

**SY09**  
**Analyse de Données et Data Mining**

Gérard Govaert

Printemps 2008



# Table des matières

Notations . . . . .	7
<b>1 Introduction</b>	<b>9</b>
<b>2 Méthodes exploratoires élémentaires</b>	<b>11</b>
2.1 Les données individus-variables . . . . .	11
2.1.1 Variables quantitatives . . . . .	11
2.1.2 Variables qualitatives . . . . .	12
2.1.3 Variables binaires . . . . .	13
2.1.4 Transformation de variables . . . . .	14
2.2 Descriptions des variables quantitatives . . . . .	15
2.2.1 Description monodimensionnelle . . . . .	15
2.2.2 Description bidimensionnelle . . . . .	17
2.2.3 Description multidimensionnelle . . . . .	19
2.3 Descriptions des variables qualitatives . . . . .	22
2.3.1 Description monodimensionnelle . . . . .	22
2.3.2 Tableaux de contingence . . . . .	23
<b>3 Vecteur aléatoire</b>	<b>25</b>
3.1 Introduction . . . . .	25
3.2 Rappels sur les variables aléatoires . . . . .	25
3.3 Vecteurs aléatoires . . . . .	27
3.3.1 Définition . . . . .	27
3.3.2 Loi jointe . . . . .	27
3.3.3 Lois marginales . . . . .	27
3.3.4 Espérance . . . . .	28
3.3.5 Matrice de variance . . . . .	28
3.3.6 Indépendance de variables aléatoires . . . . .	29
3.4 Statistiques associées à un vecteur aléatoire . . . . .	29
3.5 Loi normale multidimensionnelle . . . . .	30
3.5.1 Loi normale monodimensionnelle . . . . .	30
3.5.2 Loi normale bidimensionnelle . . . . .	30
3.5.3 Généralisation . . . . .	31
3.5.4 Propriétés . . . . .	31
3.5.5 Caractérisation de la distribution . . . . .	31
3.5.6 Simulation d'un échantillon gaussien . . . . .	31
<b>4 Distance et représentation euclidienne</b>	<b>35</b>
4.1 Tableaux de proximités . . . . .	35
4.1.1 Types de proximités . . . . .	35
4.1.2 Constitution d'un tableau de proximités . . . . .	36
4.1.3 Transformation . . . . .	36
4.1.4 Utilisation . . . . .	36
4.2 Rappels de géométrie et de mécanique . . . . .	36
4.2.1 Nuage de points . . . . .	36
4.2.2 Inertie . . . . .	37

4.2.3	Théorèmes de Huygens . . . . .	37
4.2.4	Nuage centré . . . . .	37
4.2.5	Inertie expliquée . . . . .	37
4.2.6	Expressions matricielles des inerties . . . . .	38
4.3	Représentation euclidienne des données . . . . .	38
4.4	Interprétation statistique . . . . .	38
4.4.1	Tableau centré en colonne . . . . .	38
4.4.2	Variables normées . . . . .	39
<b>5</b>	<b>L'analyse en composantes principales</b> . . . . .	<b>41</b>
5.1	Introduction . . . . .	41
5.2	Axes principaux d'inertie . . . . .	42
5.2.1	Formulation mathématique . . . . .	42
5.2.2	Résultats préalables . . . . .	42
5.2.3	Résolution du problème . . . . .	42
5.2.4	Résultats pratiques . . . . .	43
5.2.5	Inerties expliquées . . . . .	43
5.2.6	Choix du nombre d'axes à retenir . . . . .	43
5.3	Composantes principales . . . . .	44
5.3.1	Définition . . . . .	44
5.3.2	Calcul des composantes principales . . . . .	44
5.3.3	Composantes principales : nouvelles variables . . . . .	44
5.4	Formule de reconstitution . . . . .	45
5.5	Qualité de la représentation . . . . .	46
5.5.1	Qualité globale . . . . .	46
5.5.2	Contribution relative d'un axe à un individu . . . . .	46
5.5.3	Contribution relative d'un individu à un axe . . . . .	46
5.6	Représentation des variables . . . . .	46
5.7	Éléments supplémentaires . . . . .	47
5.7.1	Individu supplémentaire . . . . .	47
5.7.2	Variable supplémentaire . . . . .	47
5.7.3	Importance pratique des éléments supplémentaires . . . . .	47
5.8	Un exemple d'ACP . . . . .	47
5.8.1	Les données . . . . .	47
5.8.2	Centrage du tableau de données . . . . .	48
5.8.3	Matrice de variance . . . . .	48
5.8.4	Axes principaux d'inertie . . . . .	48
5.8.5	Qualité de la représentation . . . . .	48
5.8.6	Composantes principales . . . . .	49
5.8.7	Contributions relatives des axes aux individus . . . . .	50
5.8.8	Contributions relatives des individus aux axes . . . . .	50
5.8.9	Analyse dans $\mathbb{R}^n$ . . . . .	50
<b>6</b>	<b>Positionnement multidimensionnel</b> . . . . .	<b>53</b>
6.1	Introduction . . . . .	53
6.2	Le problème . . . . .	53
6.3	Distances euclidiennes . . . . .	53
6.3.1	Équivalence entre distances euclidiennes et produits scalaires . . . . .	53
6.3.2	Matrice de distances euclidiennes . . . . .	54
6.3.3	CNS pour qu'une matrice de dissimilarités soit euclidienne . . . . .	54
6.4	Analyse factorielle d'un tableau de distances . . . . .	55
6.4.1	$W = -\frac{1}{2}Q_n\Delta^2Q_n$ est SDP . . . . .	55
6.4.2	$W = -\frac{1}{2}Q_nD^2Q_n$ n'est pas SDP . . . . .	55
6.4.3	L'AFTD dans $\mathbb{R}$ . . . . .	56
6.4.4	Un exemple . . . . .	56
6.5	Méthodes non linéaires . . . . .	57
6.5.1	Fonctions Stress . . . . .	57

