. À cause de ce recalage on perd la positivité de l'indice de Rand. En revanche, on a toujours $P = Q$ ssi l'indice vaut 1.

Exemple 7.6. *En reprenant les données de l'exemple 7.4, on trouve* $b = 1$ *et* $c = 2$ *d'où* $\mathrm{ARI}(P, Q) = 0$.

Exemple 7.7. *En reprenant les données de l'exemple 7.5, on trouve* $b = c = n$ *et donc*

$$\begin{aligned}\mathrm{ARI}(P_n, Q_n) &= \frac{\frac{2n(2n-1)}{2}\frac{2n(2n-3)}{2} - \left(n^2 + \frac{(2n(2n-2))^2}{4}\right)}{\left(\frac{2n(2n-1)}{2}\right)^2 - \left(n^2 + \frac{(2n(2n-2))^2}{4}\right)} \\ &= \frac{(2n-1)(2n-3) - (1 + (2n-2)^2)}{(2n-1)^2 - (1 + (2n-2)^2)} \\ &= -\frac{1}{2(n-1)} \xrightarrow{n\to+\infty} 0.\end{aligned}$$

L'indice de Rand ajusté tend cette fois vers 0 *au lieu de tendre vers* 1. *On remarquera également qu'il est négatif pour tout* $n \geq 2$.

Deuxième partie

Méthodes supervisées

Chapitre 8

Introduction à l'apprentissage supervisé

1 Contenu

1.1 Problématique

Cette partie du cours présente les bases de l'*apprentissage supervisé* et de la *reconnaissance des formes*, ensembles de techniques visant à construire automatiquement des *systèmes de prédiction* à partir d'observations *étiquetées*. Il s'agit d'un domaine qui se situe à l'intersection de la statistique et de l'informatique.

De très nombreux problèmes d'apprentissage peuvent se formaliser de la manière suivante. On considère une population $\mathcal{P}$ d'individus, chacun étant décrit par p *variables explicatives* $X_1, \ldots, X_p$ (également appelées *caractéristiques*, *attributs*, ou *descripteurs*), et une variable Z *à expliquer*, c'est-à-dire dont on souhaite identifier la valeur, connaissant les valeurs prises par les variables explicatives.

Typiquement, les valeurs prises par les p variables explicatives X_j, $j = 1, \ldots, p$ sont connues ou faciles à identifier pour chaque individu de la population, tandis que la valeur de la variable à expliquer Z est difficile ou coûteuse à identifier. Il s'agit donc de déterminer une fonction qui permette d'identifier, pour tout individu, la valeur z prise par Z à partir de l'observation $\boldsymbol{x}$ le décrivant, interprétée comme une instanciation de l'ensemble des attributs (ou variables) $\boldsymbol{X}$ sur l'individu.

Dans le cours, on fera l'amalgame entre l'individu et l'ensemble des valeurs $\boldsymbol{x}$ prises par ses descripteurs (appelé « vecteur d'entrées »). On pourra appeler « exemple » ou « observation » l'ensemble des caractéristiques observées pour un même individu, soit le couple $(\boldsymbol{x}, z)$ [1].

1.2 Classification supervisée

Information de classe Dans la plupart des chapitres de ce cours, Z est une variable qualitative à g modalités ($g \geq 2$) [2], prenant ses valeurs dans un ensemble $\Omega = \{\omega_1, \ldots, \omega_g\}$: elle caractérise le *groupe* (ou *classe*) de l'individu correspondant, décrit par l'ensemble de valeurs $\boldsymbol{x}$.

1. Par abus de langage, l'exemple pourra parfois ne pas inclure l'information de classe, comme lorsqu'on parlera de « classer un exemple ».
2. On parle de classification *binaire* si $g = 2$, et *multiclasse* dès lors que $g \geq 3$.

Supervision On rappellera que la classification non supervisée, étudiée précédemment, consiste à identifier la classe d'un ensemble d'individus sur la base des seules observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Elle nécessite généralement de définir une mesure de similarité qui permette de distinguer des groupes d'individus proches, éloignés des individus des autres groupes.

La classification *supervisée* ou *discrimination* (en statistique, on parlera parfois d'*analyse discriminante*) suppose l'accès à un ensemble des observations *étiquetées*, c'est-à-dire pour lesquelles on connaît la classe. L'objectif est alors d'*entraîner* (ou *apprendre*) un modèle qui permettra de déterminer la classe à partir de ces observations $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. On parle d'apprentissage supervisé, car l'information concernant la variable de sortie est supposée fournie par un expert (ou oracle, car supposé infaillible).

Intuitivement, l'entraînement du modèle repose sur l'exploitation des données étiquetées disponibles : il s'agit en quelque sorte de reproduire le mécanisme d'étiquetage (effectué par l'expert) de manière automatique. L'objectif est de parvenir à construire un modèle performant, capable de *généraliser* la relation observée entre variables explicatives et variable à expliquer à un grand nombre de nouvelles observations.

Modèle En classification supervisée, la notion de *modèle* fait référence à un ensemble $\mathcal{F}$ de fonctions de décision (c'est-à-dire qui permettent de choisir une classe à partir d'un vecteur d'observations). Entraîner (ou « apprendre ») le modèle revient à choisir une des fonctions de cet ensemble. Le choix de la famille de fonction, tout comme la stratégie d'apprentissage (qui consiste souvent à optimiser un certain critère), influent sur les propriétés du modèle appris.

On peut prendre l'exemple du diagnostic médical, où l'on cherche à construire une règle permettant de déterminer la maladie d'un patient à partir de l'observation d'un certain nombre de symptômes. Un modèle possible serait alors l'ensemble de toutes les règles de décision basées sur une série de tests effectués les uns à la suite des autres.

1.3 Autres problèmes d'apprentissage supervisé

Régression Dans le cadre de ce cours, on abordera brièvement la problématique de la *régression* : la variable Z est alors quantitative, et apprendre un modèle revient à déterminer une fonction permettant de lier les entrées à la variable à *expliquer*.

On pourra citer comme exemples la prédiction d'un niveau de pollution (par exemple en termes de concentration d'oxyde d'azote NO et de monoxyde de carbone CO) en fonction d'indicateurs météorologiques, ou le prix d'un appartement à partir des éléments le décrivant.

Notons que l'on peut parfois résoudre un problème de discrimination par des techniques de régression, par exemple en déterminant un score, ou des probabilités d'appartenance aux classes, à partir des attributs descriptifs des individus, le processus de classement étant basé sur cette information quantitative.

Autres problèmes (qui ne seront pas abordés) Ce cours de SY09 ne saurait être exhaustif ; il ne constitue qu'une première approche du domaine de l'apprentissage automatique. Un certain nombre de problèmes, nécessitant un bagage théorique et technique plus important, seront ainsi délibérément laissés de côté. Nous pouvons en mentionner quelques-uns à titre informatif.

La détection de nouveauté a pour objectif d'identifier des individus « atypiques », jugés ainsi par opposition aux individus « normaux » ou « classiques ». Cette thématique peut donc être rapprochée de la discrimination, en ce qu'il s'agit de séparer deux groupes d'individus (normaux et atypiques) les uns des autres. Toutefois, les individus atypiques, rares, sont souvent absents de l'ensemble des données disponibles ou au mieux présents en nombre extrêmement limité ; l'emploi de techniques supervisées est alors très délicat

(les classes étant très déséquilibrées en termes d'effectifs). En conséquence, les méthodes développées présentent souvent des liens forts avec les méthodes non supervisées.

La prédiction de sorties structurées est une thématique extrêmement vaste qui correspond à une très grande diversité de problèmes. On pourra citer notamment l'apprentissage de préférences (lorsqu'on cherche par exemple à prédire un ensemble de classes ordonnées par préférences décroissantes), la prédiction de liens entre entités (comme par exemple lorsqu'on cherche à apprendre la topologie d'un réseau d'interactions entre espèces végétales), ou le traitement du langage naturel (dans lequel il est nécessaire de respecter certaines règles par exemple de syntaxe).

Mentionnons enfin les problématiques d'analyse de données temporelles, où les variables descriptives évoluent avec le temps. Il est alors nécessaire de prendre en compte cet aspect temporel pour l'apprentissage d'un modèle. On peut citer comme exemples l'économétrie, où l'on cherche à expliquer la variation d'indicateurs en fonction d'un contexte économique ; ou encore l'épidémiologie, qui peut viser à étudier l'évolution de l'état de santé d'un ensemble de patients (ou cohorte) sur une durée plus ou moins importante, pouvant aller jusqu'à plusieurs dizaines d'années.

1.4 Autour de la notion de supervision

L'apprentissage supervisé suppose l'accès à des exemples étiquetés, ces informations de classe étant parfaites, c'est-à-dire précises (une étiquette correspond à une seule classe) et certaines (la classe identifiée est toujours la vraie), puisque fournies par un oracle supposé infaillible. Cette hypothèse peut en pratique s'avérer restrictive ou fausse.

On pourrait ainsi imaginer que les données disponibles ne sont que partiellement étiquetées. L'apprentissage semi-supervisé suppose que l'on dispose d'un sous-ensemble de données parfaitement étiquetées, et d'un sous-ensemble non étiqueté. Plus généralement, chaque individu peut être associé à un sous-ensemble de classes plausibles, dont une seule est la vraie [3].

Par ailleurs, les étiquettes disponibles (qu'elles soient précises ou non) peuvent être entachées d'incertitude, la classe identifiée par l'étiquette pouvant ne pas être la vraie. Ces erreurs d'étiquetage peuvent être d'origine humaine (l'expert pouvant être faillible), ou dues à l'emploi d'un classifieur pour étiqueter les données de manière automatique. Dans ce cas, on pourra modéliser (par exemple via une distribution de probabilité sur les classes) et prendre en compte l'incertitude d'étiquetage lors de l'apprentissage du modèle.

2 Formalisation d'un problème d'apprentissage

2.1 Vecteur forme

On considère une population $\mathcal{P}$ d'individus décrits par p caractéristiques $X_1, \ldots, X_p$ et une variable Z à prédire. On note $\boldsymbol{X} = (X_1, \ldots, X_p)$ le vecteur des attributs, encore appelé *vecteur forme*. La plupart du temps, nous supposerons que les variables X_j sont des variables quantitatives à valeurs dans $\mathbb{R}$: le vecteur $\boldsymbol{X}$ prend donc ses valeurs dans $\mathcal{X} = \mathbb{R}^p$ appelé *espace de représentation* ou *espace des caractéristiques*, dans lequel un point représente un individu de la population (voir figure 8.1).

En discrimination, la variable Z prend ses valeurs dans un ensemble fini $\Omega = \{\omega_1, \ldots, \omega_g\}$, appelé ensemble des classes. En régression, l'espace de la variable dépendante Z est généralement continu — par exemple, $Z \in \mathbb{R}$.

Souvent, on interprétera $\boldsymbol{X}$ et Z comme un vecteur et une variable aléatoires : un problème d'apprentissage peut alors être formalisé comme un problème de modélisation

3. Il faut distinguer ce dernier cas de l'apprentissage multi-étiquettes, où l'on cherche à prédire un sous-ensemble de classes pour chaque individu (comme par exemple lorsqu'on cherche à identifier tous les éléments apparaissant sur une image).

d'une distribution de probabilité. On distinguera en particulier les approches dites *génératives*, où l'on cherche à modéliser la distribution jointe du couple $(\boldsymbol{X}, Z)$, des approches *prédictives* où l'on modélise la distribution conditionnelle de $Z|\boldsymbol{X}$.

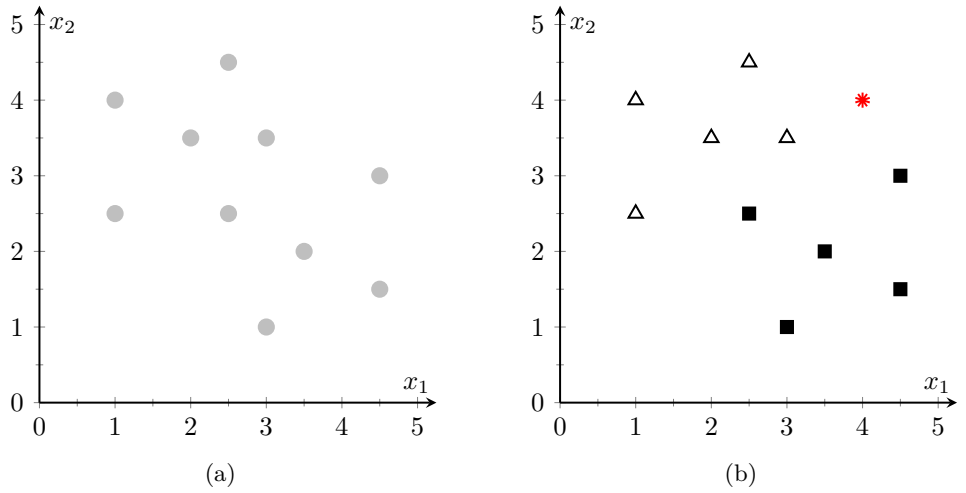


FIGURE 8.1 – Contrairement à la classification non supervisée (8.1a), visant à déterminer une partition des individus observés, la classification supervisée (8.1b) vise à construire un modèle de classement des individus à partir d'un ensemble étiqueté.

2.2 Modèle

Règle de décision L'objectif de la discrimination est de construire une *règle de décision* δ (que l'on appellera aussi *classifieur*), définie comme une fonction de $\mathbb{R}^p$ dans un ensemble d'actions $\mathcal{A}$. Le plus souvent, $\mathcal{A}$ est un ensemble à g éléments $\{a_1, \ldots, a_g\}$, où a_k est l'action d'affectation à la classe ω_k : $\delta(\boldsymbol{x}) = a_k$ signifie que la règle δ prend la décision d'affecter l'individu $\boldsymbol{x}$ à la classe ω_k, qui est donc la valeur de la variable Z prédite par la règle [4].

D'un point de vue géométrique, une règle de décision détermine une partition de l'espace de représentation $\mathbb{R}^p$ en *régions de décision* $\mathcal{R}_k$, séparées par des *frontières de décision*. Il est en général possible de définir une infinité de règles de décision. On peut alors tout d'abord choisir un modèle $\mathcal{F}$, de manière à restreindre le choix d'une règle à un sous-ensemble, ainsi qu'un critère définissant une « bonne » règle : tous deux déterminent les propriétés de la fonction de décision désirée. On pourra alors déterminer la règle $\widehat{\delta} \in \mathcal{F}$ qui optimise le critère choisi.

Par exemple, pour le problème binaire de la figure 8.1b, on pourra rechercher une frontière de décision linéaire entre les deux classes, et choisir cette frontière linéaire de manière à commettre le moins d'erreurs de classement possible sur des individus pris au hasard dans la population $\mathcal{P}$.

Fonction de régression En régression, on cherche de même à construire une fonction $f : \mathbb{R}^p \to \mathbb{R}$, que l'on pourra appeler fonction de régression. Quand bien même on se restreint à un modèle $\mathcal{F}$ (par exemple, l'ensemble des fonctions linéaires), il peut exister une infinité de fonctions de régression, de même qu'en classification.

Là encore, il faudra se doter d'un critère permettant de définir une « bonne » fonction de régression. On pourra définir un critère d'erreur E, comme par exemple l'écart quadra-

4. On peut parfois considérer des actions supplémentaires de *rejet*, qui consistent à « choisir de ne pas affecter » $\boldsymbol{x}$ à l'une des classes de Ω (par exemple lorsqu'aucune classe n'est vraisemblable, ou si au contraire plusieurs le sont). Ces actions de rejet sortent du cadre de SY09, dans lequel nous identifierons une action à une classe.

tique moyen entre la sortie $f(\boldsymbol{x})$ du modèle et la valeur z désirée pour tous les individus d'un ensemble. Apprendre le modèle reviendra alors à déterminer la fonction $\widehat{f} \in \mathcal{F}$ minimisant ce critère d'erreur (comme par exemple la fonction linéaire minimisant l'erreur quadratique moyenne sur les individus d'apprentissage).

3 Apprentissage

3.1 Ensembles d'apprentissage et de test

Pour construire une règle de décision ou une fonction de régression, on dispose typiquement de deux types d'informations : des *connaissances a priori* sur le domaine d'application, qui permettent de déterminer p attributs X_j susceptibles d'apporter une information sur la classe ; et des *données statistiques* relatives à N individus pour lesquels on connaît à la fois les valeurs des p attributs et la celle prise par la variable à expliquer.

Ces données devront être séparées en un *ensemble d'apprentissage*, dont les n éléments sont appelés *exemples d'apprentissage*, que l'on pourra noter $\mathcal{L} = \{(\boldsymbol{x}_i, z_i), i = 1, \ldots, n\}$, et qui sera utilisé pour apprendre le modèle de prédiction ; et un *ensemble de test* $\mathcal{T} = \{(\boldsymbol{x}_i, z_i), i = n+1, \ldots, n+n_t\}$, dont on utilisera les n_t éléments pour tester les performances du modèle. On supposera généralement que ces ensembles sont des échantillons iid de la population de référence $\mathcal{P}$.

La performance d'une règle de décision sera typiquement évaluée par la proportion de points de l'ensemble de test $\mathcal{T}$ mal classés par le classifieur ; nous appellerons cette quantité *taux d'erreur de test*. Nous verrons plus tard que le taux d'erreur de test calculé sur un ensemble $\mathcal{T}$ fixé est une estimation du taux d'erreur théorique du classifieur. La qualité d'une fonction de régression pourra être mesurée par la valeur prise par un critère d'erreur, mesurant par exemple l'écart moyen entre les valeurs prédites par la fonction sur les exemples de test et les valeurs de la variable Z observées pour ces exemples.

Soulignons qu'il est *primordial* d'apprendre le modèle et d'évaluer ses performances avec des individus distincts. Dans le cas contraire, l'estimation de la performance du modèle appris serait *biaisée* en faveur de ce dernier (c'est-à-dire optimiste).

3.2 Méthodologie

La méthodologie de construction d'une règle de décision ou d'une fonction de régression pour un problème donné comporte les étapes suivantes :

1. collecte des ensembles d'apprentissage et de test ;
2. choix des p attributs $X_1, \ldots, X_p$;
3. choix d'un modèle, défini ici comme une famille (un ensemble) $\mathcal{F}$ de règles de décision ou de fonctions de régression ;
4. apprentissage du modèle : choix d'une règle $\widehat{\delta} \in \mathcal{F}$ ou d'une fonction $\widehat{f} \in \mathcal{F}$;
5. évaluation des performances de $\widehat{\delta}$ ou de $\widehat{f}$.

Il s'agit d'un processus dans lequel on peut être amené à revenir à des étapes antérieures. Par exemple, si les performances d'une règle de décision ne sont pas satisfaisantes compte tenu du cahier des charges de l'application, il peut être nécessaire d'augmenter la taille de l'ensemble d'apprentissage, de définir de nouvelles variables explicatives, ou de considérer une famille plus riche de règles de décision.

Rappelons qu'il est crucial de n'utiliser l'ensemble de test qu'à la seule fin d'évaluer les performances du modèle. Pour comparer plusieurs modèles, ou sélectionner un sous-ensemble d'attributs, il sera nécessaire de comparer les performances dans chaque cas de figure au moyen de données réservées à cet effet (données de *validation*), ou d'une procédure excluant les données de test (par exemple par *validation croisée*). Le chapitre 13 constitue une introduction à cette problématique.

Soulignons enfin que les étapes 1 et 2 sont très spécifiques au domaine d'application [5], et reposent généralement sur des connaissances a priori et sur des contraintes pratiques. De même, le choix du modèle pourra être guidé par des connaissances « expertes » spécifiques au domaine. En apprentissage machine, variables et exemples sont généralement imposés à l'utilisateur qui ne dispose que d'un jeu de données. Les autres étapes sont génériques et font appel à des outils algorithmiques et statistiques que nous étudierons dans le cours.

4 Difficultés

La construction d'un « bon » modèle se heurte à un certain nombre de difficultés, liées à la complexité du modèle, à celle de l'espace des individus considéré, et à la quantité de données disponible.

4.1 Compromis entre complexité et robustesse

Il est évident qu'un modèle $\mathcal{F}$ exagérément simple ne sera pas capable de modéliser les variations de la variable à prédire Z en fonction des valeurs de $\boldsymbol{x}$ observées. Avec un modèle suffisamment complexe, il sera généralement possible d'expliquer parfaitement les variations de cette variable *observées sur l'ensemble d'apprentissage*, c'est-à-dire de construire une fonction de décision δ telle que $\delta(\boldsymbol{x}_i) = z_i$, ou de régression f telle que $f(\boldsymbol{x}_i) = z_i$, pour tout $(\boldsymbol{x}_i, z_i) \in \mathcal{L}$.

Sur-apprentissage Cela ne signifie pas pour autant que la fonction apprise sera capable de prédire la « bonne » valeur de Z pour tout individu $\boldsymbol{x}$ nouvellement observé. Le risque d'une sur-adaptation aux exemples d'apprentissage, phénomène appelé « sur-apprentissage » (ou « *overfitting* ») est de s'éloigner du modèle optimal permettant d'expliquer la variable Z sur l'ensemble de la population $\mathcal{P}$.

Plus formellement, considérons le cas de la régression, et choisissons le critère d'*erreur quadratique espérée* pour mesurer l'adéquation entre le modèle appris $\widehat{f}$ et le vrai modèle $f^\star$; ce critère peut s'exprimer comme une somme de deux termes :

$$\mathbb{E}\left[\left(\widehat{f} - f^\star\right)^2\right] = B\left(\widehat{f}\right)^2 + \mathrm{Var}\left(\widehat{f}\right).$$

Le biais $B(\widehat{f}) = \mathbb{E}[\widehat{f}] - f^\star$ du modèle $\widehat{f}$ représente l'écart entre le modèle espéré $\mathbb{E}[\widehat{f}]$ et le « vrai » modèle $f^\star$, et sa variance $\mathrm{Var}(\widehat{f})$ l'écart (quadratique) moyen entre le modèle appris $\widehat{f}$ et le modèle espéré $\mathbb{E}[\widehat{f}]$. Intuitivement, un modèle simple a tendance à avoir un biais élevé, mais une variance faible ; tandis qu'un modèle complexe, plus flexible, est susceptible d'être proche du modèle optimal (donc d'avoir un biais faible), mais sa sensibilité aux données se traduit par une variance plus élevée, un changement dans l'ensemble d'apprentissage pouvant influer grandement sur le modèle appris.

En apprentissage supervisé, l'objectif est donc de déterminer un modèle qui présente un bon compromis entre biais et variance (ou entre simplicité et adéquation aux données d'apprentissage), de manière à garantir de bonnes capacités de généralisation à de nouvelles données, en évitant le phénomène de sur-apprentissage.

Exemple 8.1 (Régression polynomiale)**.** *Considérons les données représentées dans la figure 8.2, qui représentent des mesures d'énergie cinétique en fonction de la vitesse.*

Quatre modèles de régression ont été appris à partir de cet ensemble d'apprentissage : un modèle linéaire, un modèle quadratique « restreint » (ne comprenant qu'un unique terme quadratique), un modèle quadratique « complet » (c'est-à-dire comprenant un terme linéaire et un terme quadratique), et un modèle polynomial d'ordre 5 « complet » ; tous

5. Il est évidemment possible de définir les variables descriptives avant de collecter les données.

ont un terme d'ordonnée à l'origine nul. Le vrai modèle est évidemment un modèle quadratique restreint, d'équation $E_c = mv^2/2$, la masse étant égale à $m = 1/2$ pour les données représentées.

On peut constater que le modèle linéaire n'est pas adapté, la variation de E_c en fonction de v observable sur les données étant clairement non-linéaire. Le modèle polynomial d'ordre 5, le plus complexe, est celui qui s'adapte le mieux aux données d'apprentissage. Les deux modèles quadratiques présentent un compromis entre simplicité et adéquation aux données d'apprentissage ; remarquons que le modèle quadratique complet est légèrement plus proche des données, mais le modèle restreint est le plus proche du vrai modèle.

Fléau de la dimension. Évidemment, la quantité de données disponibles a un importance cruciale vis-à-vis des propriétés du modèle appris. En effet, une augmentation de la taille n de l'ensemble d'apprentissage s'accompagnera généralement d'une diminution de la variance de la fonction de décision ou de régression déterminée. Il est donc clair que disposer d'un nombre important de données améliore la précision du modèle appris ; soulignons toutefois que la quantité de données requises pour obtenir de bonnes performances fortement de deux paramètres, liés l'un à l'autre, que sont la dimension de l'espace des caractéristiques et la complexité du modèle.

À modèle fixé, l'apprentissage d'une fonction de décision ou de régression avec de bonnes capacités de généralisation nécessitera d'autant moins de données que l'espace sera simple (c'est-à-dire engendré par un moindre nombre de variables descriptives). Par exemple, estimer une fonction de régression linéaire lorsque $p = 1$ nécessite beaucoup moins de données que lorsque $p = 10$, l'apprentissage nécessitant d'estimer deux paramètres dans le premier cas, 11 dans le second.

Pour cette raison, il peut être intéressant (et il s'avère parfois nécessaire) de maîtriser le nombre de variables descriptives utilisées pour apprendre un modèle. Certaines stratégies consistent à sélectionner les variables dans une étape préliminaire, par exemple en utilisant des méthodes factorielles similaires à l'ACP. D'autres visent à faire cette sélection après l'apprentissage du modèle, par exemple via des tests statistiques (voir chapitre 11). D'autres enfin incluent la sélection de variables dans la procédure d'apprentissage, par exemple en modifiant le critère optimisé de manière à pénaliser les modèles plus complexes.

4.2 Choix du modèle

À espace fixé, apprendre un modèle simple (comme le modèle linéaire de l'exemple 8.1) nécessitera de même moins de données qu'un modèle complexe. Même si la sensibilité au manque de données dépend de l'algorithme d'apprentissage (donc du critère) considéré, le choix d'un modèle adéquat, c'est-à-dire adapté au problème considéré (en termes de difficulté, mais aussi de quantité de données disponibles par rapport au nombre de variables) est souvent d'une importance cruciale dans la résolution d'un problème d'apprentissage.

Connaiss ance a priori Dans certains cas, une connaissance du problème abordé permet de guider le choix du modèle — ainsi, dans l'exemple 8.1, un argument de bon sens permet de supprimer le terme d'ordonnée à l'origine de chacun des modèles testés, et les lois de la physique indiquent que le modèle de régression adéquat est un modèle polynomial ne comportant qu'un terme quadratique, soit $f(x) = a\,x^2$.

Lorsqu'une telle *connaissance a priori* sur le modèle à employer n'est pas disponible, il pourra être nécessaire de tester différents modèles et de sélectionner celui donnant les meilleurs résultats. Comme nous l'avons dit au paragraphe 3.2, les données de test étant dévolues à l'évaluation des performances du modèle *final*, cette sélection nécessitera d'utiliser d'autres données, en utilisant un ensemble de validation distinct, ou en recourant à des stratégies plus complexes comme la validation croisée.

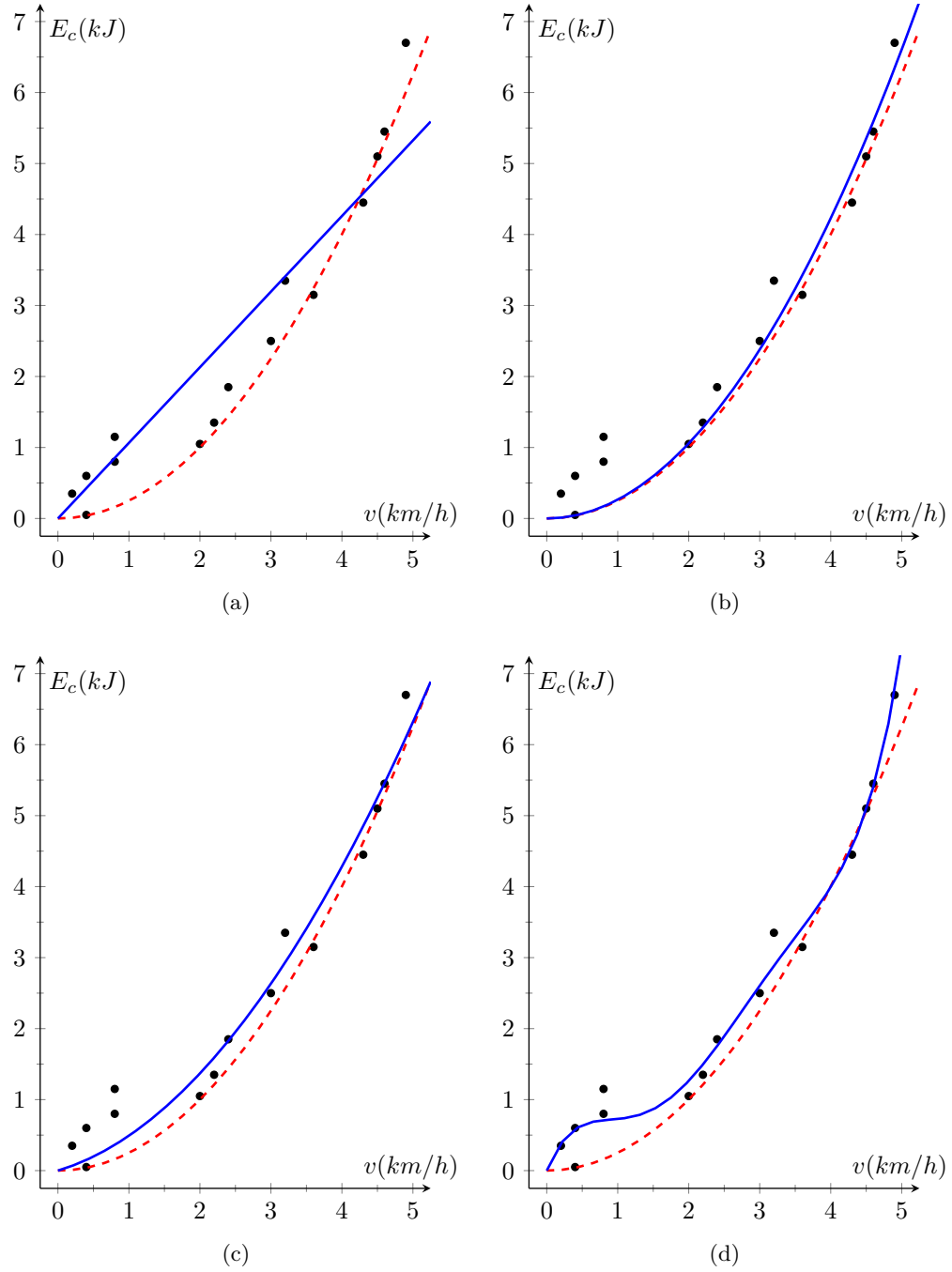


FIGURE 8.2 – Données (simulées) d'énergie cinétique (E_c, en kJ) en fonction de la vitesse (v, en km/h), et modèles de régression appris : linéaire (8.2a), quadratique « restreint » (8.2b), quadratique « complet » (8.2c), polynomial d'ordre 5 « complet » (8.2d), tous sans ordonnée à l'origine. Le vrai modèle est représenté par la courbe rouge discontinue.

Remarquons enfin que cette problématique du choix d'un modèle est liée à celle de la sélection de variables abordée au paragraphe 3.2. En effet, deux fonctions f_1 et f_2 apprises à partir d'ensembles de variables $\mathcal{X}_1$ et $\mathcal{X}_2$ différents correspondent à des modèles $\mathcal{F}_1$ et $\mathcal{F}_2$ différents, et ce même si $\mathcal{X}_1$ et $\mathcal{X}_2$ sont de même taille. Déterminer le meilleur sous-ensemble de variables entre $\mathcal{X}_1$ et $\mathcal{X}_2$ revient donc dans ce cas à sélectionner le meilleur modèle entre f_1 et f_2.

5 Deux classifieurs simples

5.1 Données et prétraitements

On suppose que l'on a observé les valeurs prises par les p variables X_j, $j = 1, \ldots, p$, et la variable de classe Z pour N individus d'une population répartis en $g = 2$ groupes. Ces données sont disposées dans un tableau à N lignes et p colonnes, de terme général x_{ij} (valeur prise par la variable X_j pour le i^{e} exemple ou individu).

Typiquement, on commence par séparer aléatoirement cet ensemble en un ensemble d'apprentissage $\mathcal{L}$ et un ensemble de test $\mathcal{T}$. Le premier (comportant par exemple 2/3 des données) servira à construire la règle de décision et le second à l'évaluer.

Lorsque les variables X_j sont hétérogènes, il est souvent utile de les centrer et de les réduire pour s'affranchir du choix des unités. On supposera ici que les données ont déjà été centrées et réduites pour éviter le recours à de nouvelles notations.

5.2 Classifieur euclidien

Apprentissage Les n vecteurs d'apprentissage $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ peuvent être vus comme n points dans $\mathbb{R}^p$. Ces n points sont répartis en deux nuages correspondant aux deux classes. L'apprentissage du classifieur euclidien repose sur le calcul des centres de gravité $\widehat{\mu}_k = \overline{\boldsymbol{x}_k}$ des nuages correspondant aux classes ω_k, $k = 1, \ldots, g$:

$$\widehat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \boldsymbol{x}_i,$$

où z_{ik} est une variable binaire indiquant l'appartenance à la classe ($z_{ik} = 1$ si $\boldsymbol{x}_i \in \omega_k$, $z_{ik} = 0$ sinon), et où $n_k = \sum_{i=1}^{n} z_{ik}$ est le nombre d'exemples d'apprentissage appartenant à la classe ω_k.

Classement La distance euclidienne entre un vecteur $\boldsymbol{x}$ et $\overline{\boldsymbol{x}_k}$ est

$$d(\boldsymbol{x}, \widehat{\mu}_k) = \|\boldsymbol{x} - \widehat{\mu}_k\| = \left[(\boldsymbol{x} - \widehat{\mu}_k)^T (\boldsymbol{x} - \widehat{\mu}_k)\right]^{1/2}.$$

Le classifieur euclidien consiste à affecter le vecteur $\boldsymbol{x}$ au groupe dont le centre est le plus proche, au sens de la distance euclidienne. La règle de décision est donc la suivante :

$$\delta(\boldsymbol{x}) = \begin{cases} a_1 & \text{si } d(\boldsymbol{x}, \widehat{\mu}_1) \leq d(\boldsymbol{x}, \widehat{\mu}_2), \\ a_2 & \text{sinon,} \end{cases}$$

soit encore, après quelques transformations :

$$\delta(\boldsymbol{x}) = \begin{cases} a_1 & \text{si } (\widehat{\mu}_2 - \widehat{\mu}_1)^T \left(\boldsymbol{x} - \dfrac{\widehat{\mu}_1 + \widehat{\mu}_2}{2} \right) \leq 0, \\ a_2 & \text{sinon.} \end{cases}$$

Cette règle sépare l'espace de représentation en deux régions de décision séparées par une frontière de décision linéaire : c'est l'hyperplan médiateur du segment joignant les centres des deux classes (voir figure 8.3a).

Propriétés Le classifieur euclidien fournit de bons résultats lorsque les nuages de points sont approximativement sphériques et de même volume : il est alors raisonnable de séparer les classes par un hyperplan. Nous verrons au chapitre 10 comment on peut obtenir ce classifieur en choisissant un modèle et un critère d'apprentissage.

Lorsque les nuages de points correspondant aux classes ont des formes quelconques ne pouvant être séparées que par des frontières non linéaires, il est nécessaire de recourir à d'autres familles de règles de décision.

5.3 K plus proches voisins (PPV)

Classement Cette règle de décision consiste à affecter le vecteur $\boldsymbol{x}$ à la classe la plus représentée parmi celles des K plus proches voisins de $\boldsymbol{x}$ (la mesure de proximité étant généralement la distance euclidienne) dans l'ensemble d'apprentissage.

Soient $d_i = d(\boldsymbol{x}, \boldsymbol{x}_i)$ la distance de $\boldsymbol{x}$ à l'exemple $\boldsymbol{x}_i$, $d_{(1)} \leq \ldots \leq d_{(n)}$ les n distances de $\boldsymbol{x}$ aux exemples d'apprentissage triées par ordre croissant, et $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}$ les n vecteurs d'apprentissage triés par ordre de distance croissante à $\boldsymbol{x}$. L'ensemble des K plus proches voisins de $\boldsymbol{x}$ est noté $N_K(\boldsymbol{x}) = \{\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(K)}\}$. En cas d'*ex aequo*, c'est-à-dire si $d_{(K)} = d_{(K+1)}$, on pourra conserver les $K+1$ plus proches voisins de $\boldsymbol{x}$ pour déterminer sa classe (en pratique, ce cas de figure est extrêmement rare).

Comme précédemment, considérons les variables binaires $z_{ik} = 1$ si $\boldsymbol{x}_i \in \omega_k$ et $z_{ik} = 0$ sinon. La règle de décision peut s'écrire :

$$\delta(\boldsymbol{x}) = \arg \max_{k=1,\ldots,g} \frac{1}{K} \sum_{\boldsymbol{x}_i \in N_K(\boldsymbol{x})} z_{ik}.$$

La figure 8.3b illustre la règle des K-PPV sur un exemple.

Choix du modèle Le choix du nombre de voisins K n'est pas un apprentissage à proprement parler, mais un choix de modèle. Il peut être effectué en calculant le taux d'erreur de test pour différentes valeur de K, et en retenant la valeur K^* correspondant au plus petit taux d'erreur. Notons que ce taux d'erreur minimum est une estimation optimiste de la probabilité d'erreur du classifieur, c'est-à-dire de la proportion de mauvais classement lorsque la règle est appliquée à la population totale. Pour obtenir une estimation non biaisée de cette probabilité d'erreur, il faudrait appliquer la règle des K^*-PPV sur un troisième ensemble de données (ensemble de validation) non utilisé pour la construction de la règle de décision.

Remarquons que dans le cas de deux classes, il est préférable de choisir K impair pour éviter la situation d'*ex aequo*. En présence d'*ex aequo* lorsque $g > 2$, on pourra tirer au hasard la classe à laquelle affecter $\boldsymbol{x}$ parmi celles en compétition. Lorsque $K = 1$, la règle des K-PPV devient particulièrement simple : elle consiste à affecter $\boldsymbol{x}$ à la classe de son plus proche voisin dans l'ensemble d'apprentissage. On parle alors parfois de *règle du plus proche voisin*.

La figure 8.4 montre un jeu de données binaire pour lequel les individus des deux classes (en symboles pleins pour ceux de l'ensemble d'apprentissage, en symboles creux pour ceux de l'ensemble de test) ont été générés suivant des lois normales de centres $\mu_1 = (-2, 0)^T$ et $\mu_2 = (2, 0)^T$, et de même matrice de covariance

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}.$$

On verra au chapitre 10 que dans ce cas, la frontière de décision minimisant la probabilité d'erreur est une droite. Il est clair que la frontière de décision obtenue avec l'algorithme du plus proche voisin est beaucoup moins régulière que celle obtenue en considérant 21 voisins, qui se rapproche davantage de la frontière optimale. Dans le premier cas, le sur-apprentissage est manifeste : à cause des deux points d'apprentissage situés dans la zone de l'espace correspondant à la classe adverse, les régions de décision sont multimodales.

Pour cette raison, il est évident que le choix du nombre de voisins ne peut se faire en calculant les performances de classification sur l'ensemble d'apprentissage : cela nécessite l'utilisation d'un ensemble de validation (distinct également de l'ensemble de test).

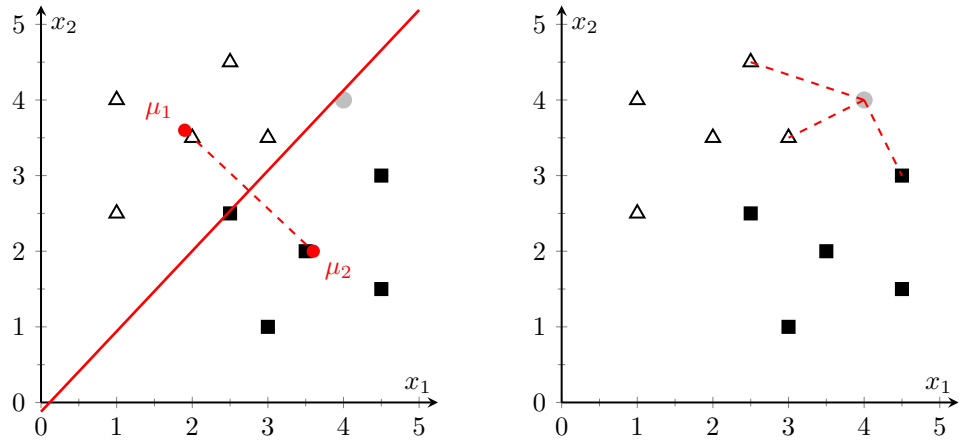
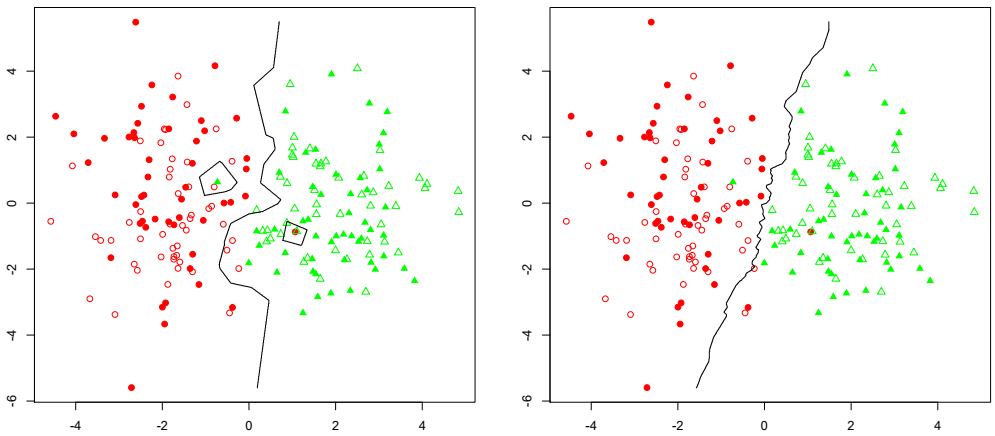


FIGURE 8.3 – Classifieur euclidien (8.3a) : les centres de gravité sont représentés par des astérisques, la frontière de décision par une droite, l'exemple de test est affecté à la classe « carré ») ; classifieur des 3 plus proches voisins (8.3b) : l'exemple de test est affecté à la classe « triangle ».



(a) Classifieur du plus proche voisin (b) Classifieur des 21 plus proches voisins

FIGURE 8.4 – Jeu de données binaire (classes gaussiennes de mêmes matrices de covariance), et frontières obtenues par l'algorithme du plus proche voisin (8.4a) et des 21 plus proches voisins (8.4b).

Chapitre 9

Théorie bayésienne de la décision

1 Introduction

Dans ce chapitre, on suppose que la distribution du vecteur de caractéristiques $\boldsymbol{X}$ dans chaque classe ω_k $(k = 1, \ldots, g)$ est connue. On rappelle qu'une règle de décision est une application $\delta : \mathbb{R}^p \to \mathcal{A}$, où $\mathcal{A} = \{a_1, \ldots, a_c\}$ est un ensemble d'actions. On souhaite trouver une règle de décision *optimale* au sens d'un certain critère.

Pour simplifier les notations, nous supposerons ici que $p = 1$, ce qui revient à ne considérer qu'un seul attribut scalaire. La généralisation au cas $p > 1$ est immédiate.

2 Règle de Neyman-Pearson

2.1 Notations et définition

On se restreint ici au cas $g = 2$. On note $f_k(x) = f(x|\omega_k)$ la densité de probabilité de X conditionnellement au fait que $x \in \omega_k$ (on parlera de densité conditionnelle à ω_k). L'intégrale

$$\int_a^b f_k(x)dx$$

représente la proportion d'individus de la classe ω_k pour lesquels $X \in [a, b]$. On considère seulement deux actions $\mathcal{A} = \{a_1, a_2\}$, a_k correspondant à l'action d'affecter x à la classe ω_k. Soit $\delta : \mathbb{R} \to \mathcal{A}$ une règle de décision, et $\mathcal{R}_k$ $(k = 1, 2)$ les régions de décision correspondantes. On a donc par définition

$$\delta(x) = \begin{cases} a_1 & \text{si } x \in \mathcal{R}_1 \\ a_2 & \text{si } x \in \mathcal{R}_2. \end{cases}$$

Remarquons que toute action a_k étant associée au choix d'une classe ω_k, nous pourrons écrire par la suite, par abus de notation, $\delta(x) = \omega_k$.

On peut caractériser la performance de δ par deux probabilités d'erreur, de première espèce α et de seconde espèce β :

$$\begin{aligned} \alpha &= \mathbb{P}\left(\delta(X) = a_2 | Z = \omega_1\right) = \int_{\mathcal{R}_2} f_1(x)dx, \\ \beta &= \mathbb{P}\left(\delta(X) = a_1 | Z = \omega_2\right) = \int_{\mathcal{R}_1} f_2(x)dx. \end{aligned}$$

En statistique médicale, ω_1 et ω_2 correspondent classiquement aux populations d'invidus sains et malades, respectivement. On définit alors la *sensibilité* de la règle de décision (test) par $1-\beta$, et sa *spécificité* par $1-\alpha$. Il s'agit de trouver une règle de décision représentant un bon compromis entre sensibilité et spécificité.

2.2 Théorème de Neyman-Pearson

On s'intéresse à la règle de décision minimisant β pour $\alpha = \alpha^*$ fixé.

Théorème 1. *La règle de décision qui minimise β sous la contrainte $\alpha = \alpha^*$ est*

$$\delta(x) = \begin{cases} a_1 & si\ \dfrac{f_1(x)}{f_2(x)} > s, \\ a_2 & sinon\,; \end{cases}$$

où s est une constante solution de l'équation

$$\mathbb{P}\left(\frac{f_1(X)}{f_2(X)} \leq s | Z = \omega_1\right) = \alpha^*.$$

Losqu'on augmente le seuil s, α et $1-\beta$ augmentent. La courbe représentant $1-\beta$ en fonction de α est appelée *courbe COR* (caractéristique opératoire du récepteur). Cette courbe est souvent utilisée pour choisir la valeur de α^*, et donc de s. Elle caractérise l'information apportée par X relativement à la classe.

3 Règle minimisant la probabilité d'erreur dans le cas de deux classes

3.1 Probabilités *a priori* et *a posteriori*

Supposons maintenant que, outre les densités conditionnelles $f_k(x)$, on connaisse la proportion $\pi_k = \mathbb{P}(x \in \omega_k)$ de chaque classe ω_k dans la population totale $\mathcal{P}$ (cette proportion sera souvent notée $\pi_k = \mathbb{P}(\omega_k)$). La proportion π_k est appelée *probabilité a priori* de la classe ω_k. On peut alors en déduire la densité de X dans la population totale, appelée *densité de mélange* de X :

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x).$$

La proportion d'individus de la classe ω_k parmi ceux vérifiant $x \leq X \leq x + dx$ peut être calculée à l'aide de la formule de Bayes :

$$\begin{aligned} \mathbb{P}(Z = \omega_k | x \leq X \leq x + dx) &= \frac{\mathbb{P}(x \leq X \leq x + dx | \omega_k)\mathbb{P}(\omega_k)}{\mathbb{P}(x \leq X \leq x + dx)} \\ &= \frac{f_k(x)dx \cdot \pi_k}{f(x)dx} = \frac{f_k(x)\pi_k}{f(x)}. \end{aligned}$$

En faisant tendre dx vers 0, on obtient :

$$\mathbb{P}(Z = \omega_k | X = x) = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\pi_1 f_1(x) + \pi_2 f_2(x)}.$$

Cette quantité, notée $\mathbb{P}(\omega_k|x)$, est appelée *probabilité a posteriori* de ω_k.

3.2 Notion de probabilité d'erreur

Comme précédemment, on considère un ensemble de deux actions $\mathcal{A} = \{a_1, a_2\}$. La règle δ commet une erreur si $\delta(X) = a_\ell$ et $Z = \omega_k$, avec $k \neq \ell$. Notons $\delta(X) \not\equiv Z$ cet

événement. La *probabilité d'erreur* de la règle δ est donc

$$\varepsilon(\delta) = \mathbb{P}(\delta(X) \not\equiv Z) = \int \varepsilon(\delta|x) f(x) dx,$$

où $\varepsilon(\delta|x) = \mathbb{P}(\delta(X) \not\equiv Z|X = x)$ est la probabilité d'erreur de la règle δ conditionnellement à l'événement $X = x$. La probabilité d'erreur peut également s'exprimer en fonction de α et β. En effet,

$$\begin{aligned}
\varepsilon(\delta) &= \int_{\mathcal{R}_1} \varepsilon(\delta|x) f(x) dx + \int_{\mathcal{R}_2} \varepsilon(\delta|x) f(x) dx \\
&= \int_{\mathcal{R}_1} \mathbb{P}(\omega_2|x) f(x) dx + \int_{\mathcal{R}_2} \mathbb{P}(\omega_1|x) f(x) dx \\
&= \int_{\mathcal{R}_1} \pi_2 f_2(x) dx + \int_{\mathcal{R}_2} \pi_1 f_1(x) dx = \pi_2 \beta + \pi_1 \alpha.
\end{aligned}$$

3.3 Minimisation de la probabilité d'erreur : règle de Bayes

On souhaite cette fois trouver la règle de décision minimisant la probabilité d'erreur. Considérons tout d'abord une valeur x fixée. La décision par la règle δ peut prendre deux valeurs possibles :

— si $\delta(x) = a_1$, alors $\varepsilon(\delta|x) = \mathbb{P}(\omega_2|x)$;
— si $\delta(x) = a_2$, alors $\varepsilon(\delta|x) = \mathbb{P}(\omega_1|x)$.

La règle δ^* minimisant $\varepsilon(\delta|x)$ pour x fixé est donc définie par :

$$\delta^*(x) = \begin{cases} a_1 & \text{si } \mathbb{P}(\omega_2|x) < \mathbb{P}(\omega_1|x), \\ a_2 & \text{sinon.} \end{cases}$$

Cette règle δ^*, appelée *règle de Bayes*, minimise $\varepsilon(\delta|x)$ pour tout x : elle minimise donc la probabilité d'erreur $\varepsilon(\delta) = \int \varepsilon(\delta|x) f(x) dx$. Notons qu'elle consiste à affecter chaque individu x à la classe de plus grande probabilité a posteriori.

Remarquons enfin que l'on peut, dans le cas de deux classes, exprimer la règle de Bayes en fonction du rapport de vraisemblance $f_1(x)/f_2(x)$. En effet,

$$\begin{aligned}
\delta^*(x) = a_1 &\Leftrightarrow \mathbb{P}(\omega_1|x) > \mathbb{P}(\omega_2|x) \\
&\Leftrightarrow \frac{f_1(x)\pi_1}{f(x)} > \frac{f_2(x)\pi_2}{f(x)} \\
&\Leftrightarrow \frac{f_1(x)}{f_2(x)} > \frac{\pi_2}{\pi_1}.
\end{aligned}$$

3.4 Probabilité d'erreur de Bayes

La probabilité d'erreur $\varepsilon(\delta^*)$ de la règle de Bayes est appelée *probabilité d'erreur de Bayes*. On la note ε^*. Dans le cas de deux classes, elle est égale à

$$\varepsilon^* = \int \min\left(\mathbb{P}(\omega_1|x), \mathbb{P}(\omega_2|x)\right) f(x) dx.$$

C'est la plus petite erreur possible que peut atteindre une règle de décision utilisant uniquement la variable explicative X.

4 Règle minimisant le risque

4.1 Notion de risque

On suppose toujours $g = 2$ et $\mathcal{A} = \{a_1, a_2\}$, mais on introduit cette fois la notion de *coût* d'une décision. On note $c(a_\ell|\omega_k) = c_{\ell k}$ le coût encouru lorsqu'on choisit l'action a_ℓ alors

que $Z = \omega_k$. C'est donc le coût lié à l'affectation à la classe ω_ℓ d'un individu de la classe ω_k. On souhaite trouver la règle de décision δ minimisant le coût *espéré* ou coût moyen, c'est-à-dire le *risque* défini par

$$r(\delta) = \mathbb{E}_{X,Z}\left[c(\delta(X)|Z)\right] = \int r(\delta|x) f(x) dx,$$

où $r(\delta|x)$ est le risque conditionnel de la règle δ sachant x, défini par

$$\begin{aligned} r(\delta|x) &= \mathbb{E}_{Z|X=x}\left[c(\delta(x)|Z)\right], \\ &= c(\delta(x)|\omega_1)\mathbb{P}(\omega_1|x) + c(\delta(x)|\omega_2)\mathbb{P}(\omega_2|x). \end{aligned}$$

Comme la probabilité d'erreur, le risque peut s'exprimer en fonction de α et β. On a

$$\begin{aligned} r(\delta) &= \int_{\mathcal{R}_1} r(\delta|x) f(x) dx + \int_{\mathcal{R}_2} r(\delta|x) f(x) dx, \\ &= \int_{\mathcal{R}_1} (c_{11}\mathbb{P}(\omega_1|x) + c_{12}\mathbb{P}(\omega_2|x)) f(x) dx + \int_{\mathcal{R}_2} (c_{21}\mathbb{P}(\omega_1|x) + c_{22}\mathbb{P}(\omega_2|x)) f(x) dx, \\ &= c_{11}\pi_1 \int_{\mathcal{R}_1} f_1(x) dx + c_{12}\pi_2 \int_{\mathcal{R}_1} f_2(x) dx + c_{21}\pi_1 \int_{\mathcal{R}_2} f_1(x) dx + c_{22}\pi_2 \int_{\mathcal{R}_2} f_2(x) dx, \end{aligned}$$

soit encore

$$r(\delta) = c_{11}\pi_1 + c_{22}\pi_2 + \pi_2\beta(c_{12} - c_{22}) + \pi_1\alpha(c_{21} - c_{11}). \tag{9.1}$$

4.2 Lien entre risque et probabilité d'erreur

Montrons que le risque introduit ci-dessus généralise la probabilité d'erreur. En effet, posons

$$c_{\ell k} = \begin{cases} 0 & \text{si } k = \ell, \\ 1 & \text{sinon.} \end{cases}$$

On suppose donc que le coût d'une erreur est égal à 1, tandis que le coût d'une bonne décision est nul. D'après l'équation (9.1) précédente, on a alors pour toute règle δ :

$$r(\delta) = \pi_2\beta + \pi_1\alpha = \varepsilon(\delta).$$

La probabilité d'erreur est donc un risque pour un choix particulier des coûts, que nous désignerons par la suite par l'expression « coûts $\{0, 1\}$ ».

4.3 Minimisation du risque

Considérons un x fixé. La décision par la règle δ peut prendre deux valeurs possibles :

— si $\delta(x) = a_1$, alors $r(\delta|x) = c_{11}\mathbb{P}(\omega_1|x) + c_{12}\mathbb{P}(\omega_2|x) = r_1(x)$;

— si $\delta(x) = a_2$, alors $r(\delta|x) = c_{21}\mathbb{P}(\omega_1|x) + c_{22}\mathbb{P}(\omega_2|x) = r_2(x)$.

La règle δ^* minimisant $r(\delta|x)$ pour x fixé est donc définie par :

$$\delta^*(x) = \begin{cases} a_1 & \text{si } r_1(x) < r_2(x) \\ a_2 & \text{sinon,} \end{cases}$$

Cette règle minimise $r(\delta|x)$ pour tout x : elle minimise donc $r(\delta) = \int r(\delta|x) f(x) dx$. La règle δ^* est appelée *règle de Bayes associée aux coûts* $c(a_\ell|\omega_k)$ $(k, \ell \in \{1, 2\})$.

Cette règle peut également s'exprimer, dans le cas de deux classes, en fonction du rapport de vraisemblance $f_1(x)/f_2(x)$. En effet,

$$\begin{aligned} \delta^*(x) = a_1 &\Leftrightarrow r_1(x) < r_2(x) \\ &\Leftrightarrow c_{11}\mathbb{P}(\omega_1|x) + c_{12}\mathbb{P}(\omega_2|x) < c_{21}\mathbb{P}(\omega_1|x) + c_{22}\mathbb{P}(\omega_2|x) \\ &\Leftrightarrow (c_{11} - c_{21})\frac{f_1(x)\pi_1}{f(x)} < (c_{22} - c_{12})\frac{f_2(x)\pi_2}{f(x)} \\ &\Leftrightarrow \frac{f_1(x)}{f_2(x)} > \frac{c_{12} - c_{22}}{c_{21} - c_{11}}\frac{\pi_2}{\pi_1}. \end{aligned}$$

On remarque que, lorsque $c_{11} = c_{22} = 0$, la règle δ^* ne dépend des coûts qu'au travers du rapport des coûts c_{12}/c_{21}. Par ailleurs, on retrouve bien la règle de Bayes minimisant la probabilité d'erreur lorsque $c_{11} = c_{22} = 0$ et $c_{12} = c_{21}$.

4.4 Extension au cas multi-classes

L'extension au cas multi-classes ($g \geq 2$) est immédiate. Soit $\mathcal{A} = \{a_1, \ldots, a_g\}$ l'ensemble des actions, a_k étant comme précédemment interprété comme l'affectation à la classe ω_k, et $c_{\ell k}$ le coût d'affectation à la classe ω_ℓ d'un individu appartenant à la classe ω_k ($\ell, k \in \{1, \ldots, g\}$). Le risque conditionnel si on choisit l'action a_ℓ, ayant observé $X = x$, est :

$$r_\ell(x) = \sum_{k=1}^{g} c_{\ell k} \mathbb{P}(\omega_k | x).$$

La règle minimisant le risque est donc définie par $\delta^*(x) = a_{\ell^*}$ avec

$$\ell^* = \arg\min_{\ell} r_\ell(x).$$

En particulier, dans le cas de coûts $\{0, 1\}$, on a $r_\ell(x) = 1 - \mathbb{P}(\omega_\ell | x)$, et donc

$$\ell^* = \arg\max_{\ell} \mathbb{P}(\omega_\ell | x).$$

La règle de Bayes consiste donc, dans ce cas, à choisir la classe de plus grande probabilité a posteriori.

Chapitre 10

Analyses discriminantes quadratique et linéaire

1 Introduction

On suppose dans ce chapitre que le vecteur de caractéristique $\boldsymbol{X}$ suit, conditionnellement à chaque classe ω_k, une loi normale multidimensionnelle d'espérance $\boldsymbol{\mu}_k$ et de variance Σ_k. En faisant différentes hypothèses sur les paramètres de ces lois (notamment sur les matrices de variance), on obtient différentes expressions de la règle de Bayes, d'où l'on déduit différentes règles de décision en remplaçant les paramètres théoriques par leurs estimations.

2 Analyse discriminante quadratique

2.1 Modèle

Considérons tout d'abord le cas général où la distribution de $\boldsymbol{x}$ dans chaque classe est caractérisée par des paramètres $\boldsymbol{\mu}_k$ et Σ_k différents. On alors

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_k)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right).$$

Pour simplifier, nous nous placerons dans ce chapitre dans le cas des coûts $\{0,1\}$ sans option de rejet. La règle de Bayes s'écrit alors $\delta^*(\boldsymbol{x}) = a_{k^*}$ avec

$$\begin{aligned} k^* &= \arg\max_k \mathbb{P}(\omega_k|\boldsymbol{x}) \\ &= \arg\max_k \pi_k f_k(\boldsymbol{x}) \\ &= \arg\max_k g_k(\boldsymbol{x}), \end{aligned}$$

avec

$$\begin{aligned} g_k(\boldsymbol{x}) &= \ln f_k(\boldsymbol{x}) + \ln \pi_k & (10.1)\\ &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\ln(\det\Sigma_k) + \ln\pi_k - \frac{p}{2}\ln(2\pi). & (10.2)\end{aligned}$$

Les fonctions $g_k(\boldsymbol{x})$ qui servent à définir la règle de décision sont appelées *fonctions discriminantes*. Ici, ce sont des formes quadratiques : on parle de *fonctions discriminantes quadratiques*. Les régions de décision sont séparées par des frontières d'équations $g_k(\boldsymbol{x}) = g_\ell(\boldsymbol{x})$. En dimension quelconque, ces variétés sont des quadratiques (hypersphères, hyperellipsoïdes, hyperparaboloïde, etc.). En dimension 2, ce sont des coniques (cercles, ellipses, paraboles, hyperboles, droites).

2.2 Estimation des paramètres

En pratique, les paramètres π_k, $\boldsymbol{\mu}_k$ et Σ_k du modèles sont inconnus, mais on dispose d'un ensemble d'apprentissage $\mathcal{L} = \{(\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n)\}$, supposé être une réalisation d'un échantillon iid du couple $(\boldsymbol{X}, Z)$. Les estimateurs du maximum de vraisemblance (EMV) des paramètres sont, pour $k = 1, \ldots, g$ (voir paragraphe 6) :

$$\begin{aligned}
\widehat{\pi}_k &= \frac{n_k}{n}, \\
\widehat{\boldsymbol{\mu}}_k &= \overline{\boldsymbol{x}_k} = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \boldsymbol{x}_i, \\
\widehat{\Sigma}_k &= V_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^T,
\end{aligned}$$

où la variable binaire z_{ik} indique l'appartenance à la classe ω_k ($z_{ik} = 1$ si $\boldsymbol{x}_i \in \omega_k$, et $z_{ik} = 0$ sinon), et où $n_k = \sum_{i=1}^{n} z_{ik}$. Notons que si $\overline{\boldsymbol{x}_k}$ est un estimateur sans biais de $\boldsymbol{\mu}_k$, V_k est en revanche biaisé : en pratique, on le remplace souvent par l'estimateur sans biais

$$V_k^* = \frac{n_k}{n_k - 1} V_k.$$

La méthode consistant à remplacer, dans le modèle précédent, les paramètres par leurs EMV (éventuellement corrigés) est appelée *analyse discriminante quadratique* (ADQ).

3 Analyse discriminante linéaire

3.1 Modèle

On suppose cette fois que la matrice de variance est commune à toute les classes (hypothèse d'*homoscédasticité*) : $\Sigma_k = \Sigma$, $k \in \{1, \ldots, g\}$. On a donc

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2} (\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \right).$$

En calculant $\ln(\pi_k f_k(\boldsymbol{x}))$ et en supprimant les termes identiques pour toutes les classes, on obtient les fonctions discriminantes suivantes :

$$g_k(\boldsymbol{x}) = -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \ln \pi_k. \tag{10.3}$$

Le terme $(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$ est le carré de la *distance de Mahalanobis* de $\boldsymbol{x}$ à $\boldsymbol{\mu}_k$. Lorsque les probabilités a priori sont égales, la règle de Bayes avec coûts $\{0, 1\}$ revient donc à affecter l'individu à la classe dont le centre est le plus proche de $\boldsymbol{x}$ au sens de la distance de Mahalanobis.

En développant le membre de droite de (10.3) et en remarquant que le terme quadratique ne dépend pas de k, on obtient les nouvelles fonctions discriminantes suivantes :

$$h_k(\boldsymbol{x}) = \left(\Sigma^{-1} \boldsymbol{\mu}_k \right)^T \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \ln \pi_k.$$

Ces fonctions discriminantes sont linéaires : la règle de Bayes est donc dans ce cas une *règle de décision linéaire*.

Les régions de décision $\mathcal{R}_k^*$ et $\mathcal{R}_\ell^*$ sont séparées par une frontière d'équation :

$$\begin{aligned}
&h_k(\boldsymbol{x}) = h_\ell(\boldsymbol{x}) \\
&\quad \Leftrightarrow \left(\Sigma^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \right)^T \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_\ell}{2} + \frac{\ln(\pi_k / \pi_\ell)}{(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)^T \Sigma^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell) \right) = 0.
\end{aligned}$$

C'est un hyperplan de vecteur normal $\Sigma^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$. Si $\pi_k = \pi_\ell$, cet hyperplan passe par le centre du segment d'extrémités $\boldsymbol{\mu}_k$ et $\boldsymbol{\mu}_\ell$.

3.2 Estimation des paramètres

Les paramètres du modèle sont les π_k, $\boldsymbol{\mu}_k$ $(k = 1, \ldots, g)$ et la matrice de variance Σ commune aux g classes. Les estimateurs du maximum de vraisemblance de ces paramètres sont :

$$\begin{aligned} \widehat{\pi}_k &= \frac{n_k}{n}, \qquad \widehat{\boldsymbol{\mu}}_k = \overline{\boldsymbol{x}_k}, \\ \widehat{\Sigma} &= \frac{1}{n}\sum_{k=1}^{g}\sum_{i=1}^{n} z_{ik}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k)^T \\ &= \frac{1}{n}\sum_{k=1}^{g}(n_k - 1)V_k^* = \frac{1}{n}\sum_{k=1}^{g} n_k V_k, \end{aligned}$$

la matrice de variance intra-classe $\widehat{\Sigma}$ étant quelquefois notée V_W.

On montre que $\mathbb{E}(\widehat{\Sigma}) = (n-g)/n\,\Sigma$. En pratique, on utilise donc plutôt l'estimateur sans biais :

$$V_W^* = \frac{1}{n-g}\sum_{k=1}^{g}(n_k - 1)V_k^*.$$

La méthode consistant à remplacer, dans le modèle précédent, les paramètres par leurs EMV (éventuellement corrigés) est appelée *analyse discriminante linéaire* (ADL).

4 Autres modèles

4.1 Hypothèse d'indépendance conditionnelle

Il est possible de définir plusieurs variantes des modèles précédents en faisant différentes hypothèses sur les matrices de variance. Par exemple, une hypothèse courante consiste à supposer l'indépendance des variables X_j conditionnellement à Z, ce qui, dans le modèle gaussien, revient à supposer les matrices Σ_k diagonales. On parle quelquefois de classifieur bayésien *naïf*. Si l'on fait cette hypothèse, on obtient une variante de l'ADQ dans laquelle les matrices de variance Σ_k sont estimées par la matrice diagonale

$$\widehat{\Sigma_k} = \mathrm{diag}(\mathrm{diag}(V_k)),$$

ce qui revient à annuler, dans la matrice V_k, les termes non diagonaux. La matrice $\widehat{\Sigma_k}$ est donc la matrice diagonale dont le j^{e} élément diagonal est la variance empirique s_{kj}^2 de la variable X^j conditionnellement à la classe ω_k.

On peut également conjuguer cette hypothèse avec celle d'homoscédasticité : dans ce cas, l'estimation de la matrice de variance commune Σ est obtenue en annulant les termes non diagonaux de V_W :

$$\widehat{\Sigma} = \mathrm{diag}(\mathrm{diag}(V_W)) = \frac{1}{n}\sum_{k=1}^{g} n_k\,\mathrm{diag}(\mathrm{diag}(V_k)).$$

4.2 Classifieur euclidien

Il s'agit du modèle le plus simple. On suppose que :

- les matrices de variance sont scalaires et communes à toutes les classes : on a donc $\Sigma_k = \sigma^2 I_p$, où σ^2 est la variance commune des variables X_j conditionnellement à chaque classe, et I_p est la matrice identité d'ordre p ;
- les probabilités a priori sont égales : $\pi_k = 1/g$, $k = 1, \ldots, g$.

TABLE 10.1 – Nombres de paramètres associés aux différents modèles.