6.5.2	Optimisation . . . . .	57
6.5.3	Projection de Sammon . . . . .	57
6.5.4	Remarques . . . . .	58
6.6	Méthodes non métriques ou ordinales . . . . .	58
6.6.1	Généralisation . . . . .	58
6.6.2	Projection de Kruskal . . . . .	58
6.7	Quelques remarques . . . . .	58
6.7.1	Dissimilarités initiales . . . . .	58
6.7.2	Autres méthodes . . . . .	58
<b>7</b>	<b>La classification automatique</b>	<b>61</b>
7.1	Introduction . . . . .	61
7.2	Structures de Classification . . . . .	62
7.2.1	Partition . . . . .	62
7.2.2	La hiérarchie indicée . . . . .	62
7.2.3	Partition et hiérarchie . . . . .	63
7.2.4	Aspects combinatoires . . . . .	63
7.3	Liens avec la notion d'ultramétrie . . . . .	64
7.3.1	Recherche de partitions associées à une mesure de dissimilarité . . . . .	64
7.3.2	Ultramétrie associée à une hiérarchie indicée : fonction $\varphi$ . . . . .	65
7.3.3	Hiérarchie indicée associée à une ultramétrie : fonction $\psi$ . . . . .	65
7.3.4	Équivalence entre hiérarchie indicée et ultramétrie . . . . .	65
7.3.5	Exemples . . . . .	66
7.4	Objectifs de la classification . . . . .	66
7.4.1	Difficultés de caractériser les objectifs . . . . .	66
7.4.2	Démarche numérique . . . . .	67
7.4.3	Démarche algorithmique . . . . .	68
7.5	La classification ascendante hiérarchique . . . . .	68
7.5.1	L'algorithme . . . . .	68
7.5.2	Les critères d'agrégation . . . . .	69
7.5.3	Formule de récurrence de Lance et Williams . . . . .	70
7.5.4	Un exemple . . . . .	70
7.5.5	Méthode de Ward . . . . .	71
7.5.6	Propriétés d'optimalité . . . . .	71
7.5.7	Utilisation des méthodes . . . . .	74
7.6	Recherche de partitions . . . . .	74
7.6.1	La méthode des centres-mobiles . . . . .	74
7.6.2	Généralisation : la méthode des nuées dynamiques . . . . .	77
7.6.3	Mise en œuvre . . . . .	79
<b>8</b>	<b>Modèles probabilistes en classification</b>	<b>81</b>
8.1	Introduction . . . . .	81
8.2	Approches probabilistes de la classification . . . . .	81
8.2.1	Approches paramétriques . . . . .	82
8.2.2	Approches non paramétriques . . . . .	82
8.2.3	Validation . . . . .	83
8.2.4	Notations . . . . .	83
8.3	Le modèle de mélange . . . . .	84
8.3.1	Introduction . . . . .	84
8.3.2	Le modèle . . . . .	84
8.3.3	Exemples . . . . .	86
8.3.4	Estimation des paramètres . . . . .	86
8.3.5	Nombre de composants . . . . .	88
8.3.6	Identifiabilité . . . . .	88
8.3.7	Estimation du maximum de vraisemblance . . . . .	88
8.4	Algorithme EM . . . . .	89
8.4.1	Données complétées et vraisemblance complétée . . . . .	89

8.4.2	Principe . . . . .	89
8.4.3	Propriétés . . . . .	89
8.4.4	Application au modèle de mélange . . . . .	90
8.4.5	Exemple des mélanges gaussiens monodimensionnel à 2 composants . . . . .	90
8.5	Classification et modèle de mélange . . . . .	91
8.5.1	Les deux approches . . . . .	91
8.5.2	La vraisemblance classifiante . . . . .	91
8.5.3	L'algorithme CEM . . . . .	92
8.5.4	Comparaison des deux approches . . . . .	92
8.5.5	Classification floue . . . . .	92
8.6	Modèle de mélange gaussien . . . . .	93
8.6.1	Le modèle . . . . .	93
8.6.2	L'algorithme CEM . . . . .	95
8.6.3	Forme sphérique, proportions et volumes identiques . . . . .	95
8.6.4	Forme sphérique, proportions identiques, volumes différents . . . . .	96
8.6.5	Formes diagonales identiques, proportions identiques . . . . .	96
8.6.6	Formes identiques, proportions identiques . . . . .	97
8.6.7	Cas général, proportion identique . . . . .	97
8.7	Mise en œuvre . . . . .	98
8.7.1	Choix du modèle et du nombre de classes . . . . .	98
8.7.2	Stratégies d'utilisation . . . . .	98
<b>A</b>	<b>Quelques résultats</b>	<b>99</b>
A.1	Trois minimisations classiques . . . . .	99
A.2	Minimisations matricielles . . . . .	100
<b>B</b>	<b>Outils d'algèbre linéaire</b>	<b>103</b>
B.1	Espace vectoriel . . . . .	103
B.2	Applications linéaires et matrices . . . . .	104
B.3	Changement de base . . . . .	104
B.4	Vecteurs et valeurs propres d'un endomorphisme . . . . .	105
B.5	Produit scalaire, norme, distance et orthogonalité . . . . .	105
B.6	Matrices symétriques et matrices Q-symétriques . . . . .	109

# Notations

$\text{diag}(A)$  vecteur colonne défini par la diagonale de  $A$  si  $A$  est une matrice carrée et matrice diagonale de diagonale  $A$  si  $A$  est un vecteur

$d_p$  vecteur colonne de dimension  $n$  des pondérations  $p_i$

$D_p$  matrice diagonale de dimension  $n$  des pondérations  $p_i$

$\mathbb{1}_n$  vecteur colonne de dimension  $n$  rempli de 1.

$I_n$  matrice unité de dimension  $n$

$U_n$  matrice carrée de dimension  $n$  remplie de 1.