Modèle	Nombre de paramètres
ADQ	$g\left(p+\frac{p(p+1)}{2}\right)+g-1$
ADQ avec indépendance conditionnelle	$2gp+g-1$
ADL	$gp+\frac{p(p+1)}{2}+g-1$
ADL avec indépendance conditionnelle	$gp+p+g-1$
Classifieur euclidien	gp

Dans ce cas, les densités conditionnelles ont pour expression

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\sigma^p}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}-\boldsymbol{\mu}_k)\right).$$

En calculant $\ln(\pi_k f_x(\boldsymbol{x}))$ et en supprimant les termes identiques pour toutes les classes, on obtient les fonctions discriminantes suivantes :

$$g_k(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}-\boldsymbol{\mu}_k). \tag{10.4}$$

Le terme $(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}-\boldsymbol{\mu}_k)$ est le carré de la *distance euclidienne* de $\boldsymbol{x}$ à $\boldsymbol{\mu}_k$. La règle de Bayes avec coûts $\{0,1\}$ revient donc dans ce cas à affecter l'individu à la classe dont le centre est le plus proche de $\boldsymbol{x}$, au sens de la distance euclidienne.

En développant le membre de droite de (10.4) et en remarquant que le terme quadratique ne dépend pas de k, on obtient les nouvelles fonctions discriminantes linéaires suivantes :

$$h_k(\boldsymbol{x}) = \boldsymbol{\mu}_k^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\mu}_k.$$

Les régions de décision $\mathcal{R}_k^*$ et $\mathcal{R}_\ell^*$ sont séparées par une frontière d'équation :

$$h_k(\boldsymbol{x}) = h_\ell(\boldsymbol{x}) \Leftrightarrow (\boldsymbol{\mu}_k-\boldsymbol{\mu}_\ell)^T\left(\boldsymbol{x}-\frac{\boldsymbol{\mu}_k+\boldsymbol{\mu}_\ell}{2}\right)=0.$$

C'est l'hyperplan médiateur du segment d'extrémités $\boldsymbol{\mu}_k$ et $\boldsymbol{\mu}_\ell$.

Notons que cette règle ne dépend que des moyennes $\boldsymbol{\mu}_k$, qui peuvent être estimées par $\widehat{\boldsymbol{\mu}}_k$. Le classifieur correspondant est appelé *classifieur euclidien*.

4.3 Choix d'un modèle d'analyse discriminante

D'une manière générale, ces modèles — à l'exception du modèle général d'analyse discriminante quadratique — ont pour but de réduire le nombre de paramètres à estimer (on parle quelquefois de modèles parcimonieux) pour gagner en robustesse, au prix d'une diminution de la flexibilité du modèle, comme le montre le tableau 10.1.

A priori, il pourrait sembler souhaitable de faire le moins d'hypothèses possible et de se placer d'emblée dans le cas le plus général. Cependant, il s'avère que, lorsque le nombre de paramètres à estimer est trop important, les erreurs d'estimation compensent le gain de flexibité du modèle. Il faut donc, en pratique, réaliser un compromis et rechercher un modèle de complexité adaptée à la taille de l'ensemble d'apprentissage.

Pour ce faire, il est possible de mettre en œuvre des méthodes de *sélection de modèle* permettant de choisir la famille de classifieurs la plus adaptée dans un ensemble donné.

4.4 Analyse discriminante régularisée (ADR)

Cette méthode permet de définir une infinité de règles de décision intermédiaires entre l'ADQ et l'ADL. Posons

$$\widehat{\Sigma}_k(\lambda) = \frac{(1-\lambda)(n_k-1)V_k+\lambda(n-g)V}{(1-\lambda)(n_k-1)+\lambda(n-g)},$$

avec $\lambda \in [0,1]$. Si $\lambda = 1$, on a $\widehat{\Sigma}_k(\lambda) = V$ et on retrouve l'ADL. Si $\lambda = 0$, on a $\widehat{\Sigma}_k(\lambda) = V_k$ et on retrouve l'ADQ. Pour $0 < \lambda < 1$, on a une infinité de solutions intermédiaires.

Si n est inférieur ou comparable à p, l'approche précédente, restant intermédiaire entre l'ADL et l'ADQ en termes de complexité, peut donner de moins bons résultats qu'un modèle plus simple comme l'ADL avec hypothèse d'indépendance conditionnelle, ou le classifieur euclidien. Une variante consiste donc à poser

$$\widehat{\Sigma}_k(\lambda,\gamma) = (1-\gamma)\widehat{\Sigma}_k(\lambda) + \gamma c_k I_p,$$

avec

$$c_k = \frac{\mathrm{Tr}(\widehat{\Sigma}_k(\lambda))}{p}.$$

Pour $\gamma = 0$, $\widehat{\Sigma}_k(\lambda,\gamma)$ est identique à l'estimateur $\widehat{\Sigma}_k(\lambda)$ précédent. Pour $\gamma = 1$, $\widehat{\Sigma}_k(\lambda,\gamma)$ est une matrice scalaire, ce qui revient à supposer que les variables X_j sont indépendantes conditionnellement à Z, et qu'elles ont la même variance conditionnelle.

À chaque valeur donnée à γ et λ correspond un estimateur des matrices de variance conditionnelles, et donc un nouveau classifieur lorsqu'on injecte ces estimateurs dans l'expression des fonctions discriminantes (10.2). Se pose donc la question du choix de ces *hyperparamètres*. Ce problème renvoie à celui, plus général, de la sélection de modèles.

5 Probabilité d'erreur de Bayes

5.1 Expression exacte ($g = 2$, $\Sigma_k = \Sigma$)

Dans certains cas simples, il est possible de calculer exactement la probabilité d'erreur de Bayes. Dans ce paragraphe, nous nous placerons dans le cas de deux classes, avec hypothèse d'homoscédasticité.

Dans ce cas, la règle de Bayes avec coûts $\{0,1\}$ peut s'écrire

$$\delta^*(\boldsymbol{x}) = \begin{cases} a_1 & \text{si } h(\boldsymbol{x}) < \ln \frac{\pi_1}{\pi_2} \\ a_2 & \text{sinon,} \end{cases}$$

avec

$$h(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)^T \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

On montre que

$$h(\boldsymbol{X}) \underset{\omega_1}{\sim} \mathcal{N}\left(-\frac{\Delta^2}{2}, \Delta^2\right),$$

et

$$h(\boldsymbol{X}) \underset{\omega_2}{\sim} \mathcal{N}\left(\frac{\Delta^2}{2}, \Delta^2\right),$$

avec $\Delta^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. La quantité Δ^2 est appelée *carré de la distance de Mahalanobis* entre les deux classes.

On en déduit l'expression de la règle de Bayes :

$$\begin{aligned} \varepsilon^* &= \mathbb{P}\left(h(\boldsymbol{X}) < \ln \frac{\pi_1}{\pi_2} | \omega_2\right) \pi_2 + \mathbb{P}\left(h(\boldsymbol{X}) \geq \ln \frac{\pi_1}{\pi_2} | \omega_1\right) \pi_1 \qquad (10.5) \\ &= \phi\left(\frac{\ln(\pi_1/\pi_2) - \Delta^2/2}{\Delta}\right) \pi_2 + \left[1 - \phi\left(\frac{\ln(\pi_1/\pi_2) + \Delta^2/2}{\Delta}\right)\right] \pi_1, \qquad (10.6) \end{aligned}$$

où ϕ représente la fonction de répartition de la loi normale (univariée) centrée réduite. Dans le cas $\pi_1 = \pi_2$, on a donc

$$\varepsilon^* = \phi\left(-\frac{\Delta}{2}\right).$$

5.2 Borne de Bhattacharyya

Dans le cas général, même en se limitant à $g = 2$, il n'est pas possible d'exprimer analytiquement l'erreur de Bayes. Cependant, on peut en donner des approximations.

Dans le cas de deux classes, on a

$$\begin{aligned} \varepsilon^* &= \int_{\mathbb{R}^p} \min\left(\mathbb{P}(\omega_1|\boldsymbol{x}), \mathbb{P}(\omega_2|\boldsymbol{x})\right) f(\boldsymbol{x}) d\boldsymbol{x} \qquad (10.7) \\ &= \int_{\mathbb{R}^p} \min\left(f_1(\boldsymbol{x})\pi_1, f_2(\boldsymbol{x})\pi_2\right) d\boldsymbol{x}. \qquad (10.8) \end{aligned}$$

Or, $\min(a, b) \leq \sqrt{ab}$ pour tous réels positifs a et b. On en déduit une borne supérieure de l'erreur de Bayes :

$$\varepsilon^* \leq \sqrt{\pi_1\pi_2} \int_{\mathbb{R}^p} \sqrt{f_1(\boldsymbol{x}) f_2(\boldsymbol{x})} d\boldsymbol{x} = \sqrt{\pi_1\pi_2}\, e^{-\Delta_B^2}.$$

La quantité Δ_B^2 est appelée *carré de la distance de Bhattacharyya* entre les deux classes. Dans le cas gaussien, on montre qu'elle est égale à :

$$\Delta_B^2 = \frac{1}{8}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} \ln \frac{\det \frac{\Sigma_1 + \Sigma_2}{2}}{\sqrt{\det \Sigma_1 \det \Sigma_2}}.$$

Cette quantité est donc composée de deux termes, dont le premier dépend de la différence des moyennes, et le second de la différence des variances. La distance de Bhattacharyya est souvent utilisée comme mesure de distance entre deux classes, même en dehors de l'hypothèse gaussienne (mais son interprétation liée à une borne supérieure de l'erreur de Bayes n'est alors plus valide).

6 Calcul des estimateurs des paramètres du modèle

Notons $Y = (\boldsymbol{X}, Z)$ le vecteur aléatoire composé du vecteur forme $\boldsymbol{X} \in \mathbb{R}^p$ et du vecteur Z des variables indicatrices de classe. Soit $\Psi = \{(\pi_k, \mu_k, \Sigma_k)_{k=1,\ldots,g}\}$ le vecteur des paramètres du modèle à estimer. La densité jointe du vecteur aléatoire Y s'écrit :

$$\begin{aligned} f_Y(y;\theta) &= \prod_{k=1}^{g} \left(\pi_k f_k(\boldsymbol{x})\right)^{z_{ik}}, \\ &= \prod_{k=1}^{g} \left(\pi_k (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu_k)^T \Sigma_k^{-1} (\boldsymbol{x} - \mu_k)\right)\right); \end{aligned}$$

la log-vraisemblance $\ln L(\Psi; y_1, \ldots, y_n)$ du vecteur de paramètres Ψ, notée plus simplement $\ln L(\Psi)$, est donc :

$$\begin{aligned} \ln L(\Psi) &= \sum_{i=1}^{n} \sum_{k=1}^{g} \ln\left(\pi_k f_k(\boldsymbol{x}_i)\right)^{z_{ik}}, \\ &= \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \ln\left(\pi_k (2\pi)^{-p/2} |\Sigma_k|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \mu_k)^T \Sigma_k^{-1} (\boldsymbol{x}_i - \mu_k)\right)\right), \\ &= \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \left(\ln \pi_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\boldsymbol{x}_i - \mu_k)^T \Sigma_k^{-1} (\boldsymbol{x}_i - \mu_k)\right). \end{aligned}$$

Pour déterminer les EMV des paramètres π_k, μ_k et Σ_k, il convient de maximiser cette log-vraisemblance par rapport à chacun de ces paramètres : la proportion π_k (scalaire), l'espérance μ_k (vecteur $g \times 1$), et la matrice de covariance Σ_k (matrice $g \times g$).

Pour ce faire, on procédera au calcul des dérivées premières de $\ln L(\Psi)$ par rapport à chacun de ces paramètres, pour ensuite les annuler (condition nécessaire d'optimalité). On admettra que les paramètres obtenus correspondent bien à des maxima de $\ln L(\Psi)$ (en d'autres termes, on admettra que la matrice des dérivées secondes par rapport aux différents paramètres est bien définie négative).

6.1 Modèle général (matrices Σ_k pleines)

EMV de π_k

La difficulté est ici de maximiser $\ln L(\Psi)$ tout en prenant en compte la contrainte $\sum_{k=1}^{g} \pi_k = 1$. La dérivée partielle de $\ln L(\Psi)$ par rapport à π_k est :

$$\frac{\partial \ln L(\Psi)}{\partial \pi_k} = \frac{1}{\pi_k} \sum_{i=1}^{n} z_{ik}.$$

Pour prendre en compte la contrainte $\sum_{k=1}^{g} \pi_k = 1$, on utilisera la formulation lagrangienne de ce problème d'optimisation de $L(\Psi)$ sous contrainte ; en introduisant le multiplicateur de Lagrange λ associé à la contrainte d'égalité [1], le lagrangien à maximiser est

$$\mathcal{L}\left(L(\Psi), \lambda\right) = \ln L(\Psi) - \lambda(\sum_{k=1}^{g} \pi_k - 1).$$

L'application des conditions d'optimalité à ce Lagrangien donnent :

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(L(\Psi), \lambda\right)}{\partial \pi_k} &= 0 \quad \Leftrightarrow \quad \frac{1}{\pi_k} \sum_{i=1}^{n} z_{ik} = \lambda \quad \Leftrightarrow \quad \frac{1}{\lambda} \sum_{i=1}^{n} z_{ik} = \pi_k, \\
\frac{\partial \mathcal{L}\left(L(\Psi), \lambda\right)}{\partial \lambda} &= 0 \quad \Leftrightarrow \quad \sum_{k=1}^{g} \pi_k = 1;
\end{aligned}$$

on en déduit :

$$\sum_{k=1}^{g} \pi_k = 1 \quad \Leftrightarrow \quad \sum_{k=1}^{g} \frac{1}{\lambda} \sum_{i=1}^{n} z_{ik} = 1 \quad \Leftrightarrow \quad \lambda = \sum_{k=1}^{g} \sum_{i=1}^{n} z_{ik} \Leftrightarrow \ \lambda = n,$$

et donc

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} z_{ik}. \tag{10.9}$$

EMV de μ_k

La matrice Σ_k (et donc Σ_k^{-1}) étant symétrique, la dérivée partielle de $\ln L(\Psi)$ par rapport à μ_k est :

$$\frac{\partial \ln L(\Psi)}{\partial \mu_k} = -\frac{1}{2} \sum_{i=1}^{n} -2 z_{ik} \Sigma_k^{-1} \left(\boldsymbol{x}_i - \mu_k\right);$$

par conséquent, Σ_k (et donc Σ_k^{-1}) étant définie positive, les conditions d'optimalité donnent :

$$\frac{\partial \ln L(\Psi)}{\partial \mu_k} = 0 \quad \Rightarrow \quad \mu_k = \frac{\sum_{i=1}^{n} z_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{n} z_{ik}}. \tag{10.10}$$

1. S'agissant d'une contrainte d'égalité, le multiplicateur est non signé, c'est-à-dire qu'il n'est pas sujet lui-même à une contrainte de positivité ou négativité.

EMV de Σ_k

Pour faciliter le calcul de l'EMV de la matrice de covariance Σ_k, commençons par spécifier quelques éléments de dérivation de matrices. Soit A la matrice carrée d'ordre p de terme général a_{ij}, et soit $f(A)$ une fonction de A. Par souci de simplicité, nous définissons la dérivée de $f(A)$ par rapport à A comme la matrice dont les éléments sont les dérivées de $f(A)$ par rapport aux éléments de A :

$$\partial f(A)/\partial A = \partial f(A)/\partial a_{ij}.$$

Rappelons tout d'abord que $\boldsymbol{x}^T A\boldsymbol{x} = \mathrm{Tr}(AB)$, avec $B = \boldsymbol{x}\boldsymbol{x}^T$. Or

$$\frac{\partial \mathrm{Tr}(AB)}{\partial A} = B^T.$$

Par ailleurs, soit $\mathrm{Cof}A$ la matrice des cofacteurs associée à A : par définition,

$$\frac{\partial |A|}{\partial A} = \mathrm{Cof}A = |A|(A^{-1})^T,$$

et il vient, d'après la propriété de dérivation d'une fonction composée, que

$$\frac{\partial \ln |A|}{\partial A} = (A^{-1})^T.$$

Réécrivons tout d'abord la fonction de log-vraisemblance :

$$\ln L(\Psi) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}\left(\ln \pi_k - \frac{p}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma_k^{-1}| - \frac{1}{2}\mathrm{Tr}(\Sigma_k^{-1}B_{ik})\right),$$

où $B_{ik} = (\boldsymbol{x}_i - \mu_k)(\boldsymbol{x}_i - \mu_k)^T$ (on rappelle également que si A est inversible, $|A^{-1}| = |A|^{-1}$). La dérivation par rapport à Σ_k^{-1} donne donc :

$$\frac{\partial \ln L(\Psi)}{\partial \Sigma_k^{-1}} = \frac{1}{2}\sum_{i=1}^{n} z_{ik}(\Sigma_k - B_{ik}),$$

et les conditions d'optimalité permettent donc d'obtenir l'expression de l'EMV de la matrice de covariance Σ_k :

$$\frac{\partial \ln L(\Psi)}{\partial \Sigma_k^{-1}} = 0 \quad \Leftrightarrow \quad \Sigma_k = \frac{\sum_{i=1}^{n} z_{ik}\widehat{B}_{ik}}{\sum_{i=1}^{n} z_{ik}}, \tag{10.11}$$

où $\widehat{B}_{ik} = (\boldsymbol{x}_i - \widehat{\mu}_k)(\boldsymbol{x}_i - \widehat{\mu}_k)^T$, et où $\widehat{\mu}_k$ est l'EMV de μ_k.

6.2 Modèles parcimonieux (matrices Σ_k contraintes)

Lorsque l'on impose des contraintes supplémentaires sur les matrices de variance pour réduire le nombre de paramètres à estimer, l'expression de leurs EMV change. Les calculs se font exactement de la même manière que dans le cas général, une fois que la vraisemblance a été réécrite de manière à prendre en compte la forme particulière que l'on désire imposer aux matrices Σ_k.

Par exemple, sous l'hypothèse d'indépendance des variables conditionnellement aux classes, Σ_k est une matrice diagonale de terme général σ_{kj}^2 (avec $j = 1, \ldots, p$) ; dans ce cas,

$$|\Sigma_k| = \prod_{j=1}^{p}\sigma_{kj}^2 \quad \text{et} \quad (\boldsymbol{x} - \mu_k)^T\Sigma_k^{-1}(\boldsymbol{x} - \mu_k) = \sum_{j=1}^{p}\frac{1}{\sigma_{kj}^2}(x_j - \mu_{kj})^2.$$

Il est facile de montrer que l'EMV de σ_{kj}^2 est alors

$$\widehat{\sigma}_{kj}^2 = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}(x_{ij} - \mu_{kj})^2 = s_{kj}^2,$$

les termes non diagonaux de la matrice étant nuls par hypothèse. Ainsi, sous l'hypothèse d'indépendance conditionnelle, l'EMV de Σ_k est donc $\widehat{\Sigma_k} = \mathrm{diag}(s_{k1}^2, \ldots, s_{kj}^2, \ldots, s_{kp}^2)$.

Chapitre 11

Régression logistique

1 Introduction

Dans le chapitre 10, nous avons vu que sous l'hypothèse que les données suivent dans chaque classe une loi normale, et lorsque l'on suppose les matrices de variance identiques, la règle de Bayes peut être exprimée à l'aide de fonctions discriminantes linéaires. Nous avons vu que cette méthode a l'avantage de fournir des estimations des probabilités a posteriori d'appartenance aux classes. Ces estimations sont d'autant plus précises que les hypothèses portant sur la distribution des données (normalité, forme ou égalité des matrices de covariance) sont vérifiées.

Plutôt que de faire des hypothèses sur les distributions conditionnelles f_k, une autre approche consiste à estimer *directement* les probabilités d'appartenance aux classes. C'est notamment le cas de la régression logistique, étudiée dans ce chapitre. Ce modèle s'exprime de manière très simple dans le cas de deux classes. Pour cette raison, il est très employé dans les applications biostatistiques où la prédiction d'une *réponse binaire* (présence/absence d'une pathologie, etc) à partir de variables explicatives est une problématique très fréquente.

2 Régression logistique binaire

2.1 Modèle général

L'idée à la base de la régression logistique consiste à modéliser les probabilités a posteriori $\mathbb{P}(\omega_k|\boldsymbol{x})$ par des fonctions de $\boldsymbol{x}$, choisies de manière à satisfaire naturellement les contraintes $\sum_{k=1}^{g} \mathbb{P}(\omega_k|\boldsymbol{x}) = 1$ et $\mathbb{P}(\omega_k|\boldsymbol{x}) \in [0;1]$ pour tout $\boldsymbol{x}$.

Plusieurs fonctions peuvent être utilisées pour ce faire. Un choix très populaire, le *modèle logit*, consiste à exprimer, pour $k = 1, \ldots, g-1$, le logarithme du rapport (ou *log-ratio*) des probabilités a posteriori $\mathbb{P}(\omega_k|\boldsymbol{x})$ et $\mathbb{P}(\omega_g|\boldsymbol{x})$ comme une fonction linéaire de $\boldsymbol{x}$:

$$\ln \frac{\mathbb{P}(\omega_k|\boldsymbol{x})}{\mathbb{P}(\omega_g|\boldsymbol{x})} = \boldsymbol{\beta}_k^T \boldsymbol{x}, \text{ pour } k = 1, \ldots, g-1.$$

(On supposera, dans tout ce chapitre, que le terme d'ordonnée à l'origine du modèle de régression est inclus aux vecteurs de paramètres, c'est-à-dire que $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})^T$ et $\boldsymbol{x} = (1, x_1, \ldots, x_p)^T$.)

Déterminons les expressions des probabilités a posteriori des classes. De la relation précédente, on obtient

$$\frac{\mathbb{P}(\omega_k|\boldsymbol{x})}{\mathbb{P}(\omega_g|\boldsymbol{x})} = \exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right), \text{ pour } k = 1, \ldots, g-1,$$

dont on peut déduire que

$$\sum_{k=1}^{g-1} \exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right) = \frac{1}{\mathbb{P}(\omega_g|\boldsymbol{x})} - 1.$$

Finalement,

$$\mathbb{P}(\omega_k|\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right)}{1 + \sum_{\ell=1}^{g-1} \exp\left(\boldsymbol{\beta}_\ell^T \boldsymbol{x}\right)}, \text{ pour } k = 1, \ldots, g-1,$$

et

$$\mathbb{P}(\omega_g|\boldsymbol{x}) = \frac{1}{1 + \sum_{k=1}^{g-1} \exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right)}.$$

Notons que l'on a choisi ici $\mathbb{P}(\omega_g|\boldsymbol{x})$ comme dénominateur des rapports de probabilités a posteriori ; ce choix est arbitraire.

2.2 Apprentissage (cas $g = 2$)

Pour l'apprentissage des paramètres on utilise la méthode du maximum de vraisemblance. Nous ne détaillerons ici que le cas le plus simple correspondant à $g = 2$. Posons

$$\begin{aligned} p(\boldsymbol{x};\boldsymbol{\beta}) &= \mathbb{P}(\omega_1|\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right)}{1 + \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right)}, \\ 1 - p(\boldsymbol{x};\boldsymbol{\beta}) &= \mathbb{P}(\omega_2|\boldsymbol{x}) = \frac{1}{1 + \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right)}. \end{aligned}$$

Dans la suite, on adoptera pour plus de simplicité la notation $p_i = p(\boldsymbol{x}_i;\boldsymbol{\beta})$.

On dispose d'un ensemble d'apprentissage $\{(\boldsymbol{x}_i, z_i), i = 1, \ldots, n\}$. On peut coder l'information de classe z_i par un indicateur binaire

$$t_i = \begin{cases} 1 & \text{si } Z_i = \omega_1, \\ 0 & \text{si } Z_i = \omega_2. \end{cases}$$

On peut voir t_i comme la réalisation d'une variable $T_i \sim \mathcal{B}(p_i)$. La fonction de vraisemblance conditionnelle [1] associée à l'échantillon $T_1, \ldots, T_n$ est donc

$$L(\boldsymbol{\beta}; t_1, \ldots, t_n) = \prod_{i=1}^{n} \mathbb{P}(T_i = t_i) = \prod_{i=1}^{n} p_i^{t_i} \left(1 - p_i\right)^{1-t_i},$$

d'où la fonction de log-vraisemblance

$$\ln L(\boldsymbol{\beta}; t_1, \ldots, t_n) = \sum_{i=1}^{n} \left(t_i \ln p_i + (1 - t_i) \ln(1 - p_i)\right).$$

Remarquons tout d'abord que

$$\begin{aligned} \frac{\partial p_i}{\partial \boldsymbol{\beta}} &= \frac{\boldsymbol{x}_i \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)(1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)) - \boldsymbol{x}_i \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i) \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\left(1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right)^2} = \boldsymbol{x}_i p_i (1 - p_i); \\ \frac{\partial \ln p_i}{\partial \boldsymbol{\beta}} &= \boldsymbol{x}_i (1 - p_i), \qquad \frac{\partial \ln(1 - p_i)}{\partial \boldsymbol{\beta}} = -\boldsymbol{x}_i p_i. \end{aligned}$$

1. puisqu'on modélise la probabilité a posteriori $\mathbb{P}(Y_i|\boldsymbol{x}_i)$ et non la probabilité jointe $\mathbb{P}(\boldsymbol{X}_i, Y_i)$

Le gradient de la log-vraisemblance s'écrit donc :

$$\begin{aligned}\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left(t_i \boldsymbol{x}_i (1-p_i) - (1-t_i)\boldsymbol{x}_i p_i\right), \\ &= \sum_{i=1}^{n} \boldsymbol{x}_i \left(t_i - p_i\right) = X^T(\boldsymbol{t}-\boldsymbol{p}),\end{aligned}$$

où $\boldsymbol{p} = (p_1, \ldots, p_n)^T$ et $\boldsymbol{t} = (t_1, \ldots, t_n)^T$. L'équation de vraisemblance

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

est un système de $p+1$ équations *non linéaires* par rapport à $\boldsymbol{\beta}$. On ne peut résoudre ce système directement : il faut donc rechercher le vecteur $\boldsymbol{\beta}$ qui maximise $\ln L$ en utilisant un algorithme d'optimisation itératif tel que *l'algorithme de Newton-Raphson*.

Cet algorithme consiste à faire, à la q^{e} itération, un développement limité de la fonction à maximiser (soit ici $\ln L(\boldsymbol{\beta})$) au voisinage de l'estimation courante $\boldsymbol{\beta}^{(q)}$ de la solution :

$$\ln L(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}^{(q)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)})^T \frac{\partial \ln L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{(q)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)})^T \underbrace{\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}(\boldsymbol{\beta}^{(q)})}_{H_{(q)}} (\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)}) + \epsilon; \quad (11.1)$$

dans cette expression, la matrice désignée par $H_{(q)}$ est la matrice hessienne (matrice des dérivées secondes) de la log-vraisemblance $\ln L$, calculée en $\boldsymbol{\beta}^{(q)}$. En dérivant par rapport à $\boldsymbol{\beta}$, on obtient :

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \approx \frac{\partial \ln L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{(q)}) + H_{(q)}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)}).$$

On a donc, en négligeant l'approximation :

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) = 0 \Leftrightarrow \boldsymbol{\beta} = \boldsymbol{\beta}^{(q)} - H_{(q)}^{-1} \frac{\partial \ln L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{(q)}).$$

La méthode de Newton-Raphson consiste à sélectionner un vecteur de poids initial $\boldsymbol{\beta}^{(0)}$, puis à calculer une séquence de vecteurs $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$ en appliquant itérativement cette formule. Chaque nouvelle estimation $\boldsymbol{\beta}^{(q+1)}$ est ainsi obtenue à partir de l'estimation précédente par

$$\boldsymbol{\beta}^{(q+1)} = \boldsymbol{\beta}^{(q)} - H_{(q)}^{-1} \frac{\partial \ln L}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^{(q)}).$$

La suite de vecteurs $\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$ converge vers un maximum local de la log-vraisemblance. En pratique, on arrêtera de calculer de nouvelles estimations $\boldsymbol{\beta}^{(q+1)}$ lorsqu'un critère d'arrêt sera vérifié (par exemple, la norme du gradient devient plus petite qu'un seuil fixé). Il est à noter qu'on utilise souvent le vecteur nul comme vecteur de poids initial $\boldsymbol{\beta}^{(0)}$.

La mise en application de cette méthode nécessite le calcul des coefficients de la matrice hessienne. Rappelons l'expression du vecteur gradient :

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \boldsymbol{x}_i \left(t_i - p_i\right);$$

on a donc

$$\begin{aligned}\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= -\sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial p_i}{\partial \boldsymbol{\beta}^T}, \\ &= -\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T p_i(1-p_i) = -X^T W X,\end{aligned}$$

où W est la matrice diagonale de terme général $W_{ii} = p_i\,(1 - p_i)$.

Notons $W_{(q)}$ l'estimation de la matrice W calculée avec le vecteur de poids $\boldsymbol{\beta}^{(q)}$ estimé à la q^{e} itération — on peut donc écrire $H_{(q)} = -X^T W_{(q)} X$. Notons de même l'estimation du vecteur des probabilités à la q^{e} itération $\boldsymbol{p}^{(q)} = (p(\boldsymbol{x}_1; \boldsymbol{\beta}^{(q)}), \ldots, p(\boldsymbol{x}_n; \boldsymbol{\beta}^{(q)}))^T$. On en déduit donc la règle de mise à jour des poids :

$$\boldsymbol{\beta}^{(q+1)} = \boldsymbol{\beta}^{(q)} + (X^T W_{(q)} X)^{-1} X^T (\boldsymbol{t} - \boldsymbol{p}^{(q)}).$$

2.3 Interprétation des coefficients

Cotes et rapports de cotes. Supposons que l'on connaisse les fréquences d'occurrence de l'événement $Y|\boldsymbol{X} = \boldsymbol{x}$: lorsque l'événement $Y = 1|\boldsymbol{X} = \boldsymbol{x}$ se produit $\text{pos}(\boldsymbol{x})$ fois, l'événement complémentaire $Y = 0|\boldsymbol{X} = \boldsymbol{x}$ se produise $\text{neg}(\boldsymbol{x})$ fois). On peut donc écrire

$$p(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\text{pos}(\boldsymbol{x})}{\text{pos}(\boldsymbol{x}) + \text{neg}(\boldsymbol{x})}, \quad 1 - p(\boldsymbol{x}; \boldsymbol{\beta}) = \frac{\text{neg}(\boldsymbol{x})}{\text{pos}(\boldsymbol{x}) + \text{neg}(\boldsymbol{x})};$$

on en déduit

$$c(\boldsymbol{x}) = \frac{p(\boldsymbol{x}; \boldsymbol{\beta})}{1 - p(\boldsymbol{x}; \boldsymbol{\beta})} = \frac{\text{pos}(\boldsymbol{x})}{\text{neg}(\boldsymbol{x})}.$$

La quantité $c(\boldsymbol{x})$ est appelée *cote*[2] de l'événement $Y = 1|\boldsymbol{X} = \boldsymbol{x}$. Il est usuel d'exprimer, dans le langage courant, des « cotes contre » (c'est-à-dire en défaveur de l'événement considéré) : ainsi, une cote de « dix contre un » signifie que $p(\boldsymbol{x}; \boldsymbol{\beta})/(1 - p(\boldsymbol{x}; \boldsymbol{\beta})) = 1/10$. Dans notre cas, la cote est une « cote pour » : plus $c(\boldsymbol{x})$ est élevée, plus l'événement $Y = 1|\boldsymbol{X} = \boldsymbol{x}$ a de chances d'arriver. Dans le modèle logistique binaire considéré, on voit très facilement que

$$c(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x}).$$

Supposons à présent que l'on fasse varier la j^{e} coordonnée du vecteur forme $\boldsymbol{x}$ observé. Notons $\boldsymbol{x}^{j+}$ le vecteur correspondant à $\boldsymbol{x}$ dont la j$^{\text{e}}$ coordonnée a été augmentée d'une unité :

$$\boldsymbol{x}^{j+} = \boldsymbol{x} + \mathbf{e}_j,$$

où $\mathbf{e}_j$ est le vecteur dont tous les éléments sont nuls à l'exception du j$^{\text{e}}$ qui est égal à 1. Le rapport de cotes pour la j$^{\text{e}}$ variable est alors

$$\frac{c(\boldsymbol{x}^{j+})}{c(\boldsymbol{x})} = \exp(\beta_j).$$

Lorsque x^j augmente d'une unité, la cote de l'événement $Y = 1|\boldsymbol{X} = \boldsymbol{x}$ est donc multipliée par $\exp(\beta_j)$: il s'agit d'une augmentation si $\beta_j > 0$ et d'une diminution si $\beta_j < 0$, la variation de x_j n'ayant aucune incidence si $\beta_j = 0$. Il est donc tentant d'interpréter le coefficient β_j comme un indicateur de la pertinence de la variable explicative correspondante dans le processus de prédiction — nous verrons plus loin quelles sont les limites de cette interprétation.