$X$  matrice des données de dimension  $(n, p)$



# 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<sup>e</sup> 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,...) et la constitution d'entrepôt de données (*data warehouse*) ont encore accentué ce besoin

d'analyse. Le perfectionnement des interfaces offrent 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, 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 attaque cardiaque, aura une seconde attaque 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 de défaillance 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 du cours

Il est classique de distinguer deux phases : une phase exploratoire et une phase d'apprentissage ou phase décisionnelle. 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 relation 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 : Flury (1997), Duda et al. (2001), Govaert (2003), Lebart et al. (1995) et Saporta (1990).

## Chapitre 2

# 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 se présentent généralement comme de tableaux rectangulaires pouvant prendre plusieurs formes. Les plus courants sont les tableaux *individus-variables* regroupent dans un tableau numérique de dimension  $(n, p)$ , les valeurs prises par un ensemble de  $n$  *individus* pour  $p$  *variables*.

### 2.1 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  $\Omega$  de  $n$  *individus* pour lesquels on connaît la valeur de  $p$  *variables*.

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

FIG. 2.1 – Tableau de données

On notera  $X = (x_{ij})$  la matrice réelle à  $n$  lignes et  $p$  colonnes associée au données et  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  et  $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})'$  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  $X_1, \dots, X_p$  sera noté  $\mathbf{X} = (X_1, \dots, X_p)'$ .

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

#### 2.1.1 Variables quantitatives

La variable est dite *quantitative* lorsqu'il s'agit d'une application de l'ensemble des individus  $\Omega$  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, comme 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* 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.

	<i>math</i>	<i>scie</i>	<i>fran</i>	<i>lati</i>	<i>d - m</i>
<i>jean</i>	6.0	6.0	5	5.5	8
<i>alin</i>	8.0	8.0	8	8.0	9
<i>anni</i>	6.0	7.0	11	9.5	11
<i>moni</i>	14.5	14.5	16	15.0	8
<i>didi</i>	14.0	14.0	12	12.5	10
<i>andr</i>	11.0	10.0	6	7.0	13
<i>pier</i>	5.5	7.0	14	11.5	10
<i>brig</i>	13.0	12.5	8	9.5	12
<i>evel</i>	9.0	9.5	12	12.0	18

TAB. 2.1 – Les données Notes

Le tableau 2.2 a été établi par les enfants d'une classe élémentaire : après avoir collecté 24 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. Ces papillons appartenaient à différentes espèces et l'objectif est d'essayer de retrouver ces espèces à partir uniquement des 4 mesures.

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

TAB. 2.2 – Les données Papillon

### 2.1.2 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 ordinaire 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 aux données.

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.