Significativité des coefficients : test de Wald. Remarquons tout d'abord que le vecteur de coefficients $\widehat{\boldsymbol{\beta}}$ estimé par la procédure décrite ci-dessus étant obtenu par maximum de vraisemblance, il est convergent, asymptotiquement sans biais, et asymptotiquement gaussien (si le modèle est correctement spécifié). En outre, la matrice hessienne étant indépendante de la variable aléatoire T, son opposée donne la matrice d'information de Fisher apportée par l'échantillon relativement au paramètre $\boldsymbol{\beta}$:

$$I_n(\boldsymbol{\beta}) = -\mathbb{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right] \simeq X^T \widehat{W} X,$$

2. Il s'agit en fait de la cote anglaise, la cote française étant $c(\boldsymbol{x}) + 1$.

où $\widehat{W}$ est la matrice W obtenue en remplaçant $\boldsymbol{\beta}$ par $\widehat{\boldsymbol{\beta}}$. Le test de Wald, qui est une stratégie courante pour tester la significativité d'un coefficient β_j, repose sur une approximation de la variance du coefficient estimé $\widehat{\beta}_j$; cela permet d'aboutir à la statistique de test suivante, appelée Z-score :

$$\mathcal{W}_j = \frac{\widehat{\beta}_j - \beta_{j0}}{\widehat{\sigma}_j} \underset{\text{app}}{\sim} \mathcal{N}(0,1).$$

Dans cette expression, l'écart-type $\widehat{\sigma}_j$ associé au coefficient β_j est la racine carrée du j$^{\text{e}}$ terme diagonal de l'inverse de la matrice d'information de Fisher. On teste la significativité de β_j en calculant $\mathcal{W}_j$ avec $\beta_{j0} = 0$, et en utilisant l'une des deux régions critiques suivantes, rigoureusement équivalentes :

$$RC = \left\{ |\mathcal{W}_j| > u_{1-\alpha^*/2} \right\} \Leftrightarrow RC = \left\{ \mathcal{W}_j^2 > \chi^2_{1;1-\alpha^*} \right\}.$$

Significativité des coefficients : test du rapport de vraisemblance. Le principe est de comparer la vraisemblance du modèle appris à partir de toutes les variables explicatives X^j à celle du modèle appris à partir d'un sous-ensemble de variables. L'hypothèse H_0 est que les variables omises n'ont aucun impact sur le modèle, et l'hypothèse alternative H_1 que l'une d'entre elles, au moins, a au contraire une influence significative :

$$\begin{cases} H_0 : \beta_j = 0, \text{ pour tout } j \in J \\ H_1 : \text{il existe } j \in J \text{ tel que } \beta_j \neq 0 \end{cases}$$

La suppression d'une ou plusieurs variables a généralement pour effet de faire décroître la vraisemblance du modèle ; il est toutefois nécessaire de déterminer si la différence de vraisemblance observée est significative ou non.

Soit $L(\boldsymbol{\beta})$ la valeur de vraisemblance du modèle incluant toutes les variables, et $L^{-J}(\boldsymbol{\beta})$ la vraisemblance du modèle omettant le sous-ensemble de variables $\{X^j, j \in J\}$, avec J l'ensemble des indices des variables écartées (généralement, on ne teste qu'une seule variable à la fois : $|J| = 1$). La statistique de test est :

$$-2\ln\Lambda = -2\ln\left(\frac{L^{-J}(\boldsymbol{\beta})}{L(\boldsymbol{\beta})}\right) = 2\left(\ln L(\boldsymbol{\beta}) - \ln L^{-J}(\boldsymbol{\beta})\right),$$

et la région critique du test est

$$RC = \left\{ -2\ln\Lambda > \chi^2_{J;1-\alpha^*} \right\}.$$

Ce test est plus coûteux à mettre en place que le test de Wald : il nécessite d'apprendre autant de modèles que de sous-ensembles de variables dont l'on souhaite évaluer la significativité. Néanmoins, étant basé sur des hypothèses moins fortes concernant la forme de la fonction de vraisemblance, il donne en général de meilleurs résultats.

Sélection de variables. Remarquons que ces tests sont généralement utilisés à des fins de *sélection de variables*. L'objectif est d'identifier les variables jugées non pertinentes pour expliquer la variable de sortie Z, afin de les écarter du modèle qui sera finalement appris.

Il existe plusieurs stratégies ; une stratégie descendante consiste à partir du modèle complet (utilisant toutes les variables) puis à supprimer les variables les unes après les autres tant que la perte (en termes de vraisemblance) n'est pas jugée significative. Une stratégie ascendante, à l'inverse, part du modèle vide (appris avec seulement une ordonnée à l'origine) puis à ajouter les variables tant que le gain est significatif. Des stratégies plus complexes existent également, que nous ne détaillerons pas ici, et qui consistent à alterner des étapes d'ajout et de suppression de variables.

Remarquons que ces stratégies sont heuristiques. Lorsque l'on compare deux modèles emboîtés (c'est-à-dire que l'un des modèles est un cas particulier de l'autre, par exemple

en ce qu'il concerne un sous-ensemble des variables considérées dans le second), la vraisemblance du modèle simple est nécessairement inférieure à celle du modèle complexe. Dans le cas général (ex. comparaison de deux modèles appris à partir de sous-ensembles de variables disjoints), cette propriété n'est pas vérifiée, et il n'est de ce fait pas possible de déterminer un sous-ensemble de variables « optimal » — à moins de tester toutes les combinaisons possibles de variables, ce qui n'est pas raisonnable si le nombre de variables est élevé.

3 Régression logistique multinomiale ($g > 2$)

Nous ne traiterons pas en détail le cas où la variable Z a plus de deux modalités. Nous ne mentionnerons ici que quelques éléments concernant l'apprentissage du modèle et évoquerons ensuite les défis principaux que pose cet apprentissage.

On utilise, comme précédemment, la méthode du maximum de vraisemblance pour estimer les $g-1$ vecteurs de coefficients. On définit à présent, pour tout $k = 1, \ldots, g-1$,

$$\begin{aligned} p_k(\boldsymbol{x}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{g-1}) &= \mathbb{P}(\omega_k|\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right)}{1 + \sum_{k=1}^{g-1} \exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right)}, \\ p_g(\boldsymbol{x}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{g-1}) &= \mathbb{P}(\omega_g|\boldsymbol{x}) = \frac{1}{1 + \sum_{k=1}^{g-1} \exp\left(\boldsymbol{\beta}_k^T \boldsymbol{x}\right)}. \end{aligned}$$

Pour chaque individu d'apprentissage, on dispose à présent d'un ensemble d'indicateurs de classe

$$t_{ik} = \begin{cases} 1 & \text{si } Z_i = \omega_k, \\ 0 & \text{sinon.} \end{cases}$$

Le vecteur $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{ig})$ d'indicateurs peut être vu comme la réalisation d'un vecteur aléatoire $\boldsymbol{T} \sim \mathcal{M}(1; p_{i1}, \ldots, p_{ig})$.

La fonction de vraisemblance conditionnelle associée à l'échantillon $T_1, \ldots, T_n$ est donc

$$L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{g-1}; \boldsymbol{t}_1, \ldots, \boldsymbol{t}_n) = \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{T}_i = \boldsymbol{t}_i) = \prod_{i=1}^{n} \prod_{k=1}^{g} (p_{ik})^{t_{ik}},$$

avec $p_{ik} = p_k(\boldsymbol{x}; \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{g-1})$, pour $k = 1, \ldots, g$; d'où la fonction de log-vraisemblance

$$\ln L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{g-1}; \boldsymbol{t}_1, \ldots, \boldsymbol{t}_n) = \sum_{i=1}^{n} \sum_{k=1}^{g} t_{ik} \ln p_{ik}.$$

On détermine les paramètres qui maximisent cette vraisemblance de la même manière que précédemment. Le vecteur gradient est à présent de dimensions $(p+1)(g-1) \times 1$, et la matrice hessienne de dimensions $(p+1)(g-1) \times (p+1)(g-1)$. Cette matrice est diagonale par blocs (elle est composée de sous-matrices de tailles $(p+1) \times (p+1)$ qui sont diagonales).

Le nombre de paramètres à déterminer croît donc linéairement en fonction du nombre de classes. On notera qu'il n'est pas possible de déterminer les vecteurs de coefficients séparément les uns des autres, ces vecteurs étant liés de par la contrainte de sommation à un des probabilités a posteriori qu'ils paramètrent.

Chapitre 12

Arbres binaires

1 Introduction

Dans ce chapitre, nous nous intéressons à une méthode d'apprentissage générique connue sous le nom d'arbres binaires, permettant de résoudre aussi bien des problèmes de discrimination que de régression. Ces méthodes consistent à calculer de manière récursive une partition l'espace des caractéristiques en régions *homogènes*, c'est-à-dire correspondant à une valeur particulière de la variable à expliquer Z.

Nous étudierons les principes généraux de construction et les principales propriétés des arbres, en faisant plus particulièrement référence à l'algorithme CART Breiman et al. (1984). Nous aborderons également la question de la régularisation, visant à éviter le phénomène de sur-apprentissage. Nous conclurons en évoquant plusieurs stratégies avancées dans lesquelles ils peuvent être exploités.

2 Principe

Nous détaillerons d'abord comment un arbre fonctionne sur des exemples de test. Rappelons que l'on considère une population $\mathcal{P}$ d'individus décrits par p variables explicatives $X^1, \ldots, X^p$ et une variable Z à expliquer qualitative (discrimination) ou quantitative (régression). Par souci de simplicité, nous supposerons les variables réelles, chaque individu étant donc associé à un vecteur $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$; toutefois, divers algorithmes permettent de traiter des descripteurs qualitatifs, voire les deux types simultanément.

2.1 Structure

Nous supposons dans ce cours que le classifieur peut être formalisé par un arbre binaire, c'est-à-dire une séquence de nœuds ayant chacun 0 ou 2 fils (certains algorithmes, comme C4.5 considèrent un nombre quelconque de fils). Chaque nœud $\mathcal{C}_\ell$ est associé à une région $\mathcal{R}_\ell$ de l'espace d'entrée. Chaque nœud interne (avec deux fils) est associé à une variable explicative X^{j_ℓ} (où j_ℓ désigne l'indice de la variable associée à $\mathcal{C}_\ell$) et un seuil s_ℓ, qui *divisent* $\mathcal{R}_\ell$ en deux sous-régions : on séparera l'espace en deux selon ou non que

$$X^{j_\ell} \leq s_\ell, \tag{12.1}$$

où s_ℓ est un seuil associé au nœud $\mathcal{C}_\ell$ déterminé par le processus d'apprentissage de l'arbre. Géométriquement, cette division correspond à une partition de la région $\mathcal{R}_\ell$ de l'espace des caractéristiques $\mathcal{X}$ associée à $\mathcal{C}_\ell$ par un hyperplan d'équation $X^{j_\ell} = s_\ell$.

Notons qu'un arbre définit une hiérarchie (voir chapitre 7) : on obtient une partition de l'espace $\mathcal{X}$ en considérant l'ensemble des régions $\mathcal{R}_\ell$ correspondant aux nœuds situés à un même niveau de profondeur de l'arbre.

2.2 Prédiction

Le processus de prédiction, pour un individu associé à un vecteur $\boldsymbol{x}$, s'effectue donc par une série de tests : à chaque nœud interne $\mathcal{C}_\ell$, si la valeur x_{j_ℓ} vérifie la condition (12.1), l'individu sera dirigé vers le fils gauche de $\mathcal{C}_\ell$, et vers son fils droit dans le cas contraire. Le processus se termine lorsqu'on atteint l'un des nœuds terminaux (ou feuille) de l'arbre.

La feuille atteinte correspond alors à une région de l'espace d'entrée, à laquelle est associée une valeur particulière de la variable Z à expliquer : on associera donc cette valeur à l'individu au terme du processus de prédiction.

L'algorithme 1 présente la procédure récursive de classement d'un individu par un arbre.

Algorithme 1 : Classement d'un exemple de test par un arbre

```
Fonction Classement(C, x_0)
    Entrées : C : arbre ou sous-arbre, x_0 : exemple de test
    Sorties : δ(x_0) : décision
    si le nœud C n'est pas terminal alors
        si x_{0,j_ℓ} ≤ s_ℓ alors
            retourner Classement(successeur de gauche de C, x_0)
        sinon
            retourner Classement(successeur de droite de C, x_0)
    sinon
        retourner décision associée à C
```

Exemple 12.1. *La figure 12.1 présente un ensemble de données dans un espace d'entrée $\mathcal{X} = \mathbb{R}^2$, comptant $n = 10$ exemples répartis en $g = 2$ classes ($n_1 = 5$ exemples dans la classe « triangle », et $n_2 = 5$ exemples dans la classe « carré »). Un arbre de classification simple de profondeur 2 permet de classer ces données. La variable à prédire est donc qualitative : $Z \in \Omega = \{$« triangle », « carré »$\}$.*

La racine $\mathcal{C}_0$ de l'arbre est associée à la région $\mathcal{R}_0 = \mathbb{R}^2$. La première division sépare donc $\mathbb{R}^2$ en deux, selon que $x_2 \leq 2.75$ ou que $x_2 > 2.75$: cette séparation est matérialisée par la droite horizontale d'équation $x_2 = 2.75$. Dans le demi-plan supérieur, la classe « triangle » est majoritaire ; dans le demi-plan inférieur, c'est la classe « carré ».

Le demi-plan inférieur est ensuite séparé en deux selon que $x_1 \leq 1.75$ ou que $x_1 > 1.75$ (séparation : demi-droite verticale d'équation $x_1 = 1.75$, pour tout $\boldsymbol{x}$ tel que $x_2 \leq 2.75$) ; le demi-plan supérieur, selon que $x_1 \leq 3.75$ ou que $x_1 > 3.75$ (séparation : demi-droite verticale d'équation $x_1 = 3.75$, pour tout $\boldsymbol{x}$ tel que $x_2 > 2.75$).

Ici, on associe les régions $\mathcal{R}_1$ et $\mathcal{R}_2$, correspondant respectivement aux feuilles $\mathcal{C}_2$ et $\mathcal{C}_5$, à la classe « triangle » (la seule représentée parmi les individus d'apprentissage présents dans ces régions), et les régions $\mathcal{R}_3$ et $\mathcal{R}_4$, correspondant aux feuilles $\mathcal{C}_3$ et $\mathcal{C}_6$, à la classe « carré ». Un individu de test $\boldsymbol{x} = (4, 4)^T$ évalué par l'arbre sera donc successivement présenté aux nœuds $\mathcal{C}_0$, $\mathcal{C}_4$, et $\mathcal{C}_6$, et sera finalement affecté à la classe « carré ».

3 Apprentissage

3.1 Objectif visé

Objectif On veut construire un arbre « performant » à partir d'un ensemble d'apprentissage $\mathcal{L} = \{(\boldsymbol{x}_1, z_1), \ldots, (\boldsymbol{x}_n, z_n)\}$. On rappelle que cette notion de performance traduit la capacité de l'arbre à prédire la variable Z pour de *nouveaux exemples*. Or il est clair qu'un arbre construit pour prédire parfaitement Z *pour les exemples d'apprentissage* (par exemple en divisant l'espace d'entrée en régions contenant chacune un unique exemple) ne généralisera pas nécessairement bien (d'autant plus si les classes sont mélangées).

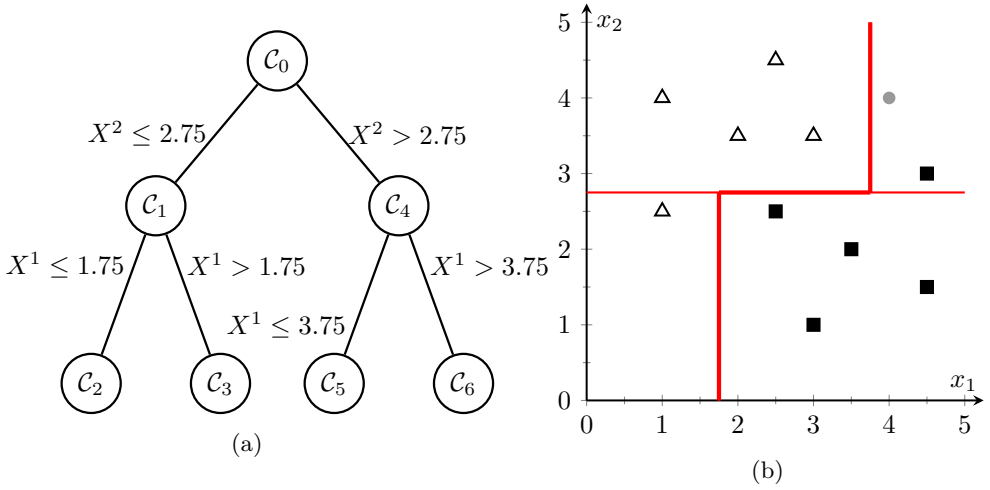


FIGURE 12.1 – Arbre de classification (12.1a) et partition correspondante de l'espace d'entrée et des données d'apprentissage (12.1b)

Compromis coût-complexité Intuitivement, la performance de l'arbre résulte d'un compromis entre sa capacité à expliquer l'ensemble d'apprentissage (qui résume la connaissance de la population $\mathcal{P}$ de référence) et sa simplicité, un arbre très complexe étant plus enclin à avoir sur-appris les données d'apprentissage.

Il existe évidemment un très grand nombre de partitions possibles de l'espace des caractéristiques en régions homogènes. La stratégie qui consisterait à les tester toutes, de manière à déterminer l'arbre donnant les meilleures performances, est beaucoup trop coûteuse calculatoirement. La *croissance* de l'arbre consiste donc à optimiser un critère d'homogénéité de manière gloutonne, en raffinant les régions de manière récursive.

Pour le *régulariser* (limiter sa complexité), plusieurs stratégies sont possibles. Une famille d'approches consiste à stopper la croissance de l'arbre lorsqu'il devient trop complexe, par exemple en arrêtant de diviser une région lorsqu'elle ne contient plus beaucoup d'exemples. Une autre consiste à développer l'arbre complètement, puis à le simplifier dans un second temps.

3.2 Croissance

Principe La procédure de croissance d'un arbre consiste à lui ajouter successivement des nœuds de manière à séparer l'espace des caractéristiques $\mathcal{X}$ en régions « homogènes », c'est-à-dire contenant des sous-ensembles d'apprentissage les plus purs possibles du point de vue de la variable à prédire Z, au sens d'un certain critère de pureté (voir ci-dessous). Ces ajouts se font de manière récursive, en raffinant progressivement les régions obtenues.

En discrimination, on cherchera des régions contenant des individus d'apprentissage de la même classe (dans l'exemple 12.1, $\mathcal{R}_1$ et $\mathcal{R}_2$ ne contiennent que des individus de la classe « triangle », et les régions $\mathcal{R}_3$ et $\mathcal{R}_4$ des individus de la classe « carré »). En régression, on cherchera des régions dans lesquelles les valeurs prises par Z sont proches, peu dispersées.

L'algorithme 2 détaille la procédure de division de l'espace $\mathcal{X}$ en sous-régions homogènes ; il fait apparaître clairement son caractère récursif. Une région $\mathcal{R}_\ell$ (non homogène), associée à un nœud interne $\mathcal{C}_\ell$ au sous-ensemble d'apprentissage $\mathcal{L}_\ell$, est divisée en $\mathcal{R}_\ell^-$ (associée à $\mathcal{C}_\ell^-$ et $\mathcal{L}_\ell^-$) et sa complémentaire $\mathcal{R}_\ell^+$ (associée à $\mathcal{C}_\ell^+$ et $\mathcal{L}_\ell^+$), de manière à maximiser le gain de pureté des sous-ensembles d'apprentissage $\mathcal{L}_\ell^-$ et $\mathcal{L}_\ell^+$.

Comme nous l'avons dit ci-dessus, ce processus de division de l'espace $\mathcal{X}$ s'arrête soit dès lors que l'arbre courant est jugé suffisamment complexe, soit quand toutes les feuilles sont associées à un sous-ensemble d'apprentissage pur au sens de Z.

Algorithme 2 : Croissance d'un arbre

Fonction `Croissance(`$\mathcal{L}$`)`
- **Entrées :** ensemble $\mathcal{L}_\ell$ d'individus étiquetés
- **Sorties :** nœud $\mathcal{C}_\ell$ (séparation (X^{j_ℓ}, s_ℓ), successeurs $\mathcal{C}_\ell^-$ et $\mathcal{C}_\ell^+$, décision δ_ℓ)
- **si** *les individus de* $\mathcal{L}_\ell$ *ne sont pas purs du point de vue de* Z **alors**
  - **pour** *chaque variable descriptive* X^j, $j = 1, 2, \ldots, p$ **faire**
    - Calculer le seuil optimal s_j pour la variable X^j, sur les individus de $\mathcal{L}_\ell$
  - En déduire la séparation optimale (X^{j_ℓ}, s_ℓ) pour le nœud $\mathcal{C}_\ell$
  - Diviser $\mathcal{L}_\ell$ en $\mathcal{L}_\ell^- \leftarrow \{(\boldsymbol{x}_i, z_i) \in \mathcal{L}_\ell : x_{ij_\ell} \leq s_\ell\}$ et $\mathcal{L}_\ell^+ \leftarrow \mathcal{L}_\ell \setminus \mathcal{L}_\ell^-$
  - Successeur $\mathcal{C}_\ell^- \leftarrow$ `Croissance(`$\mathcal{L}_\ell^-$`)`
  - Successeur $\mathcal{C}_\ell^+ \leftarrow$ `Croissance(`$\mathcal{L}_\ell^+$`)`
- **sinon**
  - Séparation optimale $(X^{j_\ell}, s_\ell) \leftarrow \varnothing$
  - Successeur $\mathcal{C}_\ell^- \leftarrow \varnothing$
  - Successeur $\mathcal{C}_\ell^+ \leftarrow \varnothing$
- Décision $\delta_\ell \leftarrow$ classe majoritaire dans $\mathcal{L}_\ell$
- **retourner** $\mathcal{C}_\ell$

Choix d'une division Pour un nœud $\mathcal{C}_\ell$, on identifie la division optimale de la région $\mathcal{R}_\ell$, c'est-à-dire la variable X^{j_ℓ} et le seuil s_ℓ donnant les sous-régions $\mathcal{R}_\ell^-$ et $\mathcal{R}_\ell^+$ les plus pures, en *testant toutes les combinaisons variable-seuil*. On se base pour cela sur les individus d'apprentissage $(\boldsymbol{x}_i, z_i) \in \mathcal{L}_\ell$, $i = 1, \ldots, n_\ell$ tels que $\boldsymbol{x}_i \in \mathcal{R}_\ell$.

Pour chaque variable X^j, les seuils possibles sont obtenus en ordonnant ses réalisations x_j sur les individus de $\mathcal{L}_\ell$. Soient $x_{(1)j}, x_{(2)j}, \ldots, x_{(n'_{\ell j})j}$ les $n'_{\ell j}$ valeurs *distinctes* ordonnées (par ordre croissant). Toute valeur s_j située entre deux valeurs successives $x_{(t)j}$ et $x_{(t+1)j}$ est une séparation admissible des individus de $\mathcal{L}_\ell$: il y en a donc au total $n'_{\ell j} - 1$ selon X^j. On pourra par exemple utiliser la moyenne des couples de valeurs successives, en considérant les seuils $s_{tj} = (x_{(t)j} + x_{(t+1)j})/2$, pour $t \in \{1, \ldots, n'_{\ell j} - 1\}$.

Pour chaque division de $\mathcal{L}_\ell$ selon X^j et s_{tj}, on peut calculer le gain associé en termes de pureté, étant donné un critère. Notons $\mathcal{L}_{\ell tj}^-$ et $\mathcal{L}_{\ell tj}^+$ les sous-ensembles obtenus en séparant $\mathcal{L}_\ell$ selon s_{tj} :

$$\mathcal{L}_{\ell tj}^- = \{(\boldsymbol{x}_i, z_i) \in \mathcal{L}_\ell : x_{ij} \leq s_{tj}\}, \quad \mathcal{L}_{\ell tj}^+ = \mathcal{L}_\ell \setminus \mathcal{L}_{\ell tj}^-;$$

notons de même $n_{\ell tj}^-$ et $n_{\ell tj}^+$ leurs tailles ($n_{\ell tj}^- + n_{\ell tj}^+ = n_\ell$), et $\boldsymbol{p}_{\ell tj}^-$ et $\boldsymbol{p}_{\ell tj}^+$ les vecteurs de proportions des classes dans ces ensembles. Pour une même variable X^j, on peut donc calculer le gain de pureté pour chacun des seuils s_{tj} candidats :

$$\text{gain}(s_{tj}) = n_\ell \, G(\boldsymbol{p}_\ell) - \left(n_{\ell tj}^- \, G(\boldsymbol{p}_{\ell tj}^-) + n_{\ell tj}^+ \, G(\boldsymbol{p}_{\ell tj}^+) \right). \tag{12.2}$$

Finalement, pour un nœud $\mathcal{C}_\ell$ de l'arbre, on identifiera donc la variable et le seuil optimaux comme étant ceux qui maximisent le gain, c'est-à-dire le couple $(X^{j_\ell}, s_\ell) = (X^{j^\star}, s_{t^\star j^\star})$ tel que

$$(j^\star, t^\star) = \underset{\substack{j=1,\ldots,p \\ t=1,\ldots,n'_\ell - 1}}{\arg\max} \ \text{gain}(s_{tj}).$$

Critère d'impureté Il existe plusieurs mesures d'*impureté* permettant de caractériser l'homogénéité des sous-ensembles résultant d'une séparation, ayant chacune des propriétés spécifiques.

Considérons tout d'abord un problème de discrimination. On requiert d'une mesure d'impureté calculée sur un ensemble d'individus qu'elle soit minimale si tous les éléments de

l'ensemble correspondent à une même modalité de Z, et maximale si toutes les modalités de Z sont représentées en quantités (ou proportions) égales. Notons $\boldsymbol{p} = (p_1, \ldots, p_g)^T$ le vecteur contenant les proportions de chacune des g modalités de Z dans l'ensemble considéré. L'*indice de Gini* G est utilisé par exemple dans l'algorithme CART :

$$G(\boldsymbol{p}) = \sum_{k=1}^{g} p_k(1 - p_k). \tag{12.3}$$

D'autres stratégies, comme l'algorithme C4.5, reposent sur l'*entropie* E :

$$E(\boldsymbol{p}) = -\sum_{k=1}^{g} p_k \ln(p_k). \tag{12.4}$$

On vérifie aisément que ces deux mesures sont minimales (nulles) lorsqu'une seule modalité de Z est présente dans l'ensemble considéré [1], et maximales lorsque toutes les modalités sont présentes en proportions égales. Notons que lorsqu'on crée une division au niveau de $\mathcal{C}_\ell$, le vecteur des proportions $\boldsymbol{p}_\ell$ est fixe : maximiser le gain (12.2) est alors équivalent à minimiser la somme pondérée des critères $n^-_{\ell tj}\, G(\boldsymbol{p}^-_{\ell tj}) + n^+_{\ell tj}\, G(\boldsymbol{p}^+_{\ell tj})$.

En régression, on pourra minimiser l'inertie intra-classe $\mathcal{I}_W$, ou de manière équivalente maximiser l'inertie inter-classe $\mathcal{I}_B$. Rappelons que lorsqu'on divise un ensemble de valeurs $z_1, \ldots, z_n$ en deux sous-ensembles,

$$\mathcal{I}_B = \frac{1}{n}\left(n_a(\overline{z}_a - \overline{z})^2 + n_b(\overline{z}_b - \overline{z})^2\right) = \frac{n_a n_b}{n^2}\left(\overline{z}_a - \overline{z}_b\right)^2, \tag{12.5}$$

où n_a et n_b représentent les effectifs des sous-ensembles, $\overline{z}_a$ et $\overline{z}_b$ leurs moyennes empiriques, et $\overline{z}$ la moyenne empirique de l'ensemble total.

Exemple 12.2. *Reprenons les données de l'exemple 12.1 pour appliquer la stratégie de développement de l'algorithme CART jusqu'à obtenir des régions pures. L'arbre est constitué initialement du nœud $\mathcal{C}_0$, associé à l'ensemble d'apprentissage $\mathcal{L}_0 = \mathcal{L}$ et à la région $\mathcal{R}_0 = \mathbb{R}^2$.*

Les gains en termes d'indice de Gini associés aux seuils possibles sont consignés dans la table 12.1. La séparation optimale associée au nœud racine $\mathcal{C}_0$ est donc obtenue en séparant l'espace selon la variable $X^{j_0} = X^2$ et le seuil $s_0 = s_{52} = 3.25$.

TABLE 12.1 – Seuils et indices de Gini correspondants pour le nœud $\mathcal{C}_0$

s_{t1}	1.50	2.25	2.75	3.25	4		
gain(s_{t1})	1.2500	2.1429	1.8000	2.1429	1.2500		
s_{t2}	1.25	1.75	2.25	2.75	3.25	3.75	4.25
gain(s_{t2})	0.5556	1.2500	2.1429	1.8000	3.3333	1.2500	0.5556

1. *On crée donc le successeur $\mathcal{C}_0^- = \mathcal{C}_1$ de $\mathcal{C}_0$, associé à la région $\mathcal{R}_1 = \{\boldsymbol{x} : x_2 \leq 3.25\}$ et au sous-ensemble d'apprentissage $\mathcal{L}_0^- = \mathcal{L}_1$; la procédure de croissance lui est appliquée de manière récursive.*

   Les seuils possibles et gains correspondants sont donnés dans la table 12.2. La séparation optimale est donc définie par la variable X^1 et le seuil $s_1 = 1.75$.

   (a) *Le successeur $\mathcal{C}_1^- = \mathcal{C}_2$ de $\mathcal{C}_1$, associé à $\mathcal{R}_2 = \{\boldsymbol{x} : x_2 \leq 3.25, x_1 \leq 1.75\}$, est associé à un sous-ensemble $\mathcal{L}_2$ pur. On arrête donc la croissance de l'arbre à son niveau.*

   (b) *De même, au nœud $\mathcal{C}_1^+ = \mathcal{C}_3$, associé à $\mathcal{R}_3 = \{\boldsymbol{x} : x_2 \leq 3.25, x_1 > 1.75\}$, on ne développe pas l'arbre, puisque $\mathcal{L}_3$ est également pur.*

1. On adoptera la convention $0 \ln(0) = 0$.

TABLE 12.2 – Seuils et indices de Gini correspondants pour le nœud $\mathcal{C}_1$

s_{t1}	1.75	2.75	3.25	4
gain(s_{t1})	1.6667	0.6667	0.3333	0.1667
s_{t2}	1.25	1.75	2.25	2.75
gain(s_{t2})	0.0667	0.1667	0.3333	0.0667

2. *On considère enfin le nœud $\mathcal{C}_0^+ = \mathcal{C}_4$ de $\mathcal{C}_0$, associé à la région $\mathcal{R}_4 = \{\boldsymbol{x} : x_2 > 3.25\}$. Le sous-ensemble $\mathcal{L}_4$ étant pur, on arrête le développement de cette branche, et donc la croissance de l'arbre.*

La figure 12.2 représente l'arbre ainsi obtenu et la partition de l'espace associée.

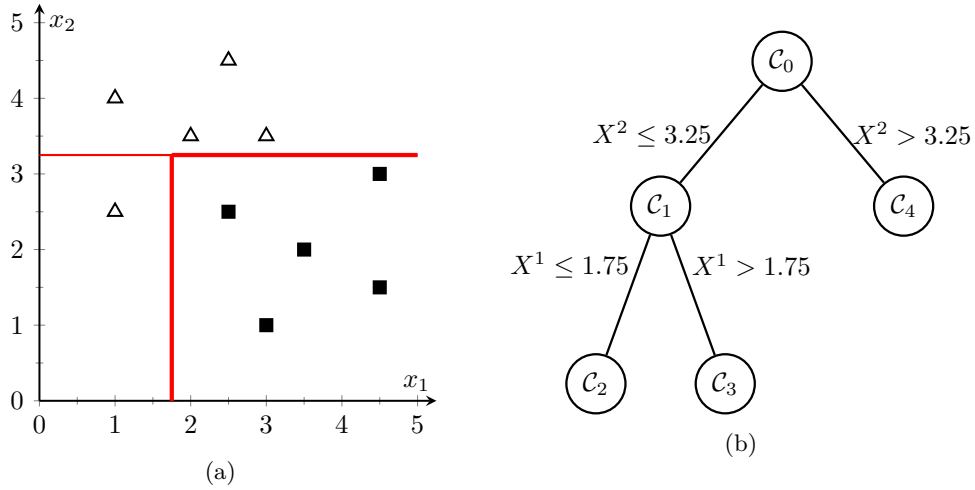


FIGURE 12.2 – Données d'apprentissage et partition (12.2a) correspondant à l'arbre de classification construit via l'algorithme CART (12.2b)

3.3 Contrôle de la complexité

Comme nous l'avons évoqué au paragraphe 3.1, un arbre séparant l'espace d'entrée en régions homogènes mais ne contenant qu'un petit nombre d'exemples d'apprentissage aura vraisemblablement une faible capacité de généralisation. Pour éviter le sur-apprentissage, on cherche un arbre ni trop simple ni trop complexe, capable de bien classer les données d'apprentissage tout en gardant une frontière de décision raisonnablement simple.

Arrêt prématuré Une première stratégie, dite de pré-élagage, ou d'arrêt prématuré (*early stopping*), consiste à arrêter la croissance de l'arbre dès lors que sa complexité est considérée suffisante, ou la nouvelle division peu intéressante.

Au niveau d'un nœud $\mathcal{C}_\ell$, on pourra arrêter de diviser $\mathcal{R}_\ell$ si l'effectif n_ℓ du sous-ensemble d'apprentissage $\mathcal{L}_\ell$ associé est jugé faible : pour cela, il est nécessaire de définir un seuil. On pourra alternativement utiliser un test statistique (comme le test du χ^2) pour déterminer si une séparation est informative en termes de prédiction, ou si au contraire la dépendance entre prédicteur X^{j_ℓ} et variable à expliquer Z n'est pas jugée significative.

Le problème est qu'une séparation au niveau d'un nœud $\mathcal{C}_\ell$ peut offrir un gain de pureté faible, mais néanmoins ouvrir la voie à de nouvelles séparations (associées à des nœuds descendants de $\mathcal{C}_\ell$) associées à des gains élevés.

Cette approche reste néanmoins très employée de par sa simplicité de mise en œuvre, ne requérant que le choix d'un seuil auquel comparer l'effectif n_ℓ d'un nœud à partir duquel on envisage le développement de l'arbre.

Post-élagage Une autre approche consiste à développer l'*arbre complet*, dont les feuilles sont associées à des sous-ensembles d'apprentissage parfaitement homogènes, puis à l'*élaguer*, c'est-à-dire limiter sa taille en coupant des branches. Cela permet de régulariser l'arbre ayant probablement sur-appris les données d'apprentissage. On ne présente ici que les principes du post-élagage, plus de détails étant donnés au paragraphe 5.