	<i>Esp.</i>	<i>LoSe</i>	<i>LaSe</i>	<i>LoPe</i>	<i>LaPe</i>		<i>Esp.</i>	<i>LoSe</i>	<i>LaSe</i>	<i>LoPe</i>	<i>LaPe</i>		<i>Esp.</i>	<i>LoSe</i>	<i>LaSe</i>	<i>LoPe</i>	<i>LaPe</i>
1	1	5.1	3.5	1.4	0.2	51	2	7.0	3.2	4.7	1.4	101	3	6.3	3.3	6.0	2.5
2	1	4.9	3.0	1.4	0.2	52	2	6.4	3.2	4.5	1.5	102	3	5.8	2.7	5.1	1.9
3	1	4.7	3.2	1.3	0.2	53	2	6.9	3.1	4.9	1.5	103	3	7.1	3.0	5.9	2.1
4	1	4.6	3.1	1.5	0.2	54	2	5.5	2.3	4.0	1.3	104	3	6.3	2.9	5.6	1.8
5	1	5.0	3.6	1.4	0.2	55	2	6.5	2.8	4.6	1.5	105	3	6.5	3.0	5.8	2.2
6	1	5.4	3.9	1.7	0.4	56	2	5.7	2.8	4.5	1.3	106	3	7.6	3.0	6.6	2.1
7	1	4.6	3.4	1.4	0.3	57	2	6.3	3.3	4.7	1.6	107	3	4.9	2.5	4.5	1.7
8	1	5.0	3.4	1.5	0.2	58	2	4.9	2.4	3.3	1.0	108	3	7.3	2.9	6.3	1.8
9	1	4.4	2.9	1.4	0.2	59	2	6.6	2.9	4.6	1.3	109	3	6.7	2.5	5.8	1.8
10	1	4.9	3.1	1.5	0.1	60	2	5.2	2.7	3.9	1.4	110	3	7.2	3.6	6.1	2.5
11	1	5.4	3.7	1.5	0.2	61	2	5.0	2.0	3.5	1.0	111	3	6.5	3.2	5.1	2.0
12	1	4.8	3.4	1.6	0.2	62	2	5.9	3.0	4.2	1.5	112	3	6.4	2.7	5.3	1.9
13	1	4.8	3.0	1.4	0.1	63	2	6.0	2.2	4.0	1.0	113	3	6.8	3.0	5.5	2.1
14	1	4.3	3.0	1.1	0.1	64	2	6.1	2.9	4.7	1.4	114	3	5.7	2.5	5.0	2.0
15	1	5.8	4.0	1.2	0.2	65	2	5.6	2.9	3.6	1.3	115	3	5.8	2.8	5.1	2.4
16	1	5.7	4.4	1.5	0.4	66	2	6.7	3.1	4.4	1.4	116	3	6.4	3.2	5.3	2.3
17	1	5.4	3.9	1.3	0.4	67	2	5.6	3.0	4.5	1.5	117	3	6.5	3.0	5.5	1.8
18	1	5.1	3.5	1.4	0.3	68	2	5.8	2.7	4.1	1.0	118	3	7.7	3.8	6.7	2.2
19	1	5.7	3.8	1.7	0.3	69	2	6.2	2.2	4.5	1.5	119	3	7.7	2.6	6.9	2.3
20	1	5.1	3.8	1.5	0.3	70	2	5.6	2.5	3.9	1.1	120	3	6.0	2.2	5.0	1.5
21	1	5.4	3.4	1.7	0.2	71	2	5.9	3.2	4.8	1.8	121	3	6.9	3.2	5.7	2.3
22	1	5.1	3.7	1.5	0.4	72	2	6.1	2.8	4.0	1.3	122	3	5.6	2.8	4.9	2.0
23	1	4.6	3.6	1.0	0.2	73	2	6.3	2.5	4.9	1.5	123	3	7.7	2.8	6.7	2.0
24	1	5.1	3.3	1.7	0.5	74	2	6.1	2.8	4.7	1.2	124	3	6.3	2.7	4.9	1.8
25	1	4.8	3.4	1.9	0.2	75	2	6.4	2.9	4.3	1.3	125	3	6.7	3.3	5.7	2.1
26	1	5.0	3.0	1.6	0.2	76	2	6.6	3.0	4.4	1.4	126	3	7.2	3.2	6.0	1.8
27	1	5.0	3.4	1.6	0.4	77	2	6.8	2.8	4.8	1.4	127	3	6.2	2.8	4.8	1.8
28	1	5.2	3.5	1.5	0.2	78	2	6.7	3.0	5.0	1.7	128	3	6.1	3.0	4.9	1.8
29	1	5.2	3.4	1.4	0.2	79	2	6.0	2.9	4.5	1.5	129	3	6.4	2.8	5.6	2.1
30	1	4.7	3.2	1.6	0.2	80	2	5.7	2.6	3.5	1.0	130	3	7.2	3.0	5.8	1.6
31	1	4.8	3.1	1.6	0.2	81	2	5.5	2.4	3.8	1.1	131	3	7.4	2.8	6.1	1.9
32	1	5.4	3.4	1.5	0.4	82	2	5.5	2.4	3.7	1.0	132	3	7.9	3.8	6.4	2.0
33	1	5.2	4.1	1.5	0.1	83	2	5.8	2.7	3.9	1.2	133	3	6.4	2.8	5.6	2.2
34	1	5.5	4.2	1.4	0.2	84	2	6.0	2.7	5.1	1.6	134	3	6.3	2.8	5.1	1.5
35	1	4.9	3.1	1.5	0.2	85	2	5.4	3.0	4.5	1.5	135	3	6.1	2.6	5.6	1.4
36	1	5.0	3.2	1.2	0.2	86	2	6.0	3.4	4.5	1.6	136	3	7.7	3.0	6.1	2.3
37	1	5.5	3.5	1.3	0.2	87	2	6.7	3.1	4.7	1.5	137	3	6.3	3.4	5.6	2.4
38	1	4.9	3.6	1.4	0.1	88	2	6.3	2.3	4.4	1.3	138	3	6.4	3.1	5.5	1.8
39	1	4.4	3.0	1.3	0.2	89	2	5.6	3.0	4.1	1.3	139	3	6.0	3.0	4.8	1.8
40	1	5.1	3.4	1.5	0.2	90	2	5.5	2.5	4.0	1.3	140	3	6.9	3.1	5.4	2.1
41	1	5.0	3.5	1.3	0.3	91	2	5.5	2.6	4.4	1.2	141	3	6.7	3.1	5.6	2.4
42	1	4.5	2.3	1.3	0.3	92	2	6.1	3.0	4.6	1.4	142	3	6.9	3.1	5.1	2.3
43	1	4.4	3.2	1.3	0.2	93	2	5.8	2.6	4.0	1.2	143	3	5.8	2.7	5.1	1.9
44	1	5.0	3.5	1.6	0.6	94	2	5.0	2.3	3.3	1.0	144	3	6.8	3.2	5.9	2.3
45	1	5.1	3.8	1.9	0.4	95	2	5.6	2.7	4.2	1.3	145	3	6.7	3.3	5.7	2.5
46	1	4.8	3.0	1.4	0.3	96	2	5.7	3.0	4.2	1.2	146	3	6.7	3.0	5.2	2.3
47	1	5.1	3.8	1.6	0.2	97	2	5.7	2.9	4.2	1.3	147	3	6.3	2.5	5.0	1.9
48	1	4.6	3.2	1.4	0.2	98	2	6.2	2.9	4.3	1.3	148	3	6.5	3.0	5.2	2.0
49	1	5.3	3.7	1.5	0.2	99	2	5.1	2.5	3.0	1.1	149	3	6.2	3.4	5.4	2.3
50	1	5.0	3.3	1.4	0.2	100	2	5.7	2.8	4.1	1.3	150	3	5.9	3.0	5.1	1.8

TAB. 2.3 – Les données Iris

### 2.1.3 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, par exemple, les entiers de 0 à 20.

### 2.1.4 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 ordinaire 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 ordinaire, 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

## 2.2 Descriptions des variables quantitatives

### 2.2.1 Description monodimensionnelle

#### Statistiques élémentaires

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

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

la médiane qui partage les valeurs ordonnées prises par la variable en deux parties égales. On peut généraliser la 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

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

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 2.2 fournit quelques unes de ces statistiques pour les variables quantitatives des données Iris.

	LoSe	LaSe	LoPe	LaPe
Min.	:4.300	:2.000	:1.000	:0.100
1st Qu.	:5.100	:2.800	:1.600	:0.300
Median	:5.800	:3.000	:4.350	:1.300
Mean	:5.843	:3.057	:3.758	:1.199
3rd Qu.	:6.400	:3.300	:5.100	:1.800
Max.	:7.900	:4.400	:6.900	:2.500

FIG. 2.2 – Description élémentaire

#### Histogramme

L'histogramme représente une estimation de la fonction de densité. Pour le tracer, 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. 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  $1 + \frac{10}{3} \log n$ ,  $n$  étant la taille de l'échantillon. La partie gauche de la figure 2.3 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. Toujours avec la variable LoPe, l'histogramme obtenu dans la partie droite de la même figure montre clairement que la variable longueur du pétale discrimine les 50 premières fleurs des suivantes.