La stratégie d'élagage a posteriori est basée sur un critère de compromis entre le coût (erreur) de l'arbre $\mathcal{A}$ et sa complexité :

$$\widehat{\eta}_\lambda(\mathcal{A}) = \widehat{\varepsilon}(\mathcal{A}) + \lambda\,\widehat{\xi}(\mathcal{A}), \tag{12.6}$$

où $\widehat{\varepsilon}(\mathcal{A})$ est typiquement un estimateur de l'erreur de l'arbre (ou de son coût de classification) et $\widehat{\xi}(\mathcal{A})$ est mesure sa taille (c'est par exemple son nombre de feuilles). Le paramètre λ permet de régler le poids relatif de ces deux termes. L'objectif est de déterminer une valeur optimale de λ, la difficulté étant que $\widehat{\eta}_\lambda$ n'est pas une fonction continue de λ.

Supposons que l'on estime le taux d'erreur $\widehat{\varepsilon}$ par le taux d'erreur d'apprentissage $\widehat{\varepsilon}_a$. Les termes $\widehat{\varepsilon}(\mathcal{A})$ et $\widehat{\xi}(\mathcal{A})$ de l'équation (12.6) sont chacun fonction monotone de la taille de $\mathcal{A}$, et ont des effets antagonistes sur $\widehat{\eta}$: en supprimant un nœud, $\widehat{\varepsilon}_a$ augmente et $\widehat{\xi}$ décroît.

— Pour $\lambda = 0$, on a $\widehat{\eta}_0 = \widehat{\varepsilon}_a$, et l'arbre optimal est le plus petit sous-arbre de $\mathcal{A}$ qui minimise l'erreur d'apprentissage, noté $\mathcal{A}_0$ (défini comme étant l'arbre complet $\mathcal{A}$).

— Si λ augmente, on pénalise davantage la complexité de $\mathcal{A}_0$: à partir d'un certain seuil λ_1, l'une des branches de $\mathcal{A}_0$ devient plus coûteuse en termes de complexité $\widehat{\xi}$ que le gain d'erreur $\widehat{\varepsilon}$ qu'elle engendre. En l'élaguant, c'est-à-dire en ne conservant que le nœud racine de cette branche qui devient donc une feuille, on obtient un sous-arbre $\mathcal{A}_1$ de $\mathcal{A}_0$ (ce que l'on note $\mathcal{A}_1 \subset \mathcal{A}_0$), tel que pour $\lambda \geq \lambda_1$, $\widehat{\eta}_{\lambda_1} < \widehat{\eta}_{\lambda_0}$.

En répétant la procédure ci-dessus sur chaque arbre obtenu de la sorte par élagage (toujours en estimant l'erreur par $\widehat{\varepsilon}_a$), on crée une séquence de sous-arbres emboîtés :

$$\mathcal{C}_0 \subset \cdots \subset \mathcal{A}_k \subset \cdots \subset \mathcal{A}_1 \subset \mathcal{A}_0; \tag{12.7}$$

les éléments « terminaux » de cette séquence sont l'arbre complet $\mathcal{A}_0$ et l'arbre racine $\mathcal{C}_0$. Chaque élément de la séquence minimise $\widehat{\eta}_\lambda$ pour un intervalle de valeurs de λ : $\mathcal{A}_0$ minimise $\widehat{\eta}_\lambda$ pour tout $\lambda \in [\lambda_0 = 0; \lambda_1[$, $\mathcal{A}_1$ pour tout $\lambda \in [\lambda_1; \lambda_2[$, etc.

Il reste à identifier l'arbre « optimal » dans la séquence définie par (12.7). Pour cela, on détermine celui qui donne les meilleurs performances. L'erreur d'apprentissage $\widehat{\varepsilon}_a$ étant un estimateur biaisé (optimiste) du coût ε de l'arbre, l'utiliser pour trouver l'arbre optimal mènerait à conclure que ce dernier est $\mathcal{A}^\star = \mathcal{A}_K$, obtenu pour $\lambda_0 = 0$. On choisira donc plutôt l'arbre de la séquence qui minimise un autre estimateur du coût. On pourra utiliser l'erreur de *validation* si l'on dispose de suffisamment de données, ou l'erreur de *validation croisée* dans le cas contraire (voir le chapitre 13).

4 Méthodes ensemblistes

4.1 Propriétés des arbres

Avantages Les arbres binaires constituent une méthode populaire d'apprentissage supervisé, pour diverses raisons.

L'une des principales qualités de cette méthode est son coût calculatoire limité : les procédures d'apprentissage et de classement ne nécessitent que des opérations arithmétiques simples et des tests de variables. Les arbres peuvent ainsi être aisément déployés pour résoudre des problèmes d'apprentissage en grande dimension. En outre, le modèle est versatile : on peut facilement prendre en compte des variables quantitatives, qualitatives, voire d'un mélange des deux. Le choix du critère d'impureté n'a en pratique que peu d'influence sur les résultats obtenus (et la procédure d'induction étant générique, tout critère admissible peut être utilisé).

Enfin, les décisions du modèle sont interprétables. En effet, le processus de classement étant réalisé par une série de tests sur les variables naturelles, il est possible de le comprendre ou de l'interpréter (pour les modèles suffisamment simples). De plus, chaque nœud $\mathcal{C}_\ell$ peut être associé à des estimations des probabilités *a posteriori* des classes, spécifiques à la région de l'espace $\mathcal{R}_\ell$ correspondante. En pratique, la précision de ces estimations dépend de la taille du sous-ensemble d'apprentissage $\mathcal{L}_\ell$ associé.

Inconvénients La procédure d'apprentissage présentée est sous-optimale : en effet, l'optimisation du critère d'impureté n'est que locale (on optimise les séparations les unes après les autres, et non conjointement). Une stratégie globale qui consisterait à optimiser l'ensemble de l'arbre est délicate à mettre en œuvre pour des raisons combinatoires : c'est un problème d'optimisation discrète dans un espace pouvant être très grand — qui reste envisageable pour des arbres de taille raisonnable en utilisant des méthodes avancées de programmation mathématique.

Si la construction de l'arbre est peu sensible au choix du critère d'impureté, elle l'est en revanche à la procédure d'élagage. En outre, on constate que de faibles changements dans l'ensemble d'apprentissage peuvent avoir une grande influence sur la structure de l'arbre obtenu : en effet, la séparation associée à un nœud de l'arbre conditionne les séparations qui seront associées à ses successeurs. Un petit changement dans le tableau de données peut donc aboutir au choix d'une séparation optimale différente pour un nœud $\mathcal{C}_\ell$ donné, et par conséquent modifier toute la structure du sous-arbre dont la racine est $\mathcal{C}_\ell$.

4.2 Combinaison d'arbres

En pratique, on peut utiliser les arbres comme algorithme d'apprentissage dans des *méthodes ensemblistes*. Cela permet de pallier leurs défauts, voire d'en tirer parti.

Ces méthodes consistent à construire un ensemble de classifieurs, qui sont ensuite agrégés pour l'étape de prédiction. Elles tirent parti de la sensibilité des classifieurs utilisés, de leur capacité à sur-apprendre, et de leur faible coût calculatoire. On mentionnera trois méthodes : le bagging et sa variante des forêts aléatoires, et le boosting.

Bagging et forêts aléatoires Le *bagging* (*bootstrap aggregating*) consiste à générer un grand nombre de réplications $\mathcal{L}^{(1)}, \ldots, \mathcal{L}^{(B)}$ de l'ensemble d'apprentissage, par tirage aléatoire avec remise dans l'ensemble original $\mathcal{L}$. Cette procédure de *bootstrap* est présentée plus en détails au chapitre 13. Chacun de ces ensembles, $\mathcal{L}^{(b)}$, peut être utilisé pour apprendre un arbre $\mathcal{C}^{(b)}$.

Ces arbres ne sont pas élagués (tout au plus pourra-t-on limiter leur complexité par arrêt prématuré) : intuitivement, ils ont donc tendance à présenter un biais très faible, mais une variance très élevée. Pour limiter cette variance (et donc améliorer la capacité de généralisation), le processus de classement d'un point $\boldsymbol{x}$ consiste à agréger les sorties des arbres par moyennage, donnant ainsi une estimation moyenne des probabilités *a posteriori* ou des décisions données par les arbres (dans ce cas, il s'agit d'un vote majoritaire) [2].

La variante des *forêts aléatoires* consiste à utiliser la procédure de bagging décrite ci-dessus en modifiant l'apprentissage des arbres de manière à *augmenter la diversité* de l'ensemble d'arbres appris. À chaque ajout d'un nouveau nœud, on tire au hasard un sous-ensemble de variables (en général de taille $\sqrt{p}$), à partir desquelles on détermine la séparation optimale (X^{j_ℓ}, s_ℓ). Les arbres obtenus à partir des réplications de $\mathcal{L}$ ont donc tendance à être plus différents les uns des autres qu'avec un bagging simple. Le reste de la procédure est identique.

Notons que l'apprentissage des arbres et la procédure de classement d'un exemple par chacun d'eux peut être parallélisée, pour une efficacité calculatoire accrue.

2. En régression, on calculera de même la moyenne des sorties des arbres.

Boosting Tandis que le bagging consiste à combiner (par moyenne) un grand nombre de classifieurs complexes (typiquement des arbres non élagués), appris sur des réplications identiquement distribuées de l'ensemble $\mathcal{L}$, le *boosting* consiste à entraîner un grand nombre de classifieurs simples $\widehat{h}_t$, parfois dits *faibles* (*weak learners*), typiquement des arbres de profondeur 1 (*stumps*) ; ces arbres sont entraînés de manière séquentielle, afin que chacun se concentre sur les erreurs commises par les arbres précédents.

Pour cela, on maintient une distribution de poids associée aux exemples d'apprentissage : le poids w_i associé à l'exemple $\boldsymbol{x}_i$ (initialisé à $1/n$) diminue ou augmente selon que $\boldsymbol{x}_i$ est bien ou mal classé par un arbre nouvellement appris. Une fois la séquence déterminée, les arbres appris sont combinés par somme pondérée de leurs sorties, le poids d'un arbre dépendant de sa propension à bien classer les données. L'algorithme 3 résume la procédure pour un problème binaire ($g = 2$), la sortie de l'ensemble pour un exemple de test étant alors

$$\widehat{h}(\boldsymbol{x}) = \sum_{t=1}^{T} \alpha_t \widehat{h}_t(\boldsymbol{x}).$$

Algorithme 3 : Apprentissage par la procédure AdaBoost discrète (cas binaire)

Fonction `AdaBoost.appr(`$\mathcal{C}$, $\boldsymbol{x}_0$`)`
- **Entrées :** Ens. d'appr. $\mathcal{L}$, poids $\boldsymbol{w} = (w_1, \ldots, w_n)$, nb. itérations T
- **Sorties :** Séquence $\widehat{h}_1, \ldots, \widehat{h}_T$, et poids $\alpha_1, \ldots, \alpha_T$ associés
- **pour** $t = 1, \ldots, T$ **faire**
  - Apprendre un arbre $\widehat{h}_t : \mathcal{X} \to \{-1, 1\}$ avec les exemples pondérés $(\mathcal{L}, \boldsymbol{w})$
  - Estimer son erreur d'apprentissage : $\widehat{\varepsilon}_{at} = \sum_{i=1}^{n} w_i \mathbb{1}_{\widehat{h}_t(\boldsymbol{x}_i) \neq z_i}$
  - Calculer son poids : $\alpha_t = 1/2 \ln\{(1 - \widehat{\varepsilon}_{at})/\widehat{\varepsilon}_{at}\}$
  - Mettre à jour les poids des exemples : $w_i \leftarrow w_i \exp\{-z_i \alpha_t \widehat{h}_t(\boldsymbol{x}_i)\}$
  - Renormaliser les poids pour que $\sum_i w_i = 1$

Pour que la procédure de boosting soit performante (bonne capacité de généralisation), les conditions sont raisonnables : il s'agit essentiellement que les arbres combinés soient un peu plus performants qu'un classifieur aléatoire (décidant au hasard). Cependant, pour un problème binaire, on peut tirer parti d'arbres avec une très faible précision (plus de 50% de taux d'erreur) : un tel arbre est en effet associé à un poids α_t négatif, son vote en faveur d'une classe diminuant le score global de cette classe.

5 Détails sur la stratégie de post-élagage

Le post-élagage est la stratégie initialement proposée pour contrôler la complexité d'un arbre. En pratique, elle est aujourd'hui peu utilisée, les arbres étant souvent utilisés comme classifieur de base dans les stratégies de bagging (ou forêt aléatoire) et de boosting. On donne néanmoins ici quelques détails techniques supplémentaires sur cette procédure.

5.1 Calcul de la séquence de sous-arbres emboîtés

Rappelons que pour chaque valeur $\lambda \geq 0$, l'arbre $\mathcal{A}_\lambda$ de taille minimale qui minimise $\widehat{\eta}_\lambda$ est unique et appartient à la séquence (12.7). L'argument principal est que $\widehat{\varepsilon}$ et $\widehat{\xi}$ sont des fonctions monotones, respectivement croissante et décroissante, de k.

L'objectif est de déterminer la séquence (12.7) en calculant les intervalles de valeurs $[\lambda_t; \lambda_{t+1}[$ pour lesquels l'arbre optimal (minimisant (12.6) pour tout $\lambda \in [\lambda_k; \lambda_{k+1}[$) reste le même, c'est-à-dire qu'on veut trouver les valeurs « seuils » λ_{k+1} correspondant à un changement « minimal » de l'arbre $\mathcal{A}_k$.

Considérons un arbre $\mathcal{A}_k$ de la séquence. Notons $\mathcal{A}_k(\mathcal{C}_\ell)$ le sous-arbre de $\mathcal{A}_k$ dont la racine est le nœud $\mathcal{C}_\ell$ de $\mathcal{A}_k$ (on a donc $\mathcal{A}_k(\mathcal{C}_0) = \mathcal{A}_k$). Comparons $\mathcal{A}_k$ et le sous-arbre $\mathcal{A} \subset \mathcal{A}_k$ obtenu en élaguant $\mathcal{A}_k$ en $\mathcal{C}_\ell$, c'est-à-dire en supprimant le sous-arbre $\mathcal{A}_k(\mathcal{C}_\ell)$ rattaché à $\mathcal{A}_k$ en $\mathcal{C}_\ell$ (le nœud $\mathcal{C}_\ell$ devenant alors une feuille) : on pourra noter $\mathcal{A} = \mathcal{A}_k \setminus \mathcal{A}_k(\mathcal{C}_\ell)$. La valeur seuil pour laquelle $\widehat{\eta}_\lambda(\mathcal{A}_k) = \widehat{\eta}_\lambda(\mathcal{A})$, c'est-à-dire pour laquelle le gain de complexité de $\mathcal{A}$ compense exactement sa perte de précision, est :

$$\begin{aligned} \widehat{\eta}_\lambda(\mathcal{A}_k) = \widehat{\eta}_\lambda(\mathcal{A}) \quad &\Leftrightarrow \quad \widehat{\varepsilon}_a(\mathcal{A}_k) + \lambda\,\widehat{\xi}(\mathcal{A}_k) = \widehat{\varepsilon}_a(\mathcal{A}) + \lambda\,\widehat{\xi}(\mathcal{A}), \\ &\Leftrightarrow \quad \lambda = \frac{\widehat{\varepsilon}_a(\mathcal{A}) - \widehat{\varepsilon}_a(\mathcal{A}_k)}{\widehat{\xi}(\mathcal{A}_k(\mathcal{C}_\ell)) - 1}, \end{aligned} \tag{12.8}$$

puisqu'on obtient $\mathcal{A}$ à partir de $\mathcal{A}_k$ en supprimant $\widehat{\xi}(\mathcal{A}_k(\mathcal{C}_\ell)) - 1$ feuilles. Notons λ^* cette valeur seuil de λ qui vérifie (12.8) : pour $\lambda < \lambda_k^*$, $\mathcal{A}_k$ est plus intéressant que $\mathcal{A}$; pour $\lambda > \lambda_k^*$, $\mathcal{A}$ devient préférable à $\mathcal{A}_k$.

Rappelons que l'on cherche les valeurs seuils $\lambda_1^*, \lambda_2^*, \ldots$ qui correspondent à des modifications minimales de l'arbre optimal. Pour déterminer λ_1, on peut calculer les valeurs de λ satisfaisant l'équation (12.8) pour tous les sous-arbres de $\mathcal{A}_0$, et conserver la plus petite de ces valeurs (en cas d'*ex aequo*, on élaguera *toutes les branches* correspondant à cette plus petite valeur) ; puis on répétera cette procédure à partir de $\mathcal{A}_1$ pour obtenir $\mathcal{A}_2$; etc, jusqu'à obtenir l'arbre-racine $\mathcal{C}_0$.

Exemple 12.3 (Calcul d'une séquence d'arbres emboîtés)**.** *La figure 12.3 montre un ensemble d'apprentissage et un arbre de classification $\mathcal{A}_0$ complet appris via l'algorithme CART. On détaille ici le calcul de la séquence d'arbres obtenue à partir de $\mathcal{A}_0$.*

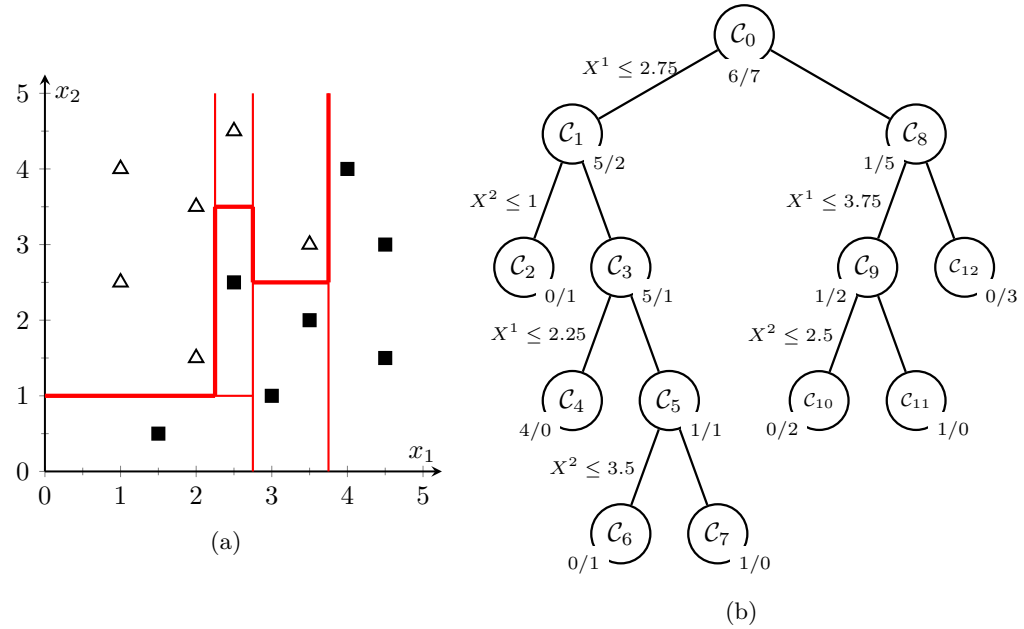


FIGURE 12.3 – Données d'apprentissage et partition (12.3a) correspondant à l'arbre construit via CART (12.3b) ; la composition du sous-ensemble d'apprentissage rattaché à chaque nœud est indiquée à côté, sous la forme [effectif ω_1]/[effectif ω_2]

On remarquera que le taux d'erreur d'apprentissage $\widehat{\varepsilon}_a(\mathcal{A})$ d'un arbre $\mathcal{A}$ est la proportion d'exemples d'apprentissage mal classés par ses feuilles : le nombre d'erreurs commises est donc l'effectif de la classe minoritaire. On obtient donc $\widehat{\varepsilon}_a(\mathcal{A})$ en divisant la somme des effectifs des classes minoritaires dans toutes les feuilles par le nombre total d'exemples d'apprentissage (évidemment, on aura $\widehat{\varepsilon}_a(\mathcal{A}_0) = 0$).

1. *En élaguant $\mathcal{A}_0$ en $\mathcal{C}_9$, on obtient l'arbre $\mathcal{A}_0 \setminus \mathcal{A}_0(\mathcal{C}_9)$ de coût $\widehat{\varepsilon}_a(\mathcal{A}_0 \setminus \mathcal{A}_0(\mathcal{C}_9)) = 1/13$*

et de complexité $\widehat{\xi}(\mathcal{A}_0 \setminus \mathcal{A}_0(\mathcal{C}_9)) = 6$. *On a*

$$\lambda = \frac{\widehat{\varepsilon}_a(\mathcal{A}_0 \setminus \mathcal{A}_0(\mathcal{C}_9)) - \widehat{\varepsilon}_a(\mathcal{A}_0)}{\widehat{\xi}(\mathcal{A}_0(\mathcal{C}_9)) - 1} = \frac{1/13 - 0}{2 - 1} = 1/13.$$

On peut de même calculer le rapport coût-complexité λ *pour un élagage en* $\mathcal{C}_8$, *en* $\mathcal{C}_5$, *en* $\mathcal{C}_3$, *en* $\mathcal{C}_1$, *et en* $\mathcal{C}_0$. *Les résultats sont présentés dans le tableau 12.3.*

TABLE 12.3 – Rapport coût-complexité pour un élagage de $\mathcal{A}_0$ en $\mathcal{C}_\ell$

$\mathcal{C}_\ell$	$\mathcal{C}_9$	$\mathcal{C}_8$	$\mathcal{C}_5$	$\mathcal{C}_3$	$\mathcal{C}_1$	$\mathcal{C}_0$
λ	1/13	1/26	1/13	1/26	2/39	1/13

La valeur de λ *minimale est donc* 1/26, *correspondant à un élagage de* $\mathcal{A}_0$ *en* $\mathcal{C}_8$ *ou en* $\mathcal{C}_3$. *Ces deux élagages sont équivalents en termes de rapport coût-complexité :* $\widehat{\xi}(\mathcal{A}_0 \setminus \mathcal{A}_0(\mathcal{C}_8)) = \widehat{\xi}(\mathcal{A}_0 \setminus \mathcal{A}_0(\mathcal{C}_3)) = 5$. *On élague alors* $\mathcal{A}_0$ en $\mathcal{C}_8$ et en $\mathcal{C}_3$, *ce qui correspond à la même valeur de* λ *(double gain de complexité et perte de précision), pour obtenir un nouvel arbre* $\mathcal{A}_1$ *correspondant à* $\lambda = 1/26$ *(voir figure 12.4a).*

2. *On peut de nouveau calculer les rapports coût-complexité pour les sous-arbres de* $\mathcal{A}_1$ *: on pourrait ainsi élaguer en* $\mathcal{C}_1$ *avec* $\lambda = (3/13 - 2/13)/(2 - 1) = 1/13$, *ou en* $\mathcal{C}_0$ *avec* $\lambda = 2/13$ *: on en déduit donc la valeur* $\lambda_2 = 1/13$ *et le sous-arbre* $\mathcal{A}_2 = \mathcal{A}_1 \setminus \mathcal{A}_1(\mathcal{C}_1)$ *correspondant (voir figure 12.4b).*
3. *Il ne reste qu'un élagage possible pour* $\mathcal{A}_2$ *donnant l'arbre-racine* $\mathcal{A}_3 = \mathcal{C}_0$ *(voir figure 12.4c), qui correspond à une valeur de* $\lambda_3 = 3/13$.

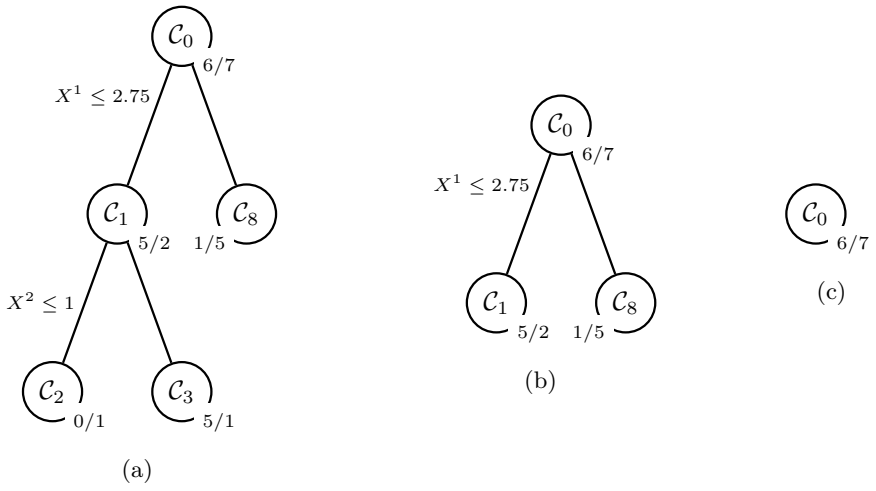


FIGURE 12.4 – Séquence de sous-arbres obtenue à partir de l'arbre complet $\mathcal{A}_0$: on obtient successivement $\mathcal{A}_1 = \mathcal{A}_0 \setminus \{\mathcal{A}_0(\mathcal{C}_3), \mathcal{A}_0(\mathcal{C}_8)\}$ correspondant à $\lambda_1 = 1/26$ (12.4a), puis $\mathcal{A}_2 = \mathcal{A}_1 \setminus (\mathcal{A}_1(\mathcal{C}_1))$ (12.4b), et enfin $\mathcal{A}_3 = \mathcal{A}_2 \setminus (\mathcal{A}_2(\mathcal{C}_0)) = \mathcal{C}_0$ (12.4c)

5.2 Calcul du sous-arbre optimal

Rappelons que le taux d'erreur d'apprentissage $\widehat{\varepsilon}_a$ utilisé dans l'équation (12.6) est un estimateur biaisé (optimiste) du coût ε de l'arbre : utiliser cet estimateur mènerait à conclure que l'arbre le plus précis est $\mathcal{A}^\star = \mathcal{A}_K$, obtenu pour $\lambda_0 = 0$. Une fois la séquence de sous-arbres emboîtés (12.7) calculée à partir de l'ensemble d'apprentissage $\mathcal{L}$ en utilisant $\widehat{\varepsilon}_a$, on sélectionnera un sous-arbre de la séquence, dit optimal, en remplaçant l'estimateur $\widehat{\varepsilon}_a$ par un autre estimateur du coût d'un arbre.

Erreur de validation On pourra utiliser l'erreur de *validation* si l'on dispose de suffisamment de données. On constituera un ensemble de validation $\mathcal{V}$ distinct de $\mathcal{L}$, et on testera tout simplement les performances de chacun des sous-arbres $\mathcal{A}_k$ de la séquence sur cet ensemble en calculant son taux d'erreur de validation $\widehat{\varepsilon}_V(\mathcal{A}_k)$. On choisira alors la valeur $\lambda^\star$ et le sous-arbre optimal $\mathcal{A}^\star$ associé qui minimisent $\widehat{\varepsilon}_V(\mathcal{A}_k)$:

$$(\lambda^\star, \mathcal{A}^\star) = \arg\min_k \widehat{\varepsilon}_V(\mathcal{A}_k).$$

Validation croisée Si l'on dispose de peu de données, on recourra généralement à une stratégie de *validation croisée* (voir le chapitre 13). Cette approche consiste à séparer l'ensemble d'apprentissage $\mathcal{L}$ en N_V sous-ensembles $\mathcal{L}^v$ (avec $v = 1, \ldots, N_V$) de tailles approximativement égales. On peut alors construire N_V arbres complets $\mathcal{C}_0^v$: chacun est obtenu en se servant de $\mathcal{L} \backslash \mathcal{L}_v$ comme ensemble d'apprentissage, et engendre une séquence d'arbres $\mathcal{A}_k^v$ emboîtés dans laquelle on peut sélectionner un arbre optimal $\mathcal{A}^{v\star}$, associé chacun à une valeur $\lambda_k^{v\star}$, en se servant de $\mathcal{L}_v$ comme ensemble de validation.

La difficulté est de choisir l'arbre optimal de la séquence (12.7) engendrée par $\mathcal{A}_0$ à partir de ces N_V arbres optimaux : pour une même valeur de k, les valeurs $\lambda_k^{v\star}$ ne sont généralement pas les mêmes — et les séquences obtenues ne sont pas toujours de même taille ! Notons que l'arbre $\mathcal{A}_k$ minimise $\widehat{\eta}_\lambda$ pour tout $\lambda \in [\lambda_k; \lambda_{k+1}[$: on associe donc à $\mathcal{A}_k$ la moyenne géométrique $\overline{\lambda}_k = \sqrt{\lambda_k \lambda_{k+1}}$, et on définit son erreur de validation croisée $\widehat{\varepsilon}_{CV}$ par

$$\widehat{\varepsilon}_{CV}(\mathcal{A}_k) = \frac{1}{N_V} \sum_{v=1}^{N_V} \widehat{\varepsilon}_v \left(\mathcal{A}^v(\overline{\lambda}_k) \right),$$

où $\mathcal{A}^v(\overline{\lambda}_k)$ est l'arbre optimal de la v^{e} séquence obtenu pour la valeur $\overline{\lambda}_k$ — et qui n'est pas nécessairement le k^{e} arbre optimal de cette séquence, dont on calcule l'erreur $\widehat{\varepsilon}_v$ avec l'ensemble de validation $\mathcal{L}_v$.

Enfin, le sous-arbre optimal $\mathcal{A}^\star$ de la séquence (12.7) engendrée par $\mathcal{A}_0$ et la valeur $\lambda^\star$ associée sont obtenus en minimisant $\widehat{\varepsilon}_{CV}$:

$$(\lambda^\star, \mathcal{A}^\star) = \arg\min_k \widehat{\varepsilon}_{CV}(\mathcal{A}_k).$$

Chapitre 13

Evaluation des performances et sélection de classifieur

1 Introduction

Comme nous l'avons vu dans le chapitre précédent, il est toujours possible de construire plusieurs classifieurs à partir d'un même ensemble de données, en se basant sur différents *modèles* ou *familles* (par exemple : modèles gaussiens avec ou sans hypothèse d'homoscédasticité, avec ou sans hypothèse d'indépendance conditionnelle, etc.). Se posent alors deux problèmes :

1. Quel modèle choisir ?
2. Une fois choisi un modèle, quelle est le niveau de performance du classifieur correspondant ?

Le niveau de performance d'un classifieur δ est défini par son risque

$$r(\delta) = \mathbb{E}_{X,Z}[c(\delta(\boldsymbol{X}), Z)].$$

Dans le cas de coûts $\{0, 1\}$, on rappelle que le risque est égal à la probabilité d'erreur.

Pour répondre aux deux questions précédentes, il faut tout d'abord disposer de méthodes permettant d'estimer le risque d'un classifieur donné, ou l'espérance du risque pour un classifieur construit à partir d'un échantillon aléatoire de taille donnée.

2 Estimation du risque

2.1 Méthode de resubstitution

La méthode de resubstitution consiste à estimer le risque $r(\delta)$ d'un classifieur δ par le coût moyen sur l'ensemble d'apprentissage :

$$\widehat{r}_R(\delta) = \frac{1}{n} \sum_{i=1}^{n} c\left(\delta(\boldsymbol{x}_i), z_i\right).$$

Dans le cas de coûts $\{0, 1\}$, $\widehat{r}_R(\delta)$ est le taux d'erreur d'apprentissage, c'est-à-dire la proportion d'individus mal classés dans l'ensemble d'apprentissage.

Cette méthode est déconseillée, car elle tend à fournir une estimation optimiste du risque (l'estimateur $\widehat{r}_R(\delta)$ est biaisé). Par exemple, dans le cas de la règle du plus proche voisin, le taux d'erreur d'apprentissage est toujours nul (puisque le plus proche voisin d'un vecteur de l'ensemble d'apprentissage est lui-même).

2.2 Méthode de l'ensemble de validation

Cette méthode consiste à partitionner aléatoirement les données initiales $\mathcal{D}$ en un ensemble d'apprentissage $\mathcal{L}$ et un *ensemble de validation* $\mathcal{V}$, avec $\mathcal{D} = \mathcal{L} \cup \mathcal{V}$ et $\mathcal{L} \cap \mathcal{V} = \varnothing$. On prend en général $\text{card}(\mathcal{L}) > \text{card}(\mathcal{V})$ car l'apprentissage de la règle de décision est un problème plus difficile que l'estimation du taux d'erreur. Typiquement, on réserve un tiers ou un quart des données initiales pour l'estimation du risque.

Soit $\mathcal{V} = \{(\boldsymbol{x}'_1, z'_1), \ldots, (\boldsymbol{x}'_{n'}, z_{n'})\}$. Ayant construit le classifieur δ à l'aide de l'ensemble $\mathcal{L}$, on pose :

$$\widehat{r}_v(\delta) = \frac{1}{n'} \sum_{i=1}^{n'} c\left(\delta(\boldsymbol{x}'_i), z'_i\right).$$

Cette méthode est rigoureuse car elle fournit un estimateur sans biais du risque. En revanche, elle nécessite un ensemble de données de taille importante, les données de validation ne pouvant être utilisées pour l'apprentissage.

2.3 Méthode de validation croisée

Dans cette méthode, on cherche à estimer l'espérance r_n du risque d'un classifieur basé sur un modèle donné, construit à l'aide d'un ensemble de données de taille n issu de la même distribution que $\mathcal{D}$:

$$r_n = \mathbb{E}_{\mathcal{D}_n}\left[r(\delta_{\mathcal{D}_n})\right],$$

$\mathcal{D}_n$ étant un échantillon aléatoire de taille n, et $\delta_{\mathcal{D}_n}$ un classifieur construit à partir de $\mathcal{D}_n$.