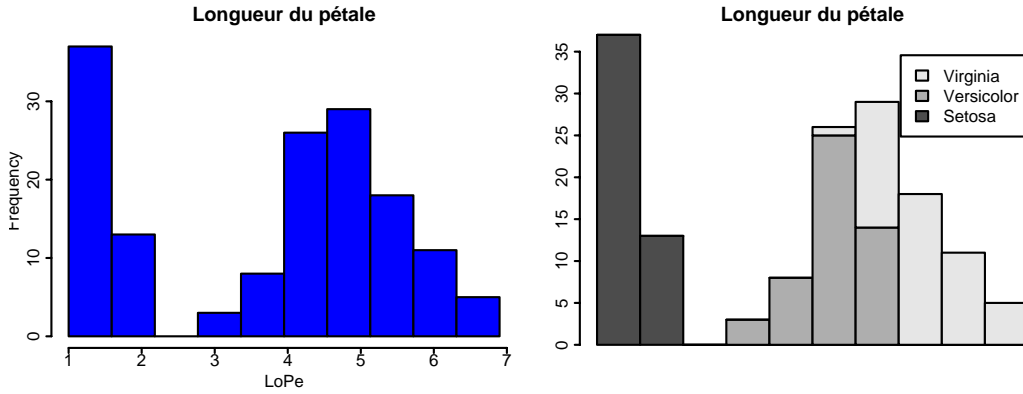


FIG. 2.3 – Histogrammes des données Iris

### Estimation de la densité par la méthode des noyaux

Sachant que l'histogramme est une estimation de la fonction de densité, généralement continue, il paraît souhaitable d'estimer cette fonction de densité par une fonction plus régulière que l'histogramme. Il est possible d'obtenir de telles estimations par la méthode des « noyaux » définie par

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

où  $K$ , appelé le noyau, est une fonction de densité centrée en 0 et  $h$  est un nombre réel qui règle le degré de régularité.

Les principaux noyaux utilisés sont les suivants :

- noyau rectangulaire :  $K(x) = \mathbb{1}_{[-0.5, +0.5]}(x)$  ;
- noyau triangulaire :  $K(x) = (1 - |x|) \cdot \mathbb{1}_{[-1, +1]}(x)$  ;
- noyau gaussien :  $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$  ;
- noyau d'Epanechnikov :  $K(x) = \frac{3}{4\sqrt{5}}(1 - x^2/5) \cdot \mathbb{1}_{[-\sqrt{5}, +\sqrt{5}]}(x)$  ;
- noyau de Lejeune :  $K(x) = \frac{105}{64}(1 - x^2)^2(1 - 3x^2) \cdot \mathbb{1}_{[-1, +1]}(x)$ .

Le choix du noyau n'est pas très important. Par contre, l'estimateur sera très sensible au choix de  $h$ . Par ailleurs, le noyau de Lejeune se comporte correctement bien que la condition de positivité ne soit pas remplie. La figure 2.4 représente un échantillon de taille 50 et l'estimation de densité obtenue avec le noyau gaussien et l'histogramme obtenu avec 9 classes.

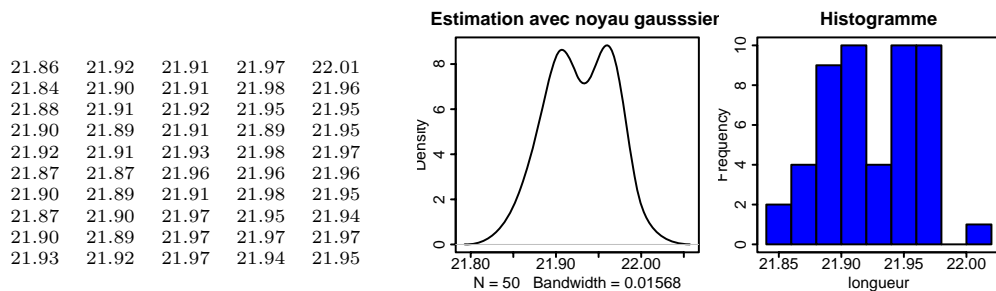
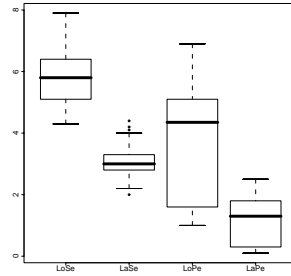


FIG. 2.4 – Estimation de densité et histogramme

### Diagramme en boîte

La figure 2.5 représente les diagrammes en boîte, encore appelés boîte à moustaches ou *boxplots*, associés à chacune des 4 variables des données Iris.

FIG. 2.5 – Diagramme en boîte pour les 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 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.

## 2.2.2 Description bidimensionnelle

### 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 2.6 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.

### Covariance et corrélation

Pour étudier les liens entre deux variables quantitatives, on utilise souvent la covariance et le coefficient de corrélation linéaire empirique et lorsque l'on a plus de deux variables, on utilise la matrice de covariance et la matrice de corrélation de l'échantillon. La table 2.4 représente les matrices de variance et de corrélation obtenues avec les données *Iris*. Le coefficient de corrélation linéaire est à utiliser avec prudence : il est en effet difficile à un seul nombre de caractériser entièrement le lien qui peut exister entre deux variables. Par exemple la figure 2.7, 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 ;