Pour estimer r_n, on construit successivement plusieurs classifieurs à partir de plusieurs ensembles d'apprentissage : on parle de *méthode de rééchantillonnage*. Les différents ensembles d'apprentissage sont construits de la manière suivante. On partitionne les données initiales $\mathcal{D}$ en N_V blocs $\mathcal{D}_{(1)}, \ldots, \mathcal{D}_{(N_V)}$ de tailles à peu près égales. Soit $n_{(k)} = \text{card}\,\mathcal{D}_{(k)}$. Soit δ^{-k} la règle de décision construite à partir de l'ensemble d'apprentissage $\mathcal{D} \setminus \mathcal{D}_{(k)}$, et $\widehat{r}_{(k)}$ son risque estimé sur $\mathcal{D}_{(k)}$:

$$\widehat{r}_{(k)} = \frac{1}{n_{(k)}} \sum_{(\boldsymbol{x}_i, z_i) \in \mathcal{D}_{(k)}} c\left(\delta^{-k}(\boldsymbol{x}_i), z_i\right).$$

L'estimateur de validation croisée de r_n est défini par

$$\widehat{r}_{CV} = \frac{1}{n} \sum_{k=1}^{N_V} n_{(k)} \widehat{r}_{(k)}.$$

On choisit typiquement N_V de l'ordre de 5 à 10. Dans le cas $N_V = n$, on parle de méthode *leave-one-out*.

Remarquons que $\widehat{r}_{CV}$ est un estimateur biaisé (pessimiste) qui tend à surestimer r_n, car il est obtenu en moyennant les risques estimés de N_V classifieurs construits à partir d'ensembles d'apprentissage de taille $n(1 - 1/N_V)$. Le biais est d'autant plus faible que N_V est grand. La méthode de validation croisée permet une utilisation « optimale » des données (puisque chaque exemple est utilisé tour à tour pour l'apprentissage et pour le test), au prix d'un temps de calcul plus important (il faut apprendre N_V classifieurs différents).

2.4 Méthode du bootstrap

Le bootstrap est une méthode générale d'estimation de la loi de probabilité d'une statistique $T(X_1, \ldots, X_n)$.

Principe général

La méthode consiste à constituer N_B échantillons de bootstrap à partir de $\mathcal{D}$. Un échantillon de bootstrap $\mathcal{D}_b^*$ a la même taille que l'échantillon initial. Il s'obtient en faisant n tirages *avec remise* dans $\mathcal{D}$. Il peut donc contenir plusieurs fois la même observation.

Soit T_b^* la statistique calculée à partir de l'échantillon de bootstrap $\mathcal{D}_b^*$. Pour estimer une caractéristique de la distribution de T, il suffit de calculer la caractéristique correspondante pour la distribution empirique $T_1^*, \ldots, T_b^*$. Par exemple, on estimera la variance σ^2 de T par

$$\widehat{\sigma}^2_{boot} = \frac{1}{N_B} \sum_{b=1}^{N_B} \left[T_b^* - \left(\frac{1}{N_B} \sum_{b=1}^{N_B} T_b^* \right) \right]^2 .$$

Application à l'estimation du risque

Comme dans la méthode de validation croisée, on cherche à estimer le risque moyen r_n d'un classifieur construit à partir d'un ensemble d'apprentissage aléatoire de taille n.

Soient N_B échantillons de bootstrap $\mathcal{D}_1^*, \ldots, \mathcal{D}_B^*$ (ce sont des ensembles d'apprentissage de même taille que $\mathcal{D}$, soit n). Soit δ_b^* le classifieur construit à partir de $\mathcal{D}_b^*$. Son risque estimé sur l'ensemble de données initial $\mathcal{D}$ est :

$$\widehat{r}_b^* = \frac{1}{n} \sum_{i=1}^{n} c(\delta_b^*(\boldsymbol{x}_i), z_i).$$

Une première application de la méthode du bootstrap consiste à estimer le risque par :

$$\widehat{r}_{boot} = \frac{1}{N_B} \sum_{b=1}^{N_B} \widehat{r}_b^* = \frac{1}{n\, N_B} \sum_{b=1}^{N_B} \sum_{i=1}^{n} c(\delta_b^*(\boldsymbol{x}_i), z_i).$$

Le problème est ici que l'estimateur $\widehat{r}_{boot}$ est optimiste, car $\mathcal{D}^b \cap \mathcal{D} \neq \varnothing$: certains exemples sont utilisés à la fois pour l'apprentissage et pour le test.

Une solution à ce problème consiste à évaluer chaque classifieur δ_b^* en utilisant uniquement les exemples $(\boldsymbol{x}_i, z_i)$ qui n'appartiennent pas à $\mathcal{D}_b^*$. Soit C^{-i} l'ensemble des indices b des échantillons de bootstrap qui ne contiennent pas l'exemple $\boldsymbol{x}_i$. On pose

$$\widehat{r}^{(1)}_{boot} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} c(\delta_b^*(\boldsymbol{x}_i), z_i).$$

Cette méthode est meilleure que la précédente, mais elle présente encore un inconvénient : dans chaque échantillon de bootstrap $\mathcal{D}_b^*$, certains $\boldsymbol{x}_i$ apparaissent plusieurs fois. Le nombre d'exemples *distincts* dans chaque échantillon de bootstrap est donc en général strictement inférieur à n : tout se passe comme si ces échantillons étaient en réalité plus petits que $\mathcal{D}$. La probabilité qu'un exemple $\boldsymbol{x}_i$ n'appartienne pas à $\mathcal{D}_b^*$ est

$$\mathbb{P}(\boldsymbol{x}_i \notin \mathcal{D}_b^*) = \left(1 - \frac{1}{n}\right)^n \approx 1/e.$$

La probabilité qu'un exemple $\boldsymbol{x}_i$ appartienne à $\mathcal{D}_b^*$ est donc :

$$\mathbb{P}(\boldsymbol{x}_i \in \mathcal{D}_b^*) \approx 1 - 1/e \approx 0.632.$$

On en déduit que le nombre moyen d'exemples distincts dans chaque échantillon de bootstrap est environ $0.632n$.

Ces considérations ont conduit à définir empiriquement l'estimateur 0.632 comme une moyenne pondérée de l'estimateur de resubstitution $\widehat{r}_R$ (optimiste) et de l'estimateur $\widehat{r}^{(1)}_{boot}$ (pessimiste) :

$$\widehat{r}^{(0.632)}_{boot} = 0.368\, \widehat{r}_R + 0.632\, \widehat{r}^{(1)}_{boot}.$$

3 Méthodologie générale de sélection de modèle

En pratique, il faut utiliser les données $\mathcal{D}$ disponibles à la fois pour sélectionner un modèle parmi un ensemble donné, puis, une fois choisi un modèle, pour construire le classifieur δ correspondant, et enfin estimer son risque.

Pour cela, une méthode générale consiste tout d'abord à mettre de côté une partie des données constituant un ensemble de test $\mathcal{T} = \{(\boldsymbol{x}''_1, z''_1), \ldots, (\boldsymbol{x}''_{n''}, z''_{n''})\}$. A l'aide des données restantes $\mathcal{D} \setminus \mathcal{T}$, on applique l'une des méthodes précédentes (le plus souvent : validation croisée ou bootstrap) pour sélectionner le modèle minimisant l'espérance estimée du risque r_n.

Ayant choisi le meilleur modèle, on refait un apprentissage sur l'ensemble $\mathcal{D} \setminus \mathcal{T}$ pour tirer partie de l'ensemble des données disponibles, ensemble de test exclu. Soit δ le classifieur obtenu. Enfin, on estime $r(\delta)$ par :

$$\widehat{r}_t(\delta) = \frac{1}{n''} \sum_{i=1}^{n''} c\left(\delta(\boldsymbol{x}''_i), z''_i\right).$$

Dans le cas de coûts $\{0, 1\}$, $\widehat{r}_t(\delta)$ est appelé *taux d'erreur de test*.

Cette méthode est utilisée, en particulier, lorsqu'on dispose d'une famille de modèles paramétrée par un ou plusieurs *hyperparamètres*, comme par exemple dans le cas de l'analyse discriminante régularisée. La validation croisée ou le bootstrap sont alors couramment utilisés pour déterminer automatiquement les valeurs des hyperparamètres.

Chapitre 14

La régression linéaire multiple

1 Introduction

La régression linéaire multiple a pour but l'étude de la relation entre une variable à expliquer quantitative Y et p variables explicatives $x_1, \ldots, x_p$. Les variables explicatives sont supposées connues sans erreur et de nature non aléatoire (on les note par des lettres minuscules). Dans certains cas, il peut s'agir de variables contrôlées par l'utilisateur (conditions expérimentales, paramètres d'un process, etc.). La variable Y est, elle, une variable aléatoire. Sa loi de probabilité est supposée dépendre des x_i selon le modèle suivant :

$$Y = b_0 + b_1 x_1 + \ldots + b_p x_p + \varepsilon$$

avec $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. L'espérance de Y est donc fonction linéaire des entrées, tandis que la variance de Y, égale à σ^2 ne dépend pas des entrées (hypothèse d'homoscédasticité). L'équation précédente peut s'écrire vectoriellement :

$$Y = \boldsymbol{x}^T \mathbf{b} + \varepsilon$$

avec $\boldsymbol{x} = (1, x_1, \ldots, x_p)^T \in \mathbb{R}^{p+1}$ et $\mathbf{b} = (b_0, b_1, \ldots, b_p)^T \in \mathbb{R}^{p+1}$.

Remarque 1. *La linéarité du modèle est essentiellement une linéarité* par rapport aux paramètres. *En effet, une relation linéaire par rapport aux paramètres mais non linéaire par rapport aux variables d'entrée peut toujours être linéarisée par changement de variable. Par exemple, si l'on a*

$$Y = b_0 + b_1 z^2 + b_2 \ln z + \varepsilon,$$

on peut toujours poser $x_1 = z^2$, $x_2 = \ln z$, *et retrouver le modèle linéaire précédent.*

Supposons que l'on ait observé les variables $x_1, \ldots, x_p, Y$ pour n individus (dans n situations différentes). Les données se présentent donc sous la forme suivante :

$$\begin{array}{cccc} x_{11} & \ldots & x_{1p} & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} & y_n \end{array}$$

On suppose que chaque valeur observée y_i sur un individu i est une réalisation d'une v.a.r. Y_i de la forme :

$$Y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + \varepsilon_i \quad i = 1, n$$

avec $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ et $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

Matriciellement, ces n équations s'écrivent

$$\boldsymbol{Y} = X\mathbf{b} + \varepsilon$$

avec

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix}$$

et

$$\mathbb{E}(\varepsilon) = 0 \quad \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Ce modèle étant posé, nous allons successivement aborder les problèmes suivants :

— estimation des paramètres $\mathbf{b}$ et σ^2 ;
— tests d'hypothèses relatives aux paramètres (« significativité » de la régression, etc.) ;
— prédiction de Y ou $\mathbb{E}(Y)$ pour une nouvelle valeur de $\boldsymbol{x}$;
— diagnostic de la régression (validation du modèle) ;
— sélection d'un ensemble de variables explicatives « pertinentes ».

2 Estimation des paramètres

2.1 Estimateur des moindres carrés de b

Soit β un estimateur du paramètre vectoriel $\mathbf{b}$. La méthode des moindres carrés consiste à choisir β de façon à minimiser la somme des carrés des écarts entre les observations Y_i et les prédictions $\widehat{Y}_i = \boldsymbol{x}_i^T \beta$. On cherche donc à minimiser le critère

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \beta)^2 \\ &= (\boldsymbol{Y} - X\beta)^T (\boldsymbol{Y} - X\beta) \\ &= \boldsymbol{Y}^T \boldsymbol{Y} - 2\beta^T X^T \boldsymbol{Y} + \beta^T X^T X \beta. \end{aligned} \tag{14.1}$$

Théorème 1. *Le minimum de la fonction $S(\beta)$ est obtenu pour $\beta = \widehat{\mathbf{b}}$ avec*

$$\widehat{\mathbf{b}} = (X^T X)^{-1} X^T \boldsymbol{Y},$$

c'est-à-dire que l'on a

$$S(\widehat{\mathbf{b}}) = \min_{\beta} S(\beta).$$

$\widehat{\mathbf{b}}$ *est appelé* estimateur des moindres carrés *de* $\mathbf{b}$.

Preuve : Il s'agit d'une fonction de $p+1$ variables. Pour en trouver le minimum, il suffit d'annuler le gradient, c'est-à-dire le vecteur des dérivées partielles :

$$\nabla S = \left(\frac{\partial S}{\partial \beta_0}, \dots, \frac{\partial S}{\partial \beta_p} \right),$$

ce qui conduit à un système de $p+1$ équations à $p+1$ inconnues. Il vient ici

$$\nabla S = -2X^T \boldsymbol{Y} + 2X^T X \beta = 0. \tag{14.2}$$

En supposant $X^T X$ inversible, on obtient directement la solution de ce système :

$$X^T X \beta = X^T \boldsymbol{Y} \Leftrightarrow \beta = (X^T X)^{-1} X^T \boldsymbol{Y}.$$

Il reste à vérifier qu'il s'agit bien d'un maximum. Soit

$$\widehat{\mathbf{b}} = (X^T X)^{-1} X^T \boldsymbol{Y}$$

la solution trouvée précédemment, et

$$\widehat{\varepsilon} = \boldsymbol{Y} - X\widehat{\mathbf{b}}$$

le vecteur des écarts (appelés résidus). On a donc

$$S(\widehat{\mathbf{b}}) = \widehat{\varepsilon}^T\widehat{\varepsilon}.$$

Soit $\widetilde{\mathbf{b}}$ une autre valeur de β. Le vecteur des écarts correspondant est

$$\widetilde{\varepsilon} = \boldsymbol{Y} - X\widetilde{\mathbf{b}} = (\boldsymbol{Y} - X\widehat{\mathbf{b}}) + (X\widehat{\mathbf{b}} - X\widetilde{\mathbf{b}}) = \widehat{\varepsilon} + X(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}}).$$

On a donc

$$\widetilde{\varepsilon}^T\widetilde{\varepsilon} = \widehat{\varepsilon}^T\widehat{\varepsilon} + 2(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})^T X^T\widehat{\varepsilon} + (\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})^T X^T X(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}}).$$

Le terme central du membre de gauche s'écrit

$$2(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})^T X^T\widehat{\varepsilon} = 2(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})^T X^T(\boldsymbol{Y} - X\widehat{\mathbf{b}}).$$

Or, d'après (14.2)

$$X^T(\boldsymbol{Y} - X\widehat{\mathbf{b}}) = 0,$$

donc ce terme est nul et l'on a finalement

$$S(\widetilde{\mathbf{b}}) = S(\widehat{\mathbf{b}}) + (\widehat{\mathbf{b}} - \widetilde{\mathbf{b}})^T X^T X(\widehat{\mathbf{b}} - \widetilde{\mathbf{b}}).$$

Le dernier terme est une somme de carrés et ne peut donc être que positif ou nul. Par conséquent $S(\widetilde{\mathbf{b}}) \geq S(\widehat{\mathbf{b}})$. En conclusion, l'estimateur des moindres carrés de $\mathbf{b}$ est donc bien $\widehat{\mathbf{b}}$.

□

On notera

$$\widehat{\boldsymbol{Y}} = X\widehat{\mathbf{b}} = X(X^TX)^{-1}X^T\boldsymbol{Y}$$

le vecteur des prédictions obtenu en remplaçant le paramètre $\mathbf{b}$ inconnu par son estimateur des moindres carrés $\widehat{\mathbf{b}}$.

Remarque 2. *On a supposé X^TX inversible, ce qui est le cas si la matrice X est de rang $p+1$. Si ce n'est pas le cas, c'est qu'une variable (une colonne de X) s'exprime comme combinaison linéaire des autres. Il suffit alors de supprimer la ou les variables redondantes.*

Remarque 3. *Si certaines variables sont très corrélées, la matrice X^TX est mal conditionnée est les calculs numériques peuvent être très imprécis. Une solution (appelée* ridge regression *en anglais) consiste à ajouter un terme sur la diagonale de X^TX :*

$$\widehat{\mathbf{b}}_\lambda = (X^TX + \lambda I)^{-1}X^T\boldsymbol{Y}$$

où λ est une constante à déterminer. On montre que l'on améliore ainsi parfois les propriétés de l'estimateur. Une autre solution consiste a faire une analyse en composante principale préalable du tableau X, et à utiliser les composantes principales comme nouvelles variables (en supprimant celles correspondant à des valeurs propres nulles ou très faible). Cette technique porte le nom de régression sur composantes principales.

Remarque 4. *En pratique, il est inutile d'inverser la matrice X^TX : il existe des algorithmes permettant de résoudre directement le système (14.2). En Matlab, on obtient directement la solution de* `Xb=Y` *par la commande* `b=X\Y`.

Remarque 5. *On a*

$$\widehat{\boldsymbol{Y}} = X\widehat{\mathbf{b}} = X(X^TX)^{-1}X^T\boldsymbol{Y} = P\boldsymbol{Y}$$

en notant $P = X(X^TX)^{-1}X^T$*. Cette matrice* P *a des propriétés remarquables. En effet,* P *est symétrique (évident), et de plus*

$$P^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = P.$$

La matrice P *est donc idempotente (c'est un opérateur de projection orthogonale, comme nous le verrons par la suite). De même, on peut écrire*

$$\widehat{\varepsilon} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (I_n - P)\boldsymbol{Y} = R\boldsymbol{Y}$$

avec $R = I_n - P$*. On vérifie aisément que* R *a les mêmes propriétés que* P *(symétrie et idempotence) : c'est également un opérateur de projection orthogonale.*

2.2 Propriétés de $\widehat{\mathbf{b}}$

Il est facile de calculer l'espérance et la variance de $\widehat{\mathbf{b}}$. On a les propriétés suivantes.

Théorème 2. $\widehat{\mathbf{b}}$ *est un estimateur sans biais de* $\mathbf{b}$*, et*

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \sigma^2(X^TX)^{-1}.$$

Preuve : En effet,

$$\begin{aligned}\widehat{\mathbf{b}} &= (X^TX)^{-1}X^T(X\mathbf{b} + \varepsilon)\\ &= (X^TX)^{-1}(X^TX)\mathbf{b} + (X^TX)^{-1}X^T\varepsilon\\ &= \mathbf{b} + (X^TX)^{-1}X^T\varepsilon\end{aligned} \tag{14.3}$$

D'où

$$\mathbb{E}(\widehat{\mathbf{b}}) = \mathbf{b} + (X^TX)^{-1}X^T\mathbb{E}(\varepsilon) = \mathbf{b}.$$

Donc $\widehat{\mathbf{b}}$ est sans biais. Calculons sa variance :

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \mathbb{E}[(\widehat{\mathbf{b}} - \mathbf{b})(\widehat{\mathbf{b}} - \mathbf{b})^T].$$

D'après ce qui précède,

$$\widehat{\mathbf{b}} - \mathbf{b} = (X^TX)^{-1}X^T\varepsilon, \tag{14.4}$$

d'où, compte-tenu du fait que X^TX est symétrique ($(X^TX)^T = X^TX$) :

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \mathbb{E}[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}] = (X^TX)^{-1}X^T\mathbb{E}[\varepsilon\varepsilon^T]X(X^TX)^{-1}.$$

Or, $\mathbb{E}[\varepsilon\varepsilon^T] = \mathrm{Var}(\varepsilon) = \sigma^2 I_n$. Donc

$$\mathrm{Var}(\widehat{\mathbf{b}}) = \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.$$

□

Théorème 3 (Théorème de Gauss-Markov)**.** *L'estimateur des moindres carrés* $\widehat{\mathbf{b}}$ *est optimal dans la classe* $\mathcal{C}$ *des estimateurs sans biais de* $\mathbf{b}$ *linéaires en* $Y_1, \ldots, Y_n$*.*

Cela signifie que, pour n'importe estimateur $\widetilde{\mathbf{b}}$ dans $\mathcal{C}$, la matrice $A = \mathrm{Var}(\widetilde{\mathbf{b}}) - \mathrm{Var}(\widehat{\mathbf{b}})$ est semi-définie positive : on a $\boldsymbol{x}^TA\boldsymbol{x} \geq 0$ pour tout $\boldsymbol{x} \in \mathbb{R}^{p+1}$. Ce résultat est admis.

Théorème 4. *On a :* $\mathrm{Cov}(\widehat{\mathbf{b}}, \widehat{\varepsilon}) = 0$.

Preuve : Par définition

$$\mathrm{Cov}(\widehat{\mathbf{b}},\widehat{\varepsilon}) = \mathbb{E}[(\widehat{\mathbf{b}}-\mathbf{b})(\widehat{\varepsilon}-\mathbb{E}(\widehat{\varepsilon}))^T].$$

Commençons par calculer $\mathbb{E}(\widehat{\varepsilon})$. On a

$$\widehat{\varepsilon} = R\boldsymbol{Y} = R(X\mathbf{b}+\varepsilon) = RX\mathbf{b} + R\epsilon.$$

Or, $RX = (I_n - X(X^TX)^{-1}X^T)X = X - X = 0$. Donc $\widehat{\varepsilon} = R\varepsilon$, et $\mathbb{E}(\widehat{\varepsilon}) = 0$. En utilisant l'équation (14.4) et sachant que $\widehat{\varepsilon} = R\varepsilon$, on a donc

$$\begin{aligned}\mathrm{Cov}(\widehat{\mathbf{b}},\widehat{\varepsilon}) &= \mathbb{E}[(X^TX)^{-1}X^T\varepsilon\varepsilon^TR^T] \\ &= \sigma^2(X^TX)^{-1}X^T[I_n - X(X^TX)^{-1}X^T] = 0.\end{aligned}$$

□

2.3 Estimation de σ^2

Il reste à estimer la variance σ^2 des perturbations. Il est naturel pour cela de s'intéresser aux résidus $\widehat{\varepsilon}_i$, et plus particulièrement à la somme des carrés des résidus.

Théorème 5. *La variance des perturbations σ^2 est estimée sans biais par*

$$\widehat{\sigma}^2 = \frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2.$$

Preuve : On a

$$\widehat{\varepsilon}^T\widehat{\varepsilon} = (R\varepsilon)^TR\varepsilon = \varepsilon^TR\varepsilon = \sum_{i=1}^{n}\sum_{j=1}^{n}R_{ij}\varepsilon_i\varepsilon_j,$$

d'où

$$\mathbb{E}(\widehat{\varepsilon}^T\widehat{\varepsilon}) = \sigma^2\sum_{i=1}^{n}R_{ii} = \sigma^2\,\mathrm{Tr}\,R.$$

Or, on sait qu'un opérateur idempotent a toutes ses valeurs propres égales à 0 ou 1, donc sa trace est égale à son rang. Le rang de R est $n-p-1$ d'où le résultat.

□

3 Analyse de la variance

3.1 Point de vue géométrique

Plaçons nous dans $\mathbb{R}^n$ et considérons les vecteurs

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \; j = 1,p \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Le modèle linéaire s'écrit avec ces notations

$$\boldsymbol{Y} = b_0\mathbb{1} + \sum_{j=1}^{p}b_j\boldsymbol{x}_j + \varepsilon.$$

La méthode des moindres carrés peut être interprétée comme la recherche de la meilleure approximation de $\boldsymbol{Y}$ dans le sous-espace $\mathcal{L}$ de $\mathbb{R}^n$ engendré par les $p+1$ vecteurs $\mathbb{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. On cherche en effet

$$\widehat{\boldsymbol{Y}} = \widehat{b}_0 \mathbb{1} + \sum_{j=1}^{p} \widehat{b}_j \boldsymbol{x}_j \in \mathcal{L}$$

tel que la distance euclidienne $\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}}\|^2$ soit minimum. On sait que la solution consiste à définir $\widehat{\boldsymbol{Y}}$ comme la projection orthogonale de $\boldsymbol{Y}$ sur $\mathcal{L}$. On a vu en effet que

$$\widehat{\boldsymbol{Y}} = P\boldsymbol{Y},$$

P étant un opérateur de projection orthogonale.

Cette représentation géométrique permet de retrouver sans calculs fastidieux plusieurs résultats intéressants. Tout d'abord, on a

$$\widehat{\varepsilon} \perp \mathbb{1} \Rightarrow \sum_{i=1}^{n} \widehat{\varepsilon}_i = 0,$$

d'où l'on déduit

$$\frac{1}{n}\widehat{Y}_i = \frac{1}{n} Y_i = \overline{Y}.$$

Par ailleurs, la projection orthogonale de $\boldsymbol{Y}$ sur l'axe dirigé par $\mathbb{1}$ a pour coordonnée

$$\frac{< \boldsymbol{Y}, \mathbb{1} >}{\|\mathbb{1}\|} = \overline{Y}.$$

Il en est de même, d'après ce qui précède, pour la projection orthogonale de $\widehat{\boldsymbol{Y}}$ sur $\mathbb{1}$. Enfin, on a de manière évidente :

$$\widehat{\boldsymbol{Y}} \perp \widehat{\varepsilon}.$$

3.2 Equation d'analyse de la variance

Notons $\overline{\boldsymbol{Y}} = \overline{Y}\mathbb{1}$. En appliquant le théorème de Pythagore au triangle $(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}, \overline{\boldsymbol{Y}})$, on obtient finalement la relation très importante suivante, appelée *équation d'analyse de la variance* :

$$\|\boldsymbol{Y} - \overline{\boldsymbol{Y}}\|^2 = \|\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\|^2 + \|\widehat{\varepsilon}\|^2,$$

ce que l'on peut encore écrire, en divisant par n :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2$$

soit encore

$$S_{YY} = S_{reg} + S_{res}.$$

Cette équation est appelée *équation d'analyse de la variance*. Le terme de gauche (S_{YY}) est la variance empirique des Y_i, il caractérise la dispersion des valeurs observées de la variable à expliquer. Le premier terme du membre de droite (S_{reg}) est la variance empirique des $\widehat{Y}_i$, que l'on appelle variance expliquée par la régression. Le second terme du membre de droite (S_{res}) est la variance des résidus, ou variance résiduelle.

Remarque 6. *La variance résiduelle est liée à l'estimateur sans biais $\widehat{\sigma}^2$ de σ^2 :*

$$\widehat{\sigma}^2 = \frac{n}{n-p-1} S_{res}.$$

Remarque 7. *A chacun des termes de l'équation d'analyse de la variance est associé un nombre de degrés de liberté (d.d.l.), égal au nombre de combinaisons linéaires des Y_i utilisées dans le calcul :*

TABLE 14.1 – Tableau d'analyse de la variance (SS : *sum of squares* ; MS : *mean square*).

source de variation	d.d.l.	SS	MS=SS/d.d.l.
régression	p	$\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$	$\frac{1}{p}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$
résiduelle	$n-p-1$	$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$\frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \widehat{\sigma}^2$
totale	$n-1$	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

— *S_{YY} dépend de n quantités $Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}$ liées par la relation*

$$\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0.$$

Ce terme a donc $n-1$ d.d.l.

— *On a $\widehat{Y}_i = \boldsymbol{x}_i^T\widehat{\mathbf{b}}$ et $\overline{Y} = \overline{\boldsymbol{x}}^T\widehat{\mathbf{b}}$. Par conséquent, le terme S_{reg} est fonction des paramètres $\widehat{b}_1, \ldots, \widehat{b}_p$ (le terme $\widehat{b}_0$ s'annule dans chacune des différences $\widehat{Y}_i - \overline{Y}$). La variance expliquée a donc p d.d.l.*

— *Par conséquent, le nombre de d.d.l associé à la variance résiduelle est $n-p-1$.*

La plupart des logiciels statistiques présentent les résultats de la régression sous forme d'un tableau (appelé *tableau d'analyse de la variance*), où figurent les différents termes de l'équation d'analyse de la variance, et les nombres de d.d.l associés (cf. tableau 14.1).

3.3 Evaluation de la qualité de l'ajustement

On définit à partir de l'équation d'analyse de la variance le *coefficient de détermination*, égal à la proportion de la variance totale expliquée par la régression :

$$R^2 = \frac{S_{reg}}{S_{YY}} = 1 - \frac{S_{res}}{S_{YY}}.$$

Ce coefficient traduit la « qualité de l'ajustement », comme on le voit en considérant les deux situations extrêmes suivantes :

— Si les résidus sont nuls, on a $S_{res} = 0$ et $R^2 = 1$. Les n points $(\boldsymbol{x}_i, Y_i) \in \mathbb{R}^{p+1}$ sont alors situés dans l'hyperplan d'équation

$$Y = \widehat{b}_0 + \widehat{b}_1 x_1 + \ldots + \widehat{b}_p x_p.$$

Cela signifie que l'on peut retrouver sans erreur les Y_i à partir des $\boldsymbol{x}_i$, c'est-à-dire que toute la variation des Y_i est expliquée par les $\boldsymbol{x}_i$.

— Si les prédictions sont constantes ($\widehat{Y}_i = \overline{Y}, \forall i$), la variance expliquée est nulle et $R^2 = 0$. Dans ce cas, les $\boldsymbol{x}_i$ n'expliquent pas du tout la variation des Y_i.

— De manière générale, on a $0 \leq R^2 \leq 1$, et la valeur du R^2 s'interprète comme un « degré de liaison » entre les variables explicatives et la variable à expliquer.

Remarque 8. *Géométriquement, R^2 est égal au carré du cosinus de l'angle θ entre les vecteurs $\boldsymbol{Y} - \overline{\boldsymbol{Y}}$ et $\widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}$: c'est donc le carré du coefficient de corrélation linéaire entre les Y_i et les $\widehat{Y}_i$.*

Remarque 9. *La définition du R^2 est telle que sa valeur augmente « mécaniquement » avec le nombre de variables explicatives : si on augmente la dimension du sous-espace $\mathcal{L}$, la distance de $\boldsymbol{Y}$ au sous-espace, et donc la variance résiduelle, ne peuvent que diminuer.*

Considérons l'exemple suivant. Supposons que l'on ait $p = 1$ variable explicative x, et $n = 4$ observations. Si l'on introduit 2 variables explicatives supplémentaires $x_2 = x^2$ et $x_3 = x^3$, le modèle devient

$$Y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \varepsilon.$$

Il est alors toujours possible de déterminer les 4 coefficients b_i de façon à avoir $\widehat{Y}_i = Y_i, \forall i$, et donc $R^2 = 1$, sans que cela ne traduise l'existence d'une relation plausible entre les variables x et y.

Pour obtenir un indicateur plus fiable de la qualité du modèle (et permettant la comparaison de plusieurs modèles ne possédant pas le même nombre de variables explicatives), on définit le R^2 ajusté comme suit :

$$\begin{aligned}\overline{R}^2 &= 1 - \frac{\frac{n}{n-p-1} S_{res}}{\frac{n}{n-1} S_{YY}} = 1 - \frac{n-1}{n-p-1}\frac{S_{res}}{S_{YY}} \\ &= 1 - \frac{n-1}{n-p-1}(1-R^2) = \frac{n-1}{n-p-1}R^2 - \frac{p}{n-p-1}.\end{aligned}$$

On remarque que l'on a toujours $\overline{R}^2 \leq R^2$.

4 Tests de significativité

4.1 Loi des estimateurs sous hypothèse gaussienne

Il est souvent intéressant de déterminer si les résultats de la régression (coefficients $\widehat{b}_j$ et R^2) sont dus au hasard, ou s'ils traduisent l'existence d'une relation « significative » entre la variable à expliquer et les variables explicatives. Pour cela, il existe des tests de significativité, qui nécessitent une hypothèse supplémentaire : la normalité des perturbations :

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

En effet, sous cette hypothèse, il est possible de préciser la loi des estimateurs de $\mathbf{b}$ et de σ^2. En effet, on a

$$\widehat{\mathbf{b}} = [(X^T X)^{-1} X^T] Y.$$

Or, $Y \sim \mathcal{N}(X\mathbf{b}, \sigma^2 I)$. Donc $\widehat{\mathbf{b}}$ suit une loi normale dans $\mathbb{R}^{p+1}$. Comme on a déjà calculé l'espérance et la variance de $\widehat{\mathbf{b}}$ dans le cas général, on a finalement

$$\widehat{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}, \sigma^2 (X^T X)^{-1}).$$

Par ailleurs, on a

$$(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{1}{\sigma^2}\varepsilon^T R \varepsilon,$$

car $\widehat{\varepsilon} = R\varepsilon$. Or, on a le théorème suivant (admis) :

Théorème 6. *Si $\boldsymbol{X}$ est un vecteur gaussien centré-réduit à composantes indépendantes, la forme quadratique $Q = \boldsymbol{X}^T A \boldsymbol{X}$ suit une loi du χ^2 si et seulement si A est un projecteur orthogonal, c'est-à-dire si $A^2 = A$. Le rang de A est alors le degré de liberté du χ^2.*

La matrice R étant un projecteur orthogonal de rang $n-p-1$ (c'est la matrice de projection sur $\mathcal{L}^{\perp}$), on a

$$(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}. \tag{14.5}$$

4.2 Test de significativité d'un coefficient de régression

Soient les hypothèses suivantes :

$$\begin{cases} H_0 : & b_j = 0 \\ H_1 : & b_j \neq 0 \end{cases}$$

L'hypothèse H_0 signifie que la variable x_j n'est pas liée à (n'apporte aucune information sur) Y. On a vu que

$$\widehat{\mathbf{b}} \sim \mathcal{N}(\mathbf{b}, \sigma^2 (X^T X)^{-1}),$$

donc

$$\widehat{b}_j \sim \mathcal{N}(b_j, \sigma^2 v_j),$$

v_j désignant le terme diagonal (j, j) de la matrice $(X^T X)^{-1}$. On peut encore écrire

$$\frac{\widehat{b}_j - b_j}{\sigma\sqrt{v_j}} \sim \mathcal{N}(0, 1).$$

D'après (14.5), et en remarquant que $\widehat{b}_j$ et $\widehat{\sigma}^2$ sont indépendants (conséquence du Théorème 4), on a

$$\frac{\widehat{b}_j - b_j}{\widehat{\sigma}\sqrt{v_j}} \sim \mathcal{T}_{n-p-1}.$$

Sous H_0, on a donc

$$\frac{\widehat{b}_j}{\widehat{\sigma}\sqrt{v_j}} \sim \mathcal{T}_{n-p-1},$$

d'où la région critique du test, au niveau de signification α :

$$W : \frac{|\widehat{b}_j|}{\widehat{\sigma}\sqrt{v_j}} > t_{n-p-1;1-\alpha/2}.$$

4.3 Test de significativité du R^2

Considérons maintenant les hypothèses :

$$\begin{cases} H_0 : & b_1 = b_2 = \ldots = b_p = 0 \\ H_1 : & \exists j \in \{1, \ldots, p\}\ b_j \neq 0 \end{cases}$$

L'hypothèse nulle signifie qu'il n'y a aucune liaison entre les variables explicatives et Y : le R^2 obtenu est donc non significatif, c'est-à-dire purement « accidentel ».

Reprenons l'équation de la variance :

$$S_{YY} = S_{reg} + S_{res}.$$

On a vu que

$$\frac{nS_{res}}{\sigma^2} = (n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1},$$

quelque soit $\mathbf{b}$. Par ailleurs, sous H_0, les v.a. Y_i ont toutes la même loi, et donc nS_{YY}/σ^2 suit un χ^2_{n-1} comme variance empirique d'un échantillon de v.a. indépendantes de même loi. Enfin, S_{reg} ne dépend que de $\widehat{\mathbf{b}}$ et S_{res} ne dépend que de $\widehat{\varepsilon}$, donc ces 2 termes sont indépendants d'après le théorème 4. On en déduit que, sous l'hypothèse H_0, $nS_{reg}/\sigma^2 \sim \chi^2_p$.

Soit la statistique

$$F = \frac{nS_{reg}/\sigma^2 p}{nS_{res}/\sigma^2(n-p-1)} = \frac{S_{reg}/p}{S_{res}/(n-p-1)}.$$

D'après ce qui précède, F suit sous H_0 une loi de Fisher $F_{p,n-p-1}$. Sous H_1, le rapport de la variance expliquée à la variance résiduelle a tendance à prendre des valeurs plus élevées, d'où la région critique :

$$W : F > F_{p,n-p-1;1-\alpha}$$

Remarquons que F peut également s'exprimer en fonction de R^2. En effet,

$$F = \frac{n-p-1}{p}\frac{S_{reg}}{S_{res}} = \frac{n-p-1}{p}\frac{S_{reg}/S_{YY}}{1-S_{reg}/S_{YY}} = \frac{n-p-1}{p}\frac{R^2}{1-R^2},$$

d'où l'autre expression de la région critique :

$$W : \frac{n-p-1}{p}\frac{R^2}{1-R^2} > F_{p,n-p-1;1-\alpha}$$

Remarque 10. *Il peut arriver (rarement), que le test F soit significatif, et que chacun des tests t pour les hypothèses $H_0 : b_j = 0$ soit non significatif. L'inverse (test F non significatif mais certains coefficients significativement non nuls) n'est pas non plus impossible (mais encore plus rare).*

4.4 Test d'une sous-hypothèse linéaire

Il s'agit de tester l'hypothèse selon laquelle q coefficients sont nuls. Moyennant une permutation des indices, on peut toujours supposer que ce sont les q premiers, et écrire ainsi les hypothèses :

$$\begin{cases} H_0 : & b_1 = b_2 = \ldots = b_q = 0 \\ H_1 : & \exists j \in \{1, \ldots, q\}\ b_j \neq 0 \end{cases}$$

Pour résoudre ce test, il faut faire deux fois la régression, avec le modèle réduit (obtenu en ne prenant que les variables $q+1, \ldots, p$) et avec le modèle complet. Appelons S^0_{res} et S^1_{res} la variance résiduelle, respectivement dans le modèle réduit et dans le modèle complet, et considérons la statistique

$$F = \frac{S^0_{res} - S^1_{res}}{S^1_{res}}\frac{n-p-1}{q}.$$

En remarquant que $S_{res} = S_{YY}(1-R^2)$, F s'écrit encore

$$F = \frac{R^2_1 - R^2_0}{1-R^2_1}\frac{n-p-1}{q}$$

R^2_0 et R^2_1 désignant le coefficient de détermination dans le modèle réduit et dans le modèle complet. On peut montrer que, sous H_0,

$$F \sim F_{q,n-p-1},$$

d'où la région critique du test :

$$W : \frac{R^2_1 - R^2_0}{1-R^2_1}\frac{n-p-1}{q} > F_{q,n-p-1;1-\alpha}$$

Ce test est très utile pour juger de la pertinence d'un ensemble de variables explicatives potentielles, en donnant un critère de significativité de l'augmentation du R^2 observée lorsqu'on complexifie le modèle initial.

5 Prédiction

Lorsque les paramètres du modèle ont été estimés, et en supposant ce modèle valide, il est possible de l'utiliser pour *prédire* la valeur que prendra la variable Y pour de nouvelles valeurs des variables explicatives.

Posons $\boldsymbol{x}_0 = (1, x_{01}, \ldots, x_{0p})^T$ le vecteur des variables d'entrée du modèle pour un nouvel individu. La sortie correspondante est

$$Y_0 = \boldsymbol{x}_0^T \mathbf{b} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

La quantité $\widehat{Y}_0 = \boldsymbol{x}_0^T \widehat{\mathbf{b}}$ fournit une prédiction non biaisée de Y_0, dans le sens où

$$\mathbb{E}(\widehat{Y}_0) = \boldsymbol{x}_0^T \mathbb{E}(\widehat{\mathbf{b}}) = \boldsymbol{x}_0^T \mathbf{b} = \mathbb{E}(Y_0).$$

Il s'agit cependant d'une prédiction ponctuelle. Dans la pratique, il est important de donner une indication sur la « fiabilité » de la prédiction, ce que l'on peut faire en donnant :

- un intervalle de confiance sur $\mathbb{E}(Y_0)$ (un intervalle aléatoire contenant la constante $\mathbb{E}(Y_0)$ dans $100(1-\alpha)$ % des cas) ;
- un intervalle de prévision (un intervalle aléatoire contenant la v.a. Y_0 dans $100(1-\alpha)$ % des cas).

Commençons par remarquer que $\widehat{Y}_0$ suit une loi normale. Il nous reste donc pour déterminer sa loi à calculer sa variance. On a

$$\mathrm{Var}(\widehat{Y}_0) = \boldsymbol{x}_0^T \,\mathrm{Var}(\widehat{\mathbf{b}}) \boldsymbol{x}_0 = \boldsymbol{x}_0^T [\sigma^2 (X^T X)^{-1}] \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0.$$

On a donc

$$\widehat{Y}_0 \sim \mathcal{N}(\boldsymbol{x}_0^T \mathbf{b}, \sigma^2 \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0).$$

On en déduit la fonction pivotale

$$\frac{\widehat{Y}_0 - \boldsymbol{x}_0^T \mathbf{b}}{\widehat{\sigma} \sqrt{\boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0}} \sim \mathcal{T}_{n-p-1},$$

qui conduit à l'intervalle de confiance suivant (au niveau de confiance $1-\alpha$) :

$$1 - \alpha = P\left[\widehat{Y}_0 - t_{n-p-1;1-\alpha/2} \widehat{\sigma} \sqrt{\boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0} < \mathbb{E}(Y_0) \right.$$
$$\left. < \widehat{Y}_0 + t_{n-p-1;1-\alpha/2} \widehat{\sigma} \sqrt{\boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0} \right].$$

Pour calculer un intervalle de prévision, on remarque que

$$Y_0 \sim \mathcal{N}(\boldsymbol{x}_0^T \mathbf{b}, \sigma^2)$$

d'où

$$\widehat{Y}_0 - Y_0 \sim \mathcal{N}(0, \sigma^2 (1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0)).$$

On en déduit la fonction pivotale

$$\frac{\widehat{Y}_0 - Y_0}{\widehat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0}} \sim \mathcal{T}_{n-p-1},$$

et l'intervalle de prévision au niveau de confiance $1-\alpha$:

$$1 - \alpha = P\left[\widehat{Y}_0 - t_{n-p-1;1-\alpha/2} \widehat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0} < Y_0 \right.$$
$$\left. < \widehat{Y}_0 + t_{n-p-1;1-\alpha/2} \widehat{\sigma} \sqrt{1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0} \right].$$

On remarque que l'intervalle de prévision est plus large que l'intervalle de confiance.

6 Diagnostic de la régression

La phase de diagnostic de la régression consiste à vérifier (de manière plus ou moins subjective) que les hypothèses du modèle (linéarité de la relation entre les x_j et y, homoscédasticité, normalité des perturbations) sont adaptées aux données.

L'examen des résidus joue un rôle fondamental. Il permet non seulement de vérifier empiriquement les hypothèses du modèle, mais également de détecter les observations atypiques (points aberrants) et de repérer les observations qui jouent un rôle important dans la détermination de la régression.

On appelle résidus bruts les quantités $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Afin de s'affranchir de facteurs d'échelle, il est utile de normaliser les résidus. Pour cela on utilise le résultat suivant :

Proposition 7. $\mathrm{Var}(\widehat{\varepsilon}) = \sigma^2 R$

Preuve : On a vu que $\widehat{\varepsilon} = R\varepsilon$ et $\mathbb{E}(\widehat{\varepsilon}) = 0$. On a donc

$$\mathrm{Var}(\widehat{\varepsilon}) = \mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}^T) = R\mathbb{E}(\varepsilon\varepsilon^T)R^T = \sigma^2 RR^T = \sigma^2 R.$$

□

Soit $r_i = R_{ii}$ le terme diagonal (i,i) de la matrice R. On a donc

$$\mathrm{Var}(\widehat{\varepsilon}_i) = r_i\sigma^2$$

qui peut être estimé par $r_i\widehat{\sigma}^2$. On appelle *résidus studentisés* les quantités

$$s_i = \frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{r_i}}.$$

Remarque 11. *Malgré l'appellation « résidus studentisés », les s_i ne suivent pas une loi de Student ($\widehat{\sigma}^2$ n'est pas indépendant de $\widehat{\varepsilon}_i$).*

Afin de vérifier les hypothèses du modèle, on croise les résidus (bruts ou studentisés) avec les variables explicatives x_j et les prédictions $\widehat{Y}$ (le croisement avec les Y_i a moins d'intérêt, car les résidus sont en général corrélés avec les Y_i). A l'examen de ces graphiques, on ne doit pas déceler de structure particulière (les points doivent être répartis de manière aléatoire à l'intérieur d'une bande de largeur à peu près constante). Sous hypothèse de normalité des perturbations, les résidus studentisés doivent par ailleurs être pratiquement tous compris entre -2 et $+2$. Si certaines hypothèses apparaissent comme non vérifiées, il faut modifier le modèle (transformation des Y_i, utilisation de modèles plus complexes qui sortent du cadre de ce cours).

L'examen des résidus n'est pas toujours suffisant pour détecter les points aberrants à cause de l'effet de levier : un point aberrant peut avoir une grande influence sur les coefficients de régression et avoir ainsi un résidu faible.

Pour mettre en évidence ce type d'effet (influence « anormale » de certaines observations sur les résultats de la régression), on introduit les quantités suivantes, appelées *distances de Cook* :

$$D_i = \frac{\|\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-i)}\|^2}{(p+1)\widehat{\sigma}^2}$$

avec $\widehat{\boldsymbol{Y}} = X\widehat{\mathbf{b}}$ et $\widehat{\boldsymbol{Y}}_{(-i)} = X\widehat{\mathbf{b}}_{(-i)}$, $\widehat{\mathbf{b}}_{(-i)}$ étant l'estimation du vecteur des coefficients de régression obtenu en supprimant de l'ensemble d'apprentissage l'individu i (la ligne i de la matrice X et du vecteur $\boldsymbol{Y}$). La quantité D_i caractérise l'influence de l'observation i sur le résultat de la régression, une valeur élevée pouvant révéler une influence « anormale ».

Remarquons que $\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(-i)} = X(\widehat{\mathbf{b}} - \widehat{\mathbf{b}}_{(-i)})$, d'où

$$D_i = \frac{(\widehat{\mathbf{b}} - \widehat{\mathbf{b}}_{(-i)})^T X^T X(\widehat{\mathbf{b}} - \widehat{\mathbf{b}}_{(-i)})}{(p+1)\widehat{\sigma}^2}$$

ce qui montre que D_i peut également s'interpréter comme le carré d'une distance entre les deux vecteurs $\widehat{\mathbf{b}}$ et $\widehat{\mathbf{b}}_{(-i)}$. On montre également que

$$D_i = \left[\frac{\widehat{\varepsilon}_i}{\widehat{\sigma}\sqrt{r_i}}\right]^2 \left[\frac{1-r_i}{r_i}\right] \frac{1}{p+1},$$

où comme précédemment r_i est le terme diagnonal (i,i) de la matrice R. Il est donc inutile pour calculer les distances de Cook de refaire n fois les calculs de la régression.

7 Sélection des variables explicatives

Un dernier point important à considérer est le choix des variables explicatives. A partir d'un ensemble de variables connues susceptibles d'influer sur Y, il est possible de construire (à partir de transformations non linéaires : log, puissance, etc.) un nombre potentiellement très grand de variables explicatives x_i. En pratique, le nombre de variables à inclure dans le modèle doit correspondre à un compromis :

— en augmentant le nombre de variables, on intègre de plus en plus d'information dans le modèle ;

— mais on augmente aussi la variance des estimations $\widehat{Y}_i$, car on augmente le nombre de paramètres à estimer.

En effet, on a

$$\mathrm{Var}(\widehat{\boldsymbol{Y}}) = \mathbb{E}[X(\widehat{\mathbf{b}}-\mathbf{b})(\widehat{\mathbf{b}}-\mathbf{b})^T X^T] = X\,\mathrm{Var}(\widehat{\mathbf{b}})X^T = \sigma^2 X(X^TX)^{-1}X^T = \sigma^2 P.$$

La variance moyenne des $\widehat{W}_i$ est donc

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\widehat{Y}_i) = \sigma^2\frac{\mathrm{Tr}(P)}{n} = \sigma^2\frac{p+1}{n}.$$

On a donc intérêt à réduire p.

Pour cela, il faut choisir (1) un critère de qualité du modèle, et (2) une stratégie de sélection.

Le critère R^2 n'est pas un bon choix en général car il est monotone (on ne peut qu'augmenter le R^2 en ajoutant de nouvelle variables). Une alternative intéressante consiste à utiliser le R^2 ajusté, qui est un critère non monotone. Cela revient également à utiliser comme critère la variance résiduelle $\widehat{\sigma}^2$. En effet, on a

$$\widehat{\sigma}^2 = \frac{n}{n-p-1}S_{res}$$

et

$$\overline{R}^2 = 1 - \frac{\frac{n}{n-p-1}S_{res}}{\frac{n}{n-1}S_{YY}}$$

d'où l'on déduit

$$\widehat{\sigma}^2 = \frac{n}{n-1}(1-\overline{R}^2)S_{YY}.$$

En ce qui concerne la stratégie de sélection de m variables parmi p variables initiales, on peut envisager, si p n'est pas trop grand, une recherche exhaustive (choix du meilleur sous-ensemble de variables parmi les p, au sens du critère retenu). Le nombre de sous-ensemble à tester est alors égal à $2^p - 1$, soit 31 pour $p = 5$, 1023 pour $p = 10$, 1048575 pour $p = 20$! En pratique, cette solution n'est donc faisable que pour une dizaine de variables initiales.

Quand p est grand, il faut par conséquent avoir recours à une démarche heuristique sous-optimale. On utilise le plus souvent une procédure pas à pas consistant en l'élimination successive ou l'ajout successif de variables. On distingue notamment :

— la sélection ascendante : on ajoute incrémentalement des variables en maximisant à chaque fois le critère $\overline{R}^2$ (on cherche à chaque pas la variables qui fait décroître le plus la variance résiduelle) ;

— la sélection descendante : on commence avec les p variables, puis on retire à chaque pas la variable dont la suppression fait croître le moins la variance résiduelle.

Troisième partie

Annexes

Annexe A

Rappels et compléments de probabilité

1 Introduction

Les méthodes étudiées au chapitre précédent visent à décrire de manière synthétique un ensemble d'observations relatives à n individus d'une population. Très souvent, cependant, ces individus ne représentent pas la totalité de la population, mais un sous-ensemble, appelé échantillon, à partir duquel on cherche à tirer des conclusions relatives à la population entière.

Les conclusions d'une telle étude dépendent évidemment de la façon dont est constitué l'échantillon. Par exemple, une étude statistique sur des habitudes de consommation donnera des résultats différents selon l'âge et le milieu social des personnes sondées. La méthode d'échantillonnage qui, à l'usage, s'est révélée offrir le maximum de garantie d'objectivité et de représentativité des résultats est l'*échantillonnage aléatoire simple*. Cette méthode consiste à choisir *au hasard* des éléments dans une population, de telle sorte que chaque individu ait autant de chance d'être sélectionné [1].

2 Rappels sur les variables aléatoires

Expérience aléatoire

On appelle *expérience aléatoire* une expérience qui, répétée plusieurs fois dans des conditions opératoires identiques, produit des résultats qui peuvent être différents. Mathématiquement, la notion d'expérience aléatoire $\mathcal{E}$ se formalise en définissant :

1. un ensemble fondamental Ω définissant l'ensemble des résultats possibles de $\mathcal{E}$, appelés *événements élémentaires*

2. un ensemble $\mathcal{A}$ de parties de Ω, appelées *événements*. Un événement aléatoire correspond à une affirmation qui peut être vraie ou fausse suivant le résultat de l'expérience aléatoire.

3. une fonction $\mathbb{P} : \mathcal{A} \to [0, 1]$, appelée *mesure* ou *distribution* de probabilité, qui à tout événement A associe un nombre $\mathbb{P}(A)$ appelé probabilité de cet événement.

1. Cette définition ne s'applique en toute rigueur qu'à une population finie ; nous admettrons qu'elle peut être étendue au cas d'une population infinie, ou même hypothétique.

Variable aléatoire

Une variable aléatoire (v.a.) est une grandeur numérique dont la valeur est fonction du résultat d'une expérience aléatoire. Mathématiquement, cette notion se formalise par une fonction

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longmapsto & X(\omega). \end{array}$$

On notera $V_X = X(\Omega)$ l'ensemble des valeurs prises par la v.a. X. On parle de v.a. *discrète* lorsque V_X est fini ou dénombrable. Dans le cas contraire, la v.a. X est dite *continue*.

Loi de probabilité d'une v.a.

Soit B un intervalle de $\mathbb{R}$. On peut définir la probabilité que la v.a. X prenne sa valeur dans B comme

$$\Pr_X(B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)),$$

quantité notée simplement $\mathbb{P}(X \in B)$. La donnée de $\Pr_X(B)$ pour tout intervalle B définit la *loi (ou distribution) de probabilité* de X.

Mathématiquement, c'est une fonction de $\mathcal{B}(\mathbb{R})$ dans $[0,1]$, $\mathcal{B}(\mathbb{R})$ étant l'ensemble des intervalles ou unions dénombrables d'intervalles de $\mathbb{R}$, appelé *tribu borélienne*. La fonction $\Pr_X$ est une mesure de probabilité sur l'espace probabilisable $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, appelée mesure image de Pr par X. Pour décrire complètement $\Pr_X$, il suffit de donner les probabilités pour des intervalles de la forme $]-\infty, x]$ pour tout $x \in \mathbb{R}$. On appelle *fonction de répartition* de X la fonction

$$\begin{array}{rccc} F_X: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & \Pr_X(]-\infty, x]), \end{array}$$

ce que l'on note $F_X(x) = \mathbb{P}(X \leq x)$.

Une loi de probabilité peut également être définie :

— dans le cas discret par la *fonction de probabilité* p_X qui à chaque élément de V_X associe sa probabilité :

$$\begin{array}{rccc} p_X: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & \Pr_X(\{x\}) \end{array}$$

et qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}),\ \Pr_X(B) = \sum_{x \in B} p_X(x)$$

— et dans le cas continu, par la *fonction de densité de probabilité* qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}), \quad \Pr_X(B) = \int_B f_X(t)dt.$$

Espérance mathématique

L'espérance mathématique d'une variable aléatoire réelle, qui représente la « valeur moyenne » prise par cette variable aléatoire, est définie par

$$\mathbb{E}(X) = \begin{cases} \displaystyle\sum_{x \in V_X} x p_X(x) & \text{si } X \text{ est une v.a. discrète,} \\ \displaystyle\int_{\mathbb{R}} x f_X(x)dx & \text{si } X \text{ est une v.a. continue} \end{cases}$$

si ces quantités existent. Dans le contraire, X n'a pas d'espérance mathématique.

Variance

La variance, qui est une mesure de dispersion de la v.a. autour de son espérance, est définie par

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

La racine carrée de la variance est appelée *écart-type* de la v.a. X et notée σ. La variance, étant une espérance, peut ne pas être définie.

Covariance

La covariance entre entre deux variables aléatoires X et Y est définie par

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

et que cette covariance vérifie les propriétés suivantes

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y)$.
- Inégalité de Cauchy-Schwarz : $[\mathrm{Cov}(X, Y)]^2 \leq \mathrm{Var}(X)\,\mathrm{Var}(Y)$ (égalité ssi $X - \mathbb{E}(X) = k(Y - \mathbb{E}(Y))$.

Corrélation

Le coefficient de corrélation $\rho_{jj'}$ entre deux variables aléatoires X_j et $X_{j'}$ est défini par

$$\rho_{jj'} = \frac{\mathrm{Cov}(X_i, X_{j'})}{\sqrt{\mathrm{Var}(X_j)\,\mathrm{Var}(X_{j'})}} = \frac{\mathrm{Cov}(X_j, X_{j'})}{\sigma_i \sigma_{j'}}.$$

3 Vecteurs aléatoires

3.1 Définition

La notion de vecteur aléatoire (réel), ou variable aléatoire vectorielle, généralise celle de variable aléatoire présentée au paragraphe précédent. On appelle vecteur aléatoire (réel) un vecteur de $\mathbb{R}^p$ dont les composants sont fonctions du résultat d'une expérience aléatoire $\mathcal{E}$. Il s'agit donc d'une application :

$$\begin{array}{rccl} \boldsymbol{X}: & \Omega & \longrightarrow & \mathbb{R}^p \\ & \omega & \longmapsto & \boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_p(\omega))^T. \end{array}$$

Un *vecteur aléatoire* (réel) est donc un vecteur $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ dont les composantes X_j sont des variables aléatoires réelles.

3.2 Loi jointe

La loi de probabilité du vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, appelée loi jointe, est définie par :

$$\mathbb{P}(\boldsymbol{X} \in A) = \Pr_{\boldsymbol{X}}(A) = \mathbb{P}(\omega \in \Omega \mid (X_1(\omega), \ldots, X_p(\omega)) \in A).$$

Cette loi de probabilité du vecteur aléatoire $\boldsymbol{X}$ peut être décrite par sa *fonction de répartition* définie pour tout $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ de $\mathbb{R}^p$ par

$$F_{\boldsymbol{X}}(x_1, \ldots, x_p) = \mathbb{P}(X_1 \leq x_1, \ldots, X_p \leq x_p).$$

Lorsque cette fonction est dérivable par rapport à chaque variable, on peut définir la *densité de probabilité* de $\boldsymbol{X}$ par

$$f_{\boldsymbol{X}}(x_1, \ldots, x_p) = \frac{\partial^p F_{\boldsymbol{X}}(x_1, \ldots, x_p)}{\partial x_1 \partial x_2 \cdots \partial x_p}.$$

Par la suite, nous noterons indifféremment $F_{\boldsymbol{X}}(x_1, \ldots, x_p)$ ou $F_{\boldsymbol{X}}(\boldsymbol{x})$, $f_{\boldsymbol{X}}(x_1, \ldots, x_p)$ ou $f_{\boldsymbol{X}}(\boldsymbol{x})$, etc.

Si $f_{\boldsymbol{X}}$ existe, on a :

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x},$$

pour tout $A \subseteq \mathbb{R}^p$ pour lequel cette intégrale est définie.

Lorsque le vecteur aléatoire $\boldsymbol{X}$ est discret, c'est-à-dire lorsque les composantes X_j sont des variables aléatoires discrètes, on peut définir la fonction de probabilité de $\boldsymbol{X}$ (équivalent de la fonction de densité) par

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(X_1 = x_1, \ldots, X_p = x_p).$$

On a alors $\mathbb{P}(\boldsymbol{X} \in A) = \sum_{\boldsymbol{x} \in A} p_{\boldsymbol{X}}(\boldsymbol{x})$ pour tout $A \subseteq \mathbb{R}^p$)

Dans la suite et s'il n'y a pas d'ambiguïté, on notera simplement p, f et F les fonctions $p_{\boldsymbol{X}}$, $f_{\boldsymbol{X}}$ et $F_{\boldsymbol{X}}$.

3.3 Lois marginales

Tout sous-vecteur du vecteur aléatoire $\boldsymbol{X}$, c'est-à-dire tout sous-ensemble de l'ensemble des variables aléatoires $X_1, \ldots, X_p$ est lui-même un vecteur aléatoire. La loi d'un tel vecteur aléatoire est appelée loi marginale. Si $X_{j_1}, \ldots, X_{j_q}$ est ce sous-ensemble, la loi marginale sera notée $f_{j_1, \ldots, j_q}$.

Si cet ensemble se réduit à une seule variable et

— si $\boldsymbol{X}$ est discret, la loi de X_j est définie par la probabilité élémentaire

$$p_j(x_j) = \sum_{x_1 \in V_1, \ldots, x_{j-1} \in V_{j-1}, x_{j+1} \in V_{j+1}, \ldots, x_p \in V_p} p(\boldsymbol{x})$$

— et si $\boldsymbol{X}$ est continu, la loi de X_j est définie par la densité

$$f_j(x_j) = \int_{\mathbb{R}^{p-1}} f(\boldsymbol{x}) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_p.$$

3.4 Espérance

L'espérance du vecteur aléatoire $\boldsymbol{X}$ est le vecteur des espérances des variables aléatoires X_j :

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{X}) = (\mathbb{E}(X_1), \ldots, E(X_p))^T.$$

$\boldsymbol{X}$ et $\boldsymbol{Y}$ étant des vecteurs aléatoires de dimension p, on a les propriétés suivantes.

Proposition 1.

$$\mathbb{E}(\boldsymbol{X} + \boldsymbol{Y}) = \mathbb{E}(\boldsymbol{X}) + \mathbb{E}(\boldsymbol{Y})$$

Proposition 2. *Pour toute matrice $A \in \mathcal{M}_{q,p}(\mathbb{R})$ et tout vecteur $\mathbf{b} \in \mathbb{R}^q$ constants, on a*

$$\mathbb{E}(A\boldsymbol{X} + \mathbf{b}) = A\,\mathbb{E}(\boldsymbol{X}) + \mathbf{b}.$$

En particulier, si $\mathbf{u} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{u}^T \boldsymbol{X}) = \mathbf{u}^T \mathbb{E}(\boldsymbol{X})$.

Espérance d'une fonction réelle d'un vecteur aléatoire

Si φ est une fonction de $\mathbb{R}^p$ dans $\mathbb{R}$, on a

$$\mathbb{E}(\varphi(\boldsymbol{X})) = \int_{\mathbb{R}^p} \varphi(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

pour un vecteur aléatoire continu et

$$\mathbb{E}(\varphi(\boldsymbol{X})) = \sum_{\boldsymbol{x} \in V_1 \times \ldots \times V_p} \varphi(\boldsymbol{x}) p(\boldsymbol{x})$$

pour un vecteur aléatoire discret.

Comme pour les variables aléatoires, ce résultat permet de calculer l'espérance d'un vecteur aléatoire $\varphi(\mathbf{X})$ sans avoir besoin de calculer sa loi.

3.5 Matrice de Variance

La variance du vecteur aléatoire $\boldsymbol{X}$, souvent appelée *matrice de variance*, est la matrice carrée symétrique Σ de dimension p de terme général

$$\begin{aligned} \sigma_{jj'} &= \mathrm{Cov}(X_j, X_{j'}) \\ &= \mathbb{E}[(X_j - \mathbb{E}(X_{j'}))(X_j - \mathbb{E}(X_{j'}))] \\ &= \mathbb{E}(X_j X_{j'}) - \mathbb{E}(X_j)\mathbb{E}(X_{j'}). \end{aligned}$$

En particulier, $\sigma_{jj} = \mathrm{Var}(X_j)$.

On peut écrire matriciellement

$$\Sigma = \mathrm{Var}(\boldsymbol{X}) = \mathbb{E}([\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})][\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})]^T).$$

Proposition 3. *Pour toute matrice $A \in \mathcal{M}_{q,p}(\mathbb{R})$ et tout vecteur $\mathbf{b} \in \mathbb{R}^q$ constants, on a*

$$\mathrm{Var}(A\boldsymbol{X} + \mathbf{b}) = A \,\mathrm{Var}(\boldsymbol{X}) A^T = A\Sigma A^T.$$

En particulier, on a donc pour tout $\mathbf{u}$ et $\mathbf{v} \in \mathbb{R}^p$

$$\mathrm{Var}(\mathbf{u}^T\boldsymbol{X}) = \mathbf{u}^T\Sigma\mathbf{u} \quad \text{et} \quad \mathrm{Cov}(\mathbf{u}^T\boldsymbol{X}, \mathbf{v}^T\boldsymbol{X}) = \mathbf{u}^T\Sigma\mathbf{v},$$

ce qui montre que la matrice Σ est définie positive ($\mathbf{u}^T\Sigma\mathbf{u} > 0, \forall \mathbf{u} \neq 0$), sauf s'il existe une relation $\mathbf{u}^T\boldsymbol{X} = c$ pour un vecteur $\mathbf{u}$ et un scalaire c constants, auquel cas $\mathrm{Var}(\mathbf{u}^T\boldsymbol{X}) = 0$.

On rappelle que toute matrice d'ordre p symétrique et définie positive a p valeurs propres strictement positives. La matrice Σ est donc inversible.

Matrice de corrélation

La matrice de corrélation R d'un vecteur aléatoire est la matrice de terme général $\rho_{jj'}$. Toutes les valeurs sont donc comprises entre -1 et $+1$ et les termes de la diagonale sont égaux à 1. Si on note D la matrice diagonale $\mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, on obtient les relations $\Sigma = DRD$ et $R = D^{-1}\Sigma D^{-1}$.

3.6 Indépendance de variables aléatoires

Les composantes $X_1, \ldots, X_p$ du vecteur aléatoire $\boldsymbol{X}$ sont *indépendantes* si la loi jointe du vecteur aléatoire $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ s'exprime comme le produit des lois marginales, c'est-à-dire si et seulement si :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{j=1}^{p} f_{X_j}(x_j).$$

Etant donnée une variable aléatoire Z à valeurs dans Ω, les composantes $X_1, \ldots, X_p$ du vecteur aléatoires $\boldsymbol{X}$ sont *indépendantes conditionnellement à Z* ssi

$$f_{\boldsymbol{X}}(\boldsymbol{x}|Z=z) = \prod_{j=1}^{p} f_{X_j}(x_j|Z=z), \quad \forall \boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p, \forall z \in \Omega.$$

Propriétés

1. Variables $X_1, \ldots, X_p$ indépendantes $\Longrightarrow$ tout sous-ensemble des v. a. est indépendant ; en particulier, les v. a. $X_1, \ldots, X_p$ sont indépendantes 2 à 2 (attention, la réciproque est fausse : l'indépendance 2 à 2 n'entraîne pas l'indépendance) ;
2. Variables $X_1, \ldots, X_p$ indépendantes $\Longrightarrow \mathbb{E}(X_1 \ldots X_p) = \mathbb{E}(X_1) \ldots \mathbb{E}(X_p)$;
3. Variables X_j et $X_{j'}$ indépendantes $\Longrightarrow \mathrm{Cov}(X_j, X_{j'}) = 0$ (la matrice de variance sera donc diagonale si les variables X_j sont indépendantes 2 à 2 ; la réciproque est fausse) ;
4. Variables $X_1, \ldots, X_p$ indépendantes $\Longrightarrow \mathrm{Var}(\sum_{j=1}^{p} X_j) = \sum_{j=1}^{p} \mathrm{Var}(X_j)$.

3.7 Transformation d'un vecteur aléatoire

Soit $\mathbf{U}$ un vecteur aléatoire de dimension p, φ une application bijective de $\mathbb{R}^p$ dans $\mathbb{R}^p$, et $\boldsymbol{X} = \varphi(\mathbf{U})$. La densité de $\boldsymbol{X}$ s'obtient en fonction de celle de $\mathbf{U}$ par l'expression :

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{f_{\mathbf{U}}(\varphi^{-1}(\boldsymbol{x}))}{|\det J_\varphi|},$$

où $\det J_\varphi$ est le jacobien de la transformation défini par

$$\det J_\varphi = \begin{vmatrix} \frac{\partial \varphi_1}{\partial u_1} & \cdots & \frac{\partial \varphi_1}{\partial u_p} \\ \vdots & & \vdots \\ \frac{\partial \varphi_p}{\partial u_1} & \cdots & \frac{\partial \varphi_p}{\partial u_p} \end{vmatrix},$$

où $\varphi = (\varphi_1, \ldots, \varphi_p)$.