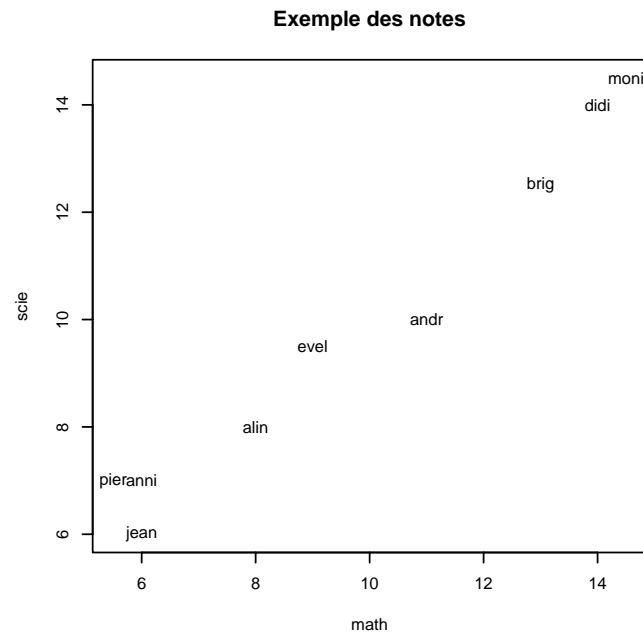


FIG. 2.6 – Graphique de dispersion pour les données Notes

	LoSe	laSe	LoPe	laPe
LoSe	0.68			
laSe	-0.04	0.19		
LoPe	1.27	-0.33	3.10	
laPe	0.51	-0.12	1.29	0.58
	LoSe	laSe	LoPe	laPe
LoSe	1.00			
laSe	-0.12	1.00		
LoPe	0.87	-0.43	1.00	
laPe	0.82	-0.37	0.96	1.00

TAB. 2.4 – Matrices de variance et de corrélation des données Iris

- 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.

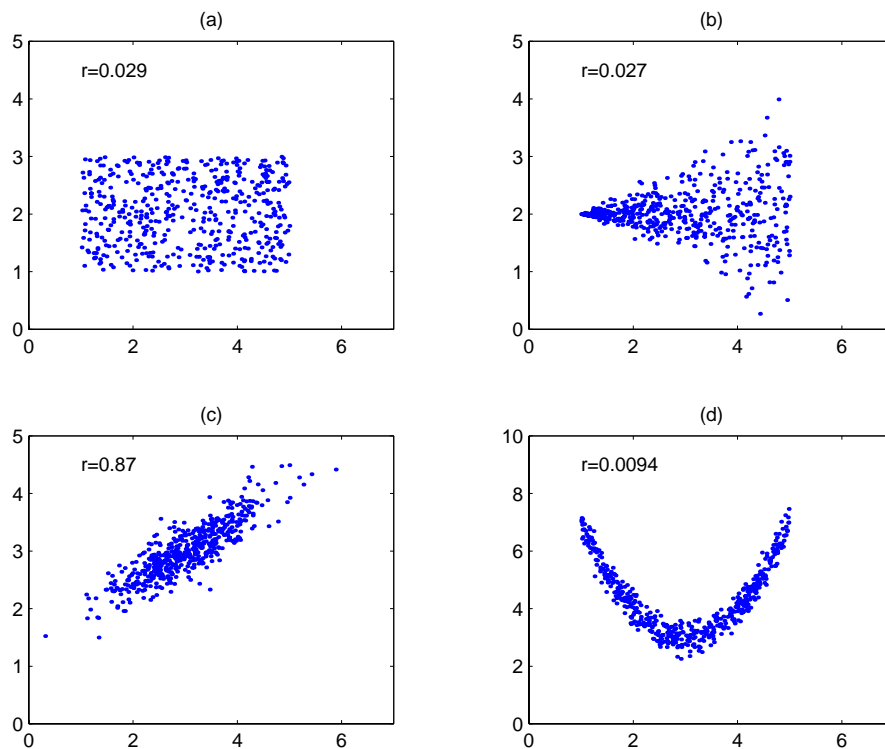


FIG. 2.7 – Exemple de corrélations

La figure 2.8 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, 3 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.

## 2.2.3 Description multidimensionnelle

### 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 2.9 représente le graphique obtenu avec les données *Iris*.

### Quelques outils variés

Enfin, parmi les nombreuses méthodes qui ont été proposées pour décrire graphiquement des données, on peut citer celles qui synthétisent chaque individu par une forme caractérisée par les valeurs des variables prises par l'individu.