4 Statistiques

Soit $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ un échantillon indépendant et identiquement distribué (iid) de vecteur aléatoire parent $\boldsymbol{X}$. On peut alors définir le *vecteur moyenne empirique*

$$\overline{\boldsymbol{X}} = \frac{1}{n} \sum_i \boldsymbol{X}_i = (\overline{X}_1, \ldots, \overline{X}_p)^T,$$

où $\overline{X}_j$ est la moyenne empirique de l'échantillon $X_{1j}, \ldots, X_{nj}$ et la *matrice de variance empirique*

$$V = \frac{1}{n} \sum_i (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T.$$

La moyenne empirique est un estimateur sans biais de l'espérance $\mathbb{E}(\boldsymbol{X})$. En revanche, la matrice de variance empirique n'est pas un estimateur sans biais de la matrice de variance car on peut montrer que $\mathbb{E}(V) = \frac{n-1}{n} \mathrm{Var}(\boldsymbol{X})$. Pour obtenir un estimateur sans biais de la matrice de variance, on définit alors la *matrice de variance empirique corrigée*

$$V^* = \frac{1}{n-1} \sum_i (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T.$$

Par abus de notation, on notera de manière identique les statistiques V et V^* et leurs réalisations

$$\frac{1}{n}\sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \quad \text{et} \quad \frac{1}{n-1}\sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T.$$

5 Loi normale multidimensionnelle

5.1 Définition

Soient $U_1, \ldots, U_p$ p v.a. réelles normales, centrées-réduites et indépendantes, et $\mathbf{U} = (U_1, \ldots, U_p)^T$. On appelle *loi normale à p dimensions* la loi suivie par $\boldsymbol{X} = \boldsymbol{\mu} + B\mathbf{U}$, où $\boldsymbol{\mu} \in \mathbb{R}^p$ et $B \in \mathcal{M}_{p,p}(\mathbb{R})$ sont des constantes.

La densité de $\mathbf{U}$ est

$$f(\mathbf{u}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{u}^T\mathbf{u}\right).$$

On peut calculer la densité de $\boldsymbol{X}$ par la méthode rappelée dans la section 3.7. On vérifie que $\det J_\varphi = \det B$. On en déduit :

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\det B|} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T(BB^T)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Or, $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ et $\Sigma = \mathrm{Var}(\boldsymbol{X}) = BB^T$, d'où $\det B = (\det \Sigma)^{1/2}$. On obtient finalement l'expression usuelle de la densité de $\boldsymbol{X}$:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

On note $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

5.2 Propriétés

1. Dans le cas $p = 1$, on retrouve l'expression de la loi normale monodimensionnelle, avec $\sigma^2 = \Sigma$.
2. La matrice Σ est diagonale ssi les variables $X_1, \ldots, X_p$ sont indépendantes.
3. Tout sous-vecteur d'un vecteur aléatoire gaussien suit une loi normale. En particulier, ses composantes sont toutes gaussiennes.
4. Les courbes d'isodensité ont pour équation $(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c$, où c est une constante. Ce sont des ellipsoïdes de centre $\boldsymbol{\mu}$. Lorsque la matrice Σ est diagonale, les axes de ces ellipsoïdes sont parallèles aux axes de coordonnées. Lorsque Σ est la matrice identité, ce sont des hypersphères.
5. Soient $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\mathbf{b} \in \mathbb{R}^q$ un vecteur constant, et $A \in \mathcal{M}_{q,p}(\mathbb{R})$ une matrice constante. Alors, $\boldsymbol{Y} = A\,\boldsymbol{X} + \mathbf{b} \sim \mathcal{N}(A\,\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

5.3 Estimation des paramètres

Si $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ est un échantillon indépendant et identiquement distribué (iid) de vecteur aléatoire parent $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, alors les estimateurs du maximum de vraisemblance $\widehat{\boldsymbol{\mu}}$ et $\widehat{\Sigma}$ de $\boldsymbol{\mu}$ et de Σ sont le vecteur moyenne empirique $\overline{\boldsymbol{X}}$ et la matrice de variance empirique V et on a $\widehat{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\Sigma)$.

Annexe B

Rappels et compléments d'algèbre linéaire et de géométrie

Dans cette annexe, les notions élémentaires d'algèbre linéaire (espace vectoriel sur un corps K, sous-espace vectoriel, combinaison linéaire de vecteurs, famille libre, famille liée, base, dimension…) seront supposées connues. Dans la suite tous les espaces vectoriels envisagés seront toujours définis sur le corps des nombres réels.

1 Espace vectoriel

L'espace $\mathbb{R}^p$

Le produit cartésien $\mathbb{R}^p$ est l'ensemble de tous les p-uplets de nombres réels. C'est un espace vectoriel très souvent utilisé, en particulier en analyse des données. La dimension de cet espace est p et on peut montrer que tous les espaces vectoriels sur $\mathbb{R}$ de dimension p sont isomorphes à $\mathbb{R}^p$. La base canonique est la famille formée des éléments suivants

$$\mathbf{e}_1 = (1, 0, \ldots, 0), \quad \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \quad \mathbf{e}_p = (0, \ldots, 0, 1)$$

Tout élément $\boldsymbol{x} = (x_1, \ldots, x_p)$ se décompose dans cette base canonique $\boldsymbol{x} = \sum_{i=1}^{p} x_i \mathbf{e}_i$

Les coordonnées par rapport à une base préalablement fixée s'écrivent traditionnellement sous forme de vecteur colonne.

Les éléments de $\mathbb{R}^p$ sont notés dans la suite sous la forme de « vecteurs colonnes » ou matrice de dimension $(p, 1)$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix},$$

et la base canonique est formée des vecteurs

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \mathbf{e}_p = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

La décomposition de x s'écrit donc $\boldsymbol{x} = \sum_{i=1}^{p} x_i \mathbf{e}_i$. Il y a donc identité entre les coordonnées de $\boldsymbol{x}$ dans la base canonique et les composantes du vecteur $\boldsymbol{x}$, élément du produit cartésien $\mathbb{R}^p$. Ceci n'est vrai que pour la base canonique.

Décomposition en somme directe

Définition 1. *Un espace vectoriel E est somme directe des sous-espaces vectoriels $E_1, \ldots, E_k$ si et seulement si tout élément $\boldsymbol{x}$ de E s'écrit de manière unique $\boldsymbol{x} = \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_k$ avec $\boldsymbol{x}_i \in E_i$.*

On note $E = E_1 \oplus \cdots \oplus E_k$. Lorsque le nombre de sous-espaces de la somme directe se réduit à deux, on parle de sous-espaces supplémentaires.

2 Applications linéaires et matrices

Application linéaire

Définition 2. *On appelle application linéaire d'un espace vectoriel E dans un espace vectoriel F, une application f de E dans F vérifiant les propriétés suivantes :*

$$\begin{array}{ll} \forall (\boldsymbol{x}, \boldsymbol{y}) \in E^2 & f(\boldsymbol{x} + \boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y}), \\ \forall \boldsymbol{x} \in E, \forall a \in \mathbb{R} & f(a\boldsymbol{x}) = af(\boldsymbol{x}). \end{array}$$

Cas particuliers : un endomorphisme est une application linéaire de E dans E et une forme linéaire est une application linéaire de E dans $\mathbb{R}$.

Matrice associée à une application linéaire

Si E et F sont de dimensions finies p et n, et si $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ et $(\mathbf{f}_1, \ldots, \mathbf{f}_n)$ sont des bases de E et F, il est possible d'associer à une application linéaire f de E dans F une matrice A de dimension (n, p). Celle-ci est construite en rangeant en colonne les coordonnées des images des vecteurs de la base de E sur la base de F : si la matrice A est notée (a_{ij}), a_{ij} est la i^{e} coordonnée de $f(\mathbf{e}_j)$.

La relation $\boldsymbol{y} = f(\boldsymbol{x})$ s'écrit alors matriciellement $\boldsymbol{y} = A\boldsymbol{x}$. Ici, pour simplifier l'écriture, les éléments de E et F et leurs vecteurs de coordonnées associées dans les bases correspondantes sont notés de la même façon. Réciproquement, toute application de E dans F se mettant sous la forme $\boldsymbol{y} = A\boldsymbol{x}$ est une application linéaire.

Opérations sur les matrices

Les opérations matricielles de base sont le produit d'une matrice par un réel, la somme de deux matrices, le produit de deux matrices et la transposition d'une matrice.

Les matrices associées aux endomorphismes sont carrées et il est alors possible de définir sur de telles matrices les notions de matrice diagonale, de matrice symétrique, de matrice identité, de déterminant, de trace et de matrice inverse.

3 Changement de base

Matrice de changement de base

Si $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ et $(\mathbf{f}_1, \ldots, \mathbf{f}_p)$ sont deux bases d'un espace vectoriel E de dimension p et si $\boldsymbol{x}_\mathbf{e}$ et $\boldsymbol{x}_\mathbf{f}$ sont les vecteurs des coordonnées d'un élément de E dans ces deux bases, on a la relation :

$$\boldsymbol{x}_\mathbf{f} = P^{-1}\boldsymbol{x}_\mathbf{e} \quad \text{et} \quad \boldsymbol{x}_\mathbf{e} = P\boldsymbol{x}_\mathbf{f},$$

où P est une matrice carrée de dimension p, appelée matrice de *changement de base* ou *matrice de passage*. Pour obtenir cette matrice de changement de base, il suffit de ranger

en colonne les coordonnées des nouveaux vecteurs de base $(\mathbf{f}_1, \ldots, \mathbf{f}_p)$ dans l'ancienne base $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$. On notera que contrairement à l'intuition, les nouvelles coordonnées $\boldsymbol{x}_\mathbf{f}$ s'obtiennent en multipliant P^{-1} (et non P) avec $\boldsymbol{x}_\mathbf{e}$.

Effet sur la matrice associée à un endomorphisme

Si f est un endomorphisme sur E, P la matrice de changement de base, A la matrice associée à f dans la base $(\mathbf{e}_j)$, B la matrice associée à f dans la base $(\mathbf{f}_j)$, alors on a la relation $B = P^{-1}AP$.

Preuve : Soit $\mathbf{a}$ un élément de E. Notons $\boldsymbol{x}_\mathbf{e}$ et $\boldsymbol{y}_\mathbf{e}$ les coordonnées de $\mathbf{a}$ et de $f(\mathbf{a})$ dans la première base et $\boldsymbol{x}_\mathbf{f}$ et $\boldsymbol{y}_\mathbf{f}$ les coordonnées des mêmes éléments dans la seconde base, nous avons :

$$\boldsymbol{x}_\mathbf{e} = P\boldsymbol{x}_\mathbf{f}, \quad \boldsymbol{y}_\mathbf{e} = P\boldsymbol{y}_\mathbf{f}, \quad \boldsymbol{y}_\mathbf{e} = A\boldsymbol{x}_\mathbf{e}, \quad \boldsymbol{y}_\mathbf{f} = A\boldsymbol{x}_\mathbf{f},$$

et donc

$$\begin{aligned} \boldsymbol{y}_\mathbf{f} &= P^{-1}\boldsymbol{y}_\mathbf{e} \\ &= P^{-1}A\boldsymbol{x}_\mathbf{e} \\ &= P^{-1}AP\boldsymbol{x}_\mathbf{f} \end{aligned}$$

On en déduit donc $B = P^{-1}AP$.

□

4 Vecteurs et valeurs propres d'un endomorphisme

Définition et propriétés

Définition 3. *On appelle vecteur propre d'un endomorphisme f sur E tout élément $\boldsymbol{x}$ de E non nul tel qu'il existe un réel λ vérifiant $f(\boldsymbol{x}) = \lambda\boldsymbol{x}$. Ce réel λ est appelé valeur propre associée au vecteur propre $\boldsymbol{x}$.*

Proposition 4. *Si $\boldsymbol{x}$ est un vecteur propre, les vecteurs $a\boldsymbol{x}$ où a est un réel non nul sont aussi des vecteurs propres et ont même valeur propre.*

Proposition 5. *L'ensemble de tous les vecteurs propres associés à une même valeur propre auquel est ajouté le vecteur nul est un espace vectoriel. Il est appelé espace propre associé à la valeur propre λ et noté E_λ.*

Recherche des valeurs propres et vecteurs propres

On suppose dans ce paragraphe que E est de dimension finie. Soient $\mathbf{e}_i$ une base de E et A la matrice carrée associée à un endomorphisme f dans cette base, on a alors :

$$\boldsymbol{x} \text{ vecteur propre de } f \iff \boldsymbol{x} \neq 0 \text{ et } A\boldsymbol{x} = \lambda\boldsymbol{x} \iff \boldsymbol{x} \neq 0 \text{ et } (A - \lambda I)\boldsymbol{x} = 0$$

Le système de p équations à p inconnues ainsi défini ne doit donc pas être un système de Cramer, sinon la solution unique serait 0. Les solutions λ doivent donc annuler le déterminant de la matrice $(A - \lambda I)$. Il suffit ensuite pour chaque valeur λ réalisant cette condition de trouver les vecteurs $\boldsymbol{x}$ vérifiant le système $A\boldsymbol{x} = \lambda\boldsymbol{x}$.

Application : diagonalisation d'une matrice

Le problème

Si E est un espace vectoriel de dimension finie muni d'une base $(\mathbf{e}_j)$ et f un endomorphisme sur E dont la matrice associée dans cette base est A , on cherche une nouvelle base $(\mathbf{f}_j)$ telle que la matrice associée à f soit diagonale.

Résolution

Il est facile de montrer que les vecteurs de la nouvelle base sont nécessairement des vecteurs propres de f et que les termes de la diagonale sont les valeurs propres associées. Réciproquement, si on a une base formée de vecteurs propres de f, la matrice associée à f est diagonale et les valeurs propres sont les termes de la diagonale : diagonaliser une matrice revient donc à trouver une base de vecteurs propres.

Remarque

Toute matrice carrée n'est pas diagonalisable, mais on peut montrer que toutes les matrices symétriques le sont (voir paragraphe 6).

5 Produit scalaire, norme, distance et orthogonalité

Définition 6. *On appelle produit scalaire sur un espace vectoriel E une application de $E \times E$ dans $\mathbb{R}$:*

(i) bilinéaire

(ii) symétrique : $\forall \boldsymbol{x}, \boldsymbol{y} \in E \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$,

(iii) définie : $\forall \boldsymbol{x}, \boldsymbol{y} \in E \quad \langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0 \Rightarrow \boldsymbol{x} = 0$,

(iv) positive : $\forall \boldsymbol{x} \in E \quad \langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$.

Expression matricielle On se place dans l'espace $\mathbb{R}^p$ muni de sa base canonique. On montre facilement que tout produit scalaire $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ s'écrit sous la forme $\boldsymbol{x}^T M \boldsymbol{y}$ où M est une matrice :

(i) symétrique : $M^T = M$,

(ii) définie : $\forall \boldsymbol{x} \in \mathbb{R}^p \quad \boldsymbol{x}^T M \boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0$,

(iii) positive : $\forall \boldsymbol{x} \in \mathbb{R}^p \quad \boldsymbol{x}^T M \boldsymbol{x} \geq 0$.

On note souvent $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_M$ ce produit scalaire. Le produit scalaire habituel correspond à la matrice identité.

Définition 7. *On appelle norme sur un espace vectoriel E une application de E dans $\mathbb{R}^+$ vérifiant :*

$$\begin{array}{ll} \forall \boldsymbol{x} \in E, \forall \lambda \in \mathbb{R} & \|\lambda \boldsymbol{x}\| = |\lambda| \|\boldsymbol{x}\|, \\ \forall \boldsymbol{x} \in E & \|\boldsymbol{x}\| = 0 \Rightarrow \boldsymbol{x} = 0, \\ \forall \boldsymbol{x}, \boldsymbol{y} \in E & \|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|. \end{array}$$

Norme euclidienne Lorsque E est muni d'un produit scalaire, on montre que l'application qui associe à un élément de E la racine carrée du produit scalaire de cet élément avec lui-même est une norme sur A. Elle est appelée norme euclidienne et notée $\|\boldsymbol{x}\|_M = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_M}$.

Vecteur normé Un vecteur est normé si sa norme est égale à 1.

Définition 8. *On appelle distance sur un ensemble quelconque A une application d de $A \times A$ dans $\mathbb{R}^+$ vérifiant :*

$$\begin{array}{ll} \forall \boldsymbol{x}, \boldsymbol{y} \in A & d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \\ \forall \boldsymbol{x}, \boldsymbol{y} \in A & d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}, \\ \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in A & d(\boldsymbol{x}, \boldsymbol{y}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}). \end{array}$$

Distance associée à une norme Lorsque A est un espace vectoriel muni d'un produit scalaire, on peut montrer que l'application d définie par $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ est une distance sur A.

Distance euclidienne Si la distance est associée à une norme euclidienne, la distance est euclidienne et on a dans ce cas :

$$d_M(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_M = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle_M} = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T M (\boldsymbol{x} - \boldsymbol{y})}$$

Orthogonalité

Dans tout ce paragraphe, l'espace vectoriel E est muni d'un produit scalaire et donc d'une norme et d'une distance.

Vecteurs orthogonaux Deux éléments $\boldsymbol{x}$ et $\boldsymbol{y}$ de E sont orthogonaux si leur produit scalaire est nul :

$$\boldsymbol{x} \perp \boldsymbol{y} \Longleftrightarrow \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0.$$

Sous-espaces vectoriels orthogonaux Deux sous-espaces vectoriels F et G sont orthogonaux si tous les éléments de l'un sont orthogonaux à tous les éléments de l'autre :

$$F \perp G \Longleftrightarrow (\forall \boldsymbol{x} \in F, \forall \boldsymbol{y} \in G, \boldsymbol{x} \perp \boldsymbol{y})$$

Sous-espace orthogonal supplémentaire Le sous-espace orthogonal supplémentaire $F^\perp$ d'un sous-espace vectoriel F est l'ensemble des éléments de E orthogonaux à tous les éléments de F :

$$F^\perp = \{\boldsymbol{x} \in E : \forall \boldsymbol{y} \in F, \boldsymbol{x} \perp \boldsymbol{y}\}$$

On peut montrer que les deux sous-espaces F et $F^\perp$ sont orthogonaux et supplémentaires.

Décomposition en somme directe d'espaces orthogonaux Une décomposition en somme directe de sous-espaces orthogonaux est une décomposition en somme directe de sous-espaces orthogonaux deux à deux.

Théorème 9 (Théorème de Pythagore)**.** *Si un élément $\boldsymbol{x}$ de E se décompose suivant deux sous-espaces supplémentaires orthogonaux en $\boldsymbol{y}$ et $\boldsymbol{z}$, on a*

$$\|\boldsymbol{x}\|^2 = \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2$$

Preuve. Soit $E = F \oplus G$. On a $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ avec $\boldsymbol{x} \in F$ et $\boldsymbol{y} \in G$

$$\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \langle \boldsymbol{y} + \boldsymbol{z}, \boldsymbol{y} + \boldsymbol{z} \rangle = \langle \boldsymbol{y}, \boldsymbol{y} \rangle + 2\langle \boldsymbol{y}, \boldsymbol{z} \rangle + \langle \boldsymbol{z}, \boldsymbol{z} \rangle = \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2.$$

□

Généralisation du théorème de Pythagore Si $\boldsymbol{x} = \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_r$ est la décomposition d'un élément suivant une somme directe de sous-espaces orthogonaux, on a

$$\|\boldsymbol{x}\|^2 = \sum_{i=1}^{r} \|\boldsymbol{x}_i\|^2.$$

Base orthonormée Une base est orthonormée si les vecteurs de la base sont orthogonaux deux à deux et s'ils sont normés. Par exemple, on peut facilement montrer que la base canonique est orthonormée pour le produit scalaire usuel. Si $x_1, \ldots, x_p$ sont les coordonnées d'un vecteur $\boldsymbol{x}$ dans une base orthonormée, on montre facilement, en utilisant le théorème de Pythagore, la relation :

$$\|\boldsymbol{x}\|^2 = \sum_{j=1}^{p} x_j^2.$$

La matrice de passage entre deux bases orthonormées est orthogonale, c'est-à-dire vérifie la relation $P^T = P^{-1}$ (sa transposée est aussi son inverse).

Définition

Soit $E = F \oplus G$ une décomposition de E en deux sous-espaces supplémentaires, la décomposition unique $\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{z}$ avec $\boldsymbol{y} \in F$ et $\boldsymbol{z} \in G$ permet alors de définir deux applications :

— la première, qui associe au vecteur $\boldsymbol{x}$ de E le vecteur $\boldsymbol{y}$ de F, est appelée projection sur F parallèlement à G ;

— la seconde, qui associe au vecteur $\boldsymbol{x}$ de E le vecteur $\boldsymbol{z}$ de G, est appelée projection sur G parallèlement à F.

On peut montrer qu'une projection est une application linéaire et qu'elle est idempotente ($p \circ p = p$). Réciproquement, toute application linéaire idempotente est une projection.

Projection orthogonale sur un sous-espace vectoriel

Définition 10. *On appelle projection orthogonale sur un sous-espace vectoriel F la projection sur F parallèlement à $F^\perp$.*

Proposition 11. *F étant un sous-espace vectoriel, $\boldsymbol{x}$ un point quelconque de E, $\boldsymbol{y}$ sa projection orthogonale sur F et $\boldsymbol{t}$ un point quelconque de F, alors on a*

$$(\boldsymbol{x} - \boldsymbol{t}) = (\boldsymbol{x} - \boldsymbol{y}) + (\boldsymbol{y} - \boldsymbol{t}).$$

Cette relation représente la décomposition de $(\boldsymbol{x} - \boldsymbol{t})$ sur F et $F^\perp$ (figure B.1) et le théorème de Pythagore peut donc s'appliquer :

$$\|\boldsymbol{x} - \boldsymbol{t}\|^2 = \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \|\boldsymbol{y} - \boldsymbol{t}\|^2$$

$$d^2(\boldsymbol{x}, \boldsymbol{t}) = d^2(\boldsymbol{x}, \boldsymbol{y}) + d^2(\boldsymbol{y}, \boldsymbol{t})$$

La quantité $d^2(\boldsymbol{y}, \boldsymbol{t})$ étant toujours positive, cette relation permet d'affirmer que $\boldsymbol{y}$ est l'élément de F le plus proche de $\boldsymbol{x}$. Finalement, les trois relations suivantes sont équivalentes :

$$\begin{array}{l} \boldsymbol{y} \text{ est la projection orthogonale de } \boldsymbol{x} \text{ sur } F \\ \forall \boldsymbol{t} \in F \quad (\boldsymbol{x} - \boldsymbol{y}) \perp \boldsymbol{t} \\ d(\boldsymbol{x}, \boldsymbol{y}) = \inf\{d(\boldsymbol{x}, \boldsymbol{t}) / \boldsymbol{t} \in F\} \end{array}$$

La quantité $d(\boldsymbol{x}, \boldsymbol{y})$ est souvent appelée "distance" de $\boldsymbol{x}$ au sous-espace F et notée $d(\boldsymbol{x}, F)$.

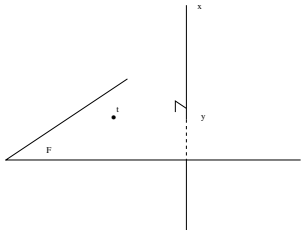


FIGURE B.1 – Projection sur un plan

Projection orthogonale sur une variété linéaire

Définition 12 (sous-espace affine)**.** *Si G est un sous-espace vectoriel de E et $\mathbf{a}$ un élément de E, on appelle variété linéaire l'ensemble F des éléments $\boldsymbol{x}$ de E tels que $\boldsymbol{x} = \mathbf{a} + \boldsymbol{y}$ avec $\boldsymbol{y}$ élément de G. On note $F = \mathbf{a} + G$. Le sous-espace vectoriel G est appelé direction de F. Si G est de dimension r, on dit que F est un espace affine de dimension r.*

Définition 13 (Variétés linéaires orthogonales)**.** *Deux variétés linéaires sont orthogonales si les sous-espaces vectoriels qui les définissent le sont.*

Projection orthogonale sur une variété linéaire

Soit $F = \mathbf{a} + G$ une variété linéaire, $\boldsymbol{x}$ un point quelconque de E et $H = \boldsymbol{x} + G^{\perp}$ (fig B.2. On peut montrer que l'intersection de H et F est réduite à un seul élément noté ici $\boldsymbol{y}$. Cet élément est appelé projection orthogonale de $\boldsymbol{x}$ sur F. On peut alors étendre les résultats obtenus précédemment :

1. $\boldsymbol{y}$ est la projection orthogonale de $\boldsymbol{x}$ sur F
2. $\forall \boldsymbol{t}, \mathbf{u} \in F \qquad (\boldsymbol{x} - \boldsymbol{y}) \perp (\mathbf{t} - \mathbf{u})$ ou encore $(\boldsymbol{x} - \boldsymbol{y})^T M(\mathbf{t} - \mathbf{u}) = 0$
3. $d(\boldsymbol{x}, \boldsymbol{y}) = \inf\{d(\boldsymbol{x}, \mathbf{t}) / \mathbf{t} \in F\}$

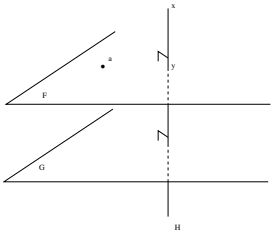


FIGURE B.2 – Projection sur un espace affine

La quantité $d(\boldsymbol{x}, \boldsymbol{y})$ est appelée « distance » de $\boldsymbol{x}$ au sous-espace F et notée $d(\boldsymbol{x}, F)$.

6 Matrices symétriques et matrices Q-symétriques

Proposition 14. *Toute matrice B symétrique possède une base orthonormée (au sens du produit scalaire usuel) de vecteurs propres et est donc diagonalisable. La matrice P de changement de base est orthogonale ($P^T = P^{-1}$ ou $P^T P = PP^T = I$). De plus, si B est positive alors toutes les valeurs propres sont positives ou nulles.*

Proposition 15. *Si Q est une matrice définissant un produit scalaire, toute matrice B Q-symétrique (c'est-à-dire QB symétrique) possède une base Q-orthonormée de vecteurs propres et est donc diagonalisable. La matrice P de changement de base vérifie $P^T QP = PP^T Q = QPP^T = I$ et $P^T QBP$ est la matrice diagonale des valeurs propres. De plus, si B est Q-positive (c'est-à-dire QB positive) alors toutes les valeurs propres sont positives ou nulles.*

Théorème 16 (décomposition d'une matrice)**.** *L'orthogonalité et la norme étant définies à l'aide d'une matrice Q, si B est une matrice Q-symétrique et Q-positive et $(\mathbf{u}_1, \ldots, \mathbf{u}_p)$ une base Q-orthonormée de vecteurs propres de la matrice B rangés suivant l'ordre décroissant des valeurs propres λ_k associés, alors :*

- *Le vecteur de norme 1 maximisant $\langle \mathbf{u}, B\mathbf{u} \rangle$ est le vecteur $\mathbf{u}_1$ et la valeur maximisée est λ_1.*
- *$\forall k, 1 < k \leq 1$, le vecteur de norme 1, orthogonal au sous-espace engendré par les vecteurs $\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}$ maximisant $\langle \mathbf{u}, B\mathbf{u} \rangle$ est le vecteur $\mathbf{u}_k$ et la valeur maximisée est λ_k.*

7 Espace euclidien

Un *espace euclidien* est un espace vectoriel réel de dimension finie muni d'un produit scalaire. Si on note $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ ce produit scalaire, la norme et la distance associées, appelées *norme euclidienne* et *distance euclidienne*, sont respectivement définies par

$$||\boldsymbol{x}|| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle} \qquad \text{et} \qquad d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}|| = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle}.$$

Rappelons que le produit scalaire permet aussi de définir les notions d'*othogonalité* et de *projection orthogonale* qui seront utilisées dans ce paragraphe.

8 Nuage de points et centre de gravité

Si Ω est un ensemble fini de points d'un espace euclidien $\mathcal{E}$ et si chaque point $\boldsymbol{x}$ de Ω est muni d'une pondération $\mu_{\boldsymbol{x}} > 0$, alors l'ensemble

$$\mathcal{N}(\Omega) = \{(\boldsymbol{x}, \mu_{\boldsymbol{x}}) / \boldsymbol{x} \in \Omega\}$$

est appelé *nuage de points* de $\mathcal{E}$ et son *centre de gravité* est défini par

$$\mathbf{g} = \frac{1}{\mu} \sum_{\boldsymbol{x} \in \Omega} \mu_{\boldsymbol{x}} \boldsymbol{x}$$

où μ est la somme des pondérations $\sum_{\boldsymbol{x}} \mu_{\boldsymbol{x}}$.

9 Inerties

L'*inertie de $\mathcal{N}(\Omega)$ par rapport à un point* $\mathbf{a}$ est définie par

$$\mathcal{I}_{\mathbf{a}} = \sum_{\boldsymbol{x} \in \Omega} \mu_{\boldsymbol{x}} d^2(\mathbf{a}, \boldsymbol{x})$$

et l'*inertie du nuage $\mathcal{N}(\Omega)$ par rapport à une variété linéaire F* par

$$\mathcal{I}_F = \sum_{\boldsymbol{x} \in \Omega} \mu_{\boldsymbol{x}} d^2(\boldsymbol{x}, F)$$

où $d(\boldsymbol{x}, F) = d(\boldsymbol{x}, \boldsymbol{y})$ avec $\boldsymbol{y}$ projection orthogonale de $\boldsymbol{x}$ sur F.

L'inertie $\mathcal{I}_{\mathbf{g}}$ du nuage Ω par rapport à son centre de gravité est appelée simplement *Inertie du nuage* et notée $\mathcal{I}$.

9.1 Théorèmes de Huygens

— Version 1 :

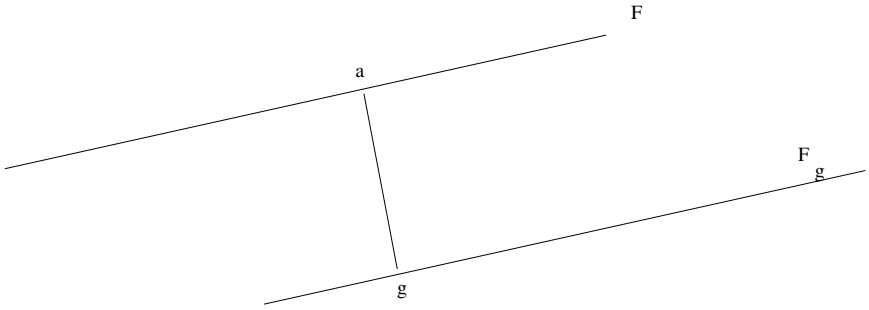
$$\mathcal{I}_{\mathbf{a}} = \mathcal{I}_{\mathbf{g}} + \mu d^2(\mathbf{a}, \mathbf{g}) \qquad \forall \mathbf{a} \in \mathbb{R}^p$$

Le centre de gravité est donc le point d'inertie minimum.

— Version 2 :

$$\mathcal{I}_F = \mathcal{I}_{F_{\mathbf{g}}} + \mu d^2(\mathbf{a}, \mathbf{g}) \qquad \forall \quad \text{variété linéaire} \quad F$$

où $F_{\mathbf{g}}$ est la variété linéaire parallèle à F passant par $\mathbf{g}$ et $\mathbf{a}$ la projection orthogonale de $\mathbf{g}$ sur F.



Le sous-espace affine parallèle à F d'inertie minimum est donc $F_{\mathbf{g}}$.

9.2 Inertie expliquée

Les propriétés d'optimalité du centre de gravité vis-à-vis de l'inertie conduisent souvent à placer celui-ci à l'origine à l'aide d'une translation. On dit alors que le nuage est centré. C'est ce que l'on supposera dans ce paragraphe. Si $\mathbb{R}^p = F \oplus F^\perp$ est une décomposition de $\mathbb{R}^p$ en 2 sous-espaces vectoriels supplémentaires orthogonaux, on peut alors montrer que l'inertie $\mathcal{I}$ se décompose suivant la relation

$$\mathcal{I} = \mathcal{I}_F + \mathcal{I}_{F^\perp}.$$

En outre, l'inertie $\mathcal{I}_{F^\perp}$, inertie du nuage par rapport à $F^\perp$, peut s'interpréter comme l'inertie du nuage des points projetés orthogonalement sur F. Pour cette raison, cette inertie est aussi appelée *inertie expliquée* par le sous-espace vectoriel F. On peut alors montrer la décomposition suivante :

$$A = B \oplus C \quad \text{et} \quad B \perp C \Rightarrow \mathcal{I}_{A^\perp} = \mathcal{I}_{B^\perp} + \mathcal{I}_{C^\perp}.$$

Bibliographie

Ball, G. H. and Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavorial Science*, 12(2):153–155.

Benzecri, J.-P. (1973). *L'analyse des données tome 1 : la taxinomie*. Dunod, Paris.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling : Theory and applications*. Springer Science & Business Media.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall / CRC, New York.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Chapman and Hall, London.

Cleveland, W. S. (1994a). *The Elements of Graphical Data*. Hobart Press, Summit, New Jersey, USA.

Cleveland, W. S. (1994b). *Visualizing Data*. Hobart Press, Summit, New Jersey, USA.

Cox, T. and Cox, M. (1994). *Multidimensional Scaling*. Chapman and Hall, London.

De Rham, C. (1980). La classification hiérarchique ascendante selon la méthode des voisins réciproques. *Les Cahiers de l'Analyse des Données*, 135:144.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification, 2nd Edition*. Wiley Interscience, New York.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, 38(2):467–474.

Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer, New York.

Govaert, G. (2003). *Analyse de données*. Hermes.

Govaert, G. (2009). *Data Analysis*. Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning - Data Mining, Inference and Prediction*. Springer, New York.

Jackson (1991). *A User's Guide to Principal Components*. Wiley, New York.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies, 1 : hierarchical systems. *Computer Journal*, 12.

Lebart, L., Morineau, A., and Piron, M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod, Paris.

MacQueen, J. B. (1967). Some methods for classification and analysis of cluster analysis. In LeCam, L. M. and Neyman, J., editors, *Proceedings of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 281–297, CA. University of California Press.

Prim, R. C. (1957). Shortest connection network and some generalizations. *Bell System Tech. Journal*, 36.

Ruspini, E. H. (1969). A new approach to clustering. *Information and Control*, 15:22–32.

Saporta, G. (2006). *Probabilités, analyse de données et statistique, 2e édition révisée et augmentée*. Technip, Paris.

Sutcliffe, J. (1994). On the logical necessity and priority of a monothetic conception of class, and on the consequent inadequacy of polythetic accounts of category and categorisation. In Diday, E., editor, *New approaches in Classification and data analysis*, pages 53–63, Berlin. Springer-Verlag.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachussets.

Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Chapman & Hall, London.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association"*, 58:236–244.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Index