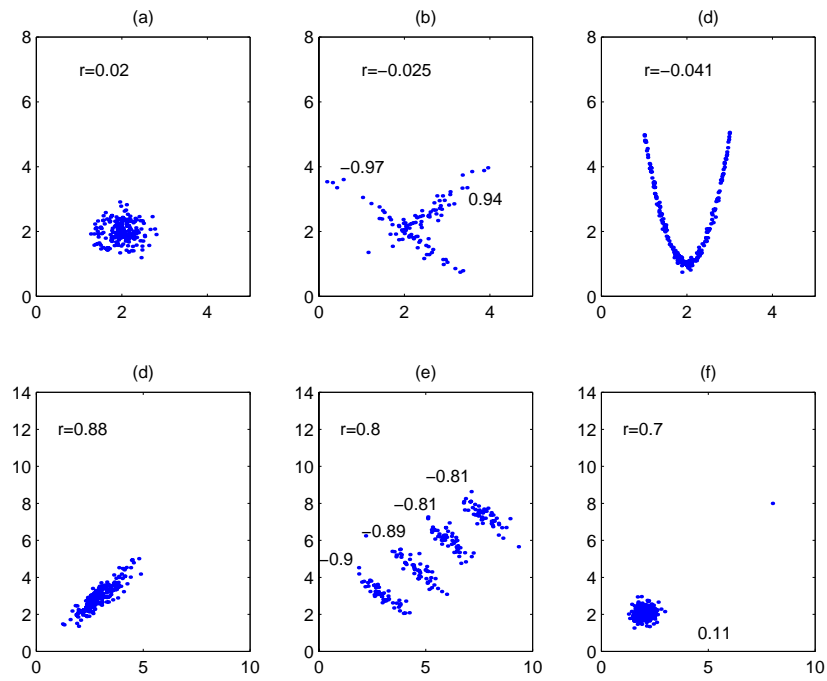


FIG. 2.8 – Exemple de corrélations

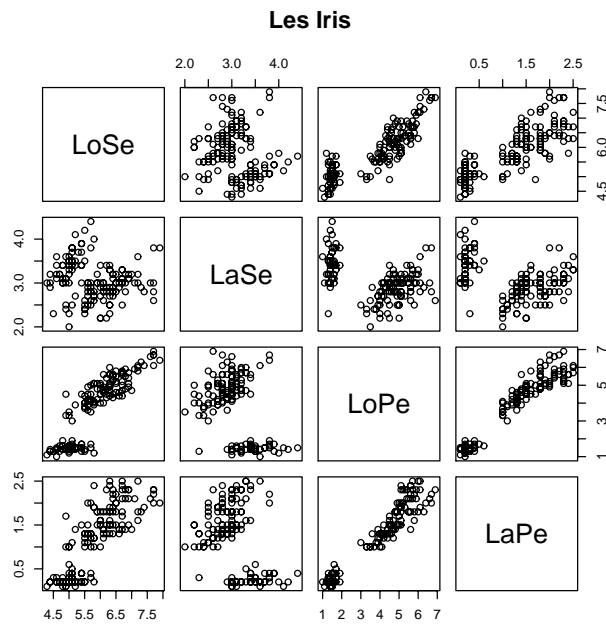


FIG. 2.9 – Graphique matriciel

- Les Visages de Chernoff : la méthode des visages de Chernoff associe un visage à chaque individu en utilisant les variables pour définir les caractéristiques de ce visage (forme du visage, sourcil, sourire, ...); la figure 2.10 suivante montre ce que l'on obtient avec les données notes;

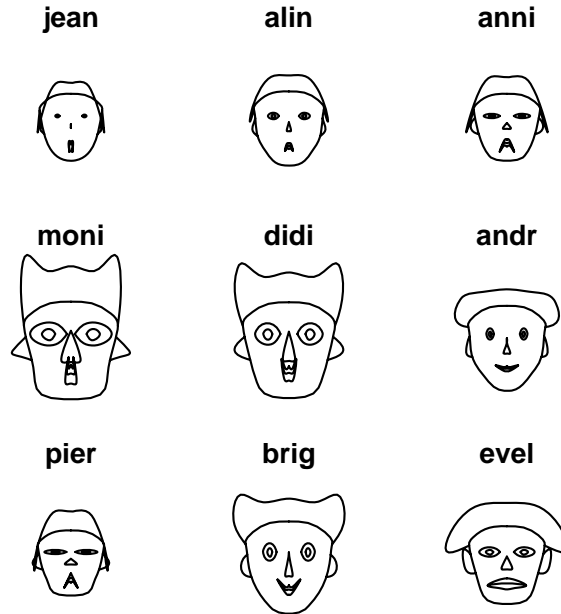


FIG. 2.10 – Visage de Chernoff

- Les Polygones : il s'agit d'une méthode représentant chaque individu par un polygone à  $p$  sommets défini de la manière suivante : les sommets sont obtenus en traçant  $p$  rayons régulièrement répartis d'un cercle et en prenant comme sommets les points  $M_j$  de ces rayons définis par  $OM_j = \frac{x_j - \min(x^j)}{\max(x^j) - \min(x^j)}$ ; la figure 2.11 montre ce que l'on obtient toujours avec les mêmes données notes.

### 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.

### 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.

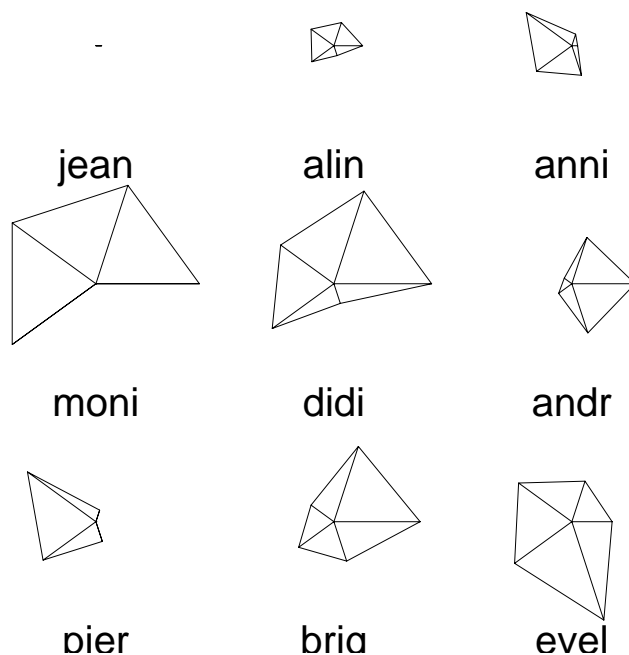


FIG. 2.11 – Polygones

**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 2.12 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.

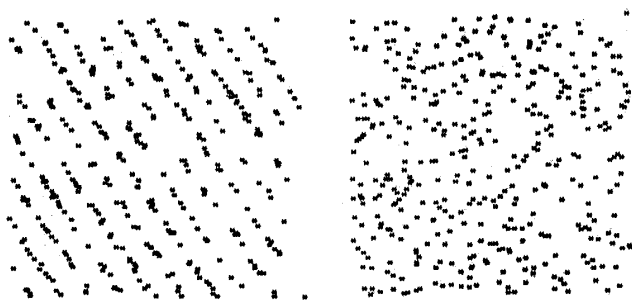


FIG. 2.12 – Deux projections voisines

## 2.3 Descriptions des variables qualitatives

### 2.3.1 Description monodimensionnelle

Une distribution de  $n$  observations associée à une variable qualitative peut être présentée sous forme d'un tableau de fréquences où figurent pour chaque modalité  $\xi_k$ , le nombre  $n_k$  (appelé effectif ou fréquence) d'observations ayant la valeur  $\xi_k$ , la fréquence relative  $f_k = n_k/n$  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. Une autre représentation graphique souvent utilisée est

le diagramme en « camembert ». L'exemple de la figure 2.13 représente les descriptions ainsi obtenues de la variable `espèce` pour le sous-échantillon des iris dont la longueur du sépale est supérieure à 5.

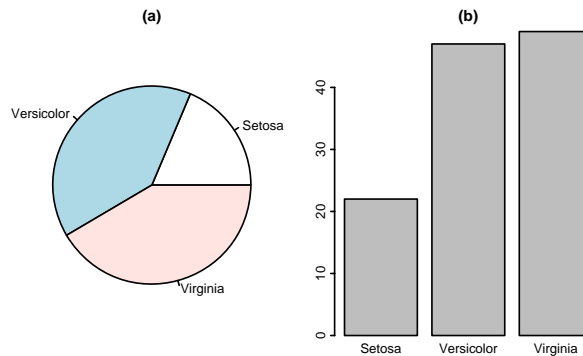


FIG. 2.13 – Description de la variable Espèce pour les Iris

### 2.3.2 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. L'exemple suivant fournit le tableau de contingence obtenu en croisant deux variables ayant respectivement 2 et 3 modalités.

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

	1	2	3
1	1	1	0
2	1	1	2

#### 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, est appelé tableau de contingence. Ainsi le tableau suivant (données « budget-temps »), qui regroupe le nombre d'heures passées à pratiquer une des 10 classes d'activité (profession, transport, ménage, enfants, courses, toilette, repas, sommeil, télévision, 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, constitue un tableau de contingence.

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

Les utilisateurs de l'analyse de données vont quelque fois plus loin et appliquent les méthodes destinées aux tableaux de contingence comme l'analyse des correspondances dès lors que toutes les valeurs sont positives et homogènes (ensemble de mesures de même nature par exemple). Par ailleurs, certains tableaux binaires peuvent être considérés comme des tableaux de contingence et peuvent même en être issus, en remplaçant la fréquence par la valeur 0 ou 1 indiquant l'absence ou la présence. Remarquons qu'un tableau binaire peut donc être considéré à la fois comme un tableau individus-variables binaires et comme un tableau de contingence particulier. Ce second aspect a le mérite de mettre en avant une symétrie possible des données.

# Chapitre 3

## Vecteur aléatoire

### 3.1 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 révélé 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<sup>1</sup>.

### 3.2 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  $\Omega$  définissant l'ensemble des résultats possibles de  $\mathcal{E}$ , appelés *événements élémentaires*;
2. un ensemble  $\mathcal{A}$  de parties de  $\Omega$ , 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  $\mathbb{P} : \mathcal{A} \rightarrow [0, 1]$ , appelée *mesure* ou *distribution* de probabilité, qui à tout événement  $A$  associe un nombre  $\mathbb{P}(A)$  appelé probabilité de cet événement.

#### 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

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega). \end{aligned}$$

---

<sup>1</sup>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.

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

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in B\}) = \mathbb{P}(X^{-1}(B)),$$

quantité notée simplement  $\mathbb{P}(X \in B)$ . La donnée de  $\mathbb{P}_X(B)$  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  $\mathbb{P}_X$  est une mesure de probabilité sur l'espace probabilisable  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , appelée mesure image de  $\mathbb{P}$  par  $X$ . Pour décrire complètement  $\mathbb{P}_X$ , 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

$$F_X : \begin{array}{l} \mathbb{R} \rightarrow [0, 1] \\ x \rightarrow \mathbb{P}_X(] - \infty, x]), \end{array}$$

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 :

$$p_X : \begin{array}{l} \mathbb{R} \rightarrow [0, 1] \\ x \rightarrow \mathbb{P}_X(\{x\}) \end{array}$$

et qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}), \mathbb{P}_X(B) = \sum_{x \in B} p_X(x)$$

- et dans le cas continu, par la *fonction de densité de probabilité* qui vérifie

$$\forall B \in \mathcal{B}(\mathbb{R}), \mathbb{P}_X(B) = \int_B f_X(t) dt.$$

### Espérance mathématique

L'espérance mathématique d'une variable aléatoire réelle, qui représente la « valeur moyenne » prise par cette variable aléatoire, est définie par

$$\mathbb{E}(X) = \begin{cases} \sum_{x \in V_X} xp_X(x) & \text{si } X \text{ est une v.a. discrète,} \\ \int_{\mathbb{R}} xf_X(x) dx & \text{si } X \text{ est une v.a. continue} \end{cases}$$

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

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

La racine carrée de la variance est appelée *écart-type* de la v.a.  $X$  et notée  $\sigma$ . La variance, étant une espérance, peut ne pas être définie.

## 3.3 Vecteurs aléatoires

### 3.3.1 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  $\mathbb{R}^n$  dont les composants sont fonctions du résultat d'une expérience aléatoire  $\mathcal{E}$ . Il s'agit donc d'une application :

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R}^n \\ \omega &\mapsto X(\omega) = (X_1(\omega), \dots, X_n(\omega)). \end{aligned}$$

### 3.3.2 Loi jointe

La loi de probabilité du vecteur aléatoire  $\mathbf{X} = (X_1, \dots, X_p)'$ , appelée loi jointe, est définie par :

$$\mathbb{P}(\mathbf{X} \in A) = \mathbb{P}_{\mathbf{X}}(A) = \mathbb{P}(\omega \in \Omega | (X_1(\omega), \dots, X_p(\omega)) \in A).$$

Comme dans le cas monodimensionnel, on peut décrire  $\mathbb{P}_{\mathbf{X}}$  de deux façons :

1. Par la fonction de répartition de  $\mathbf{X}$   $F_{\mathbf{X}} : \mathbb{R}^n \rightarrow [0, 1]$ , définie par

$$(x_1, \dots, x_n) \mapsto \mathbb{P}_{\mathbf{X}}([-\infty, x_1] \times \dots \times [-\infty, x_n]),$$

ce que l'on note  $F_{\mathbf{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1; X_2 \leq x_2; \dots; X_n \leq x_n)$ .

2. Par la fonction de probabilité (cas discret) ou de densité de probabilité (cas continu) de  $\mathbf{X}$ .

Si le vecteur aléatoire  $\mathbf{X}$  est discret, la loi de probabilité du vecteur aléatoire est entièrement définie par la fonction de probabilité

$$p_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} = \mathbf{x}) = \mathbb{P}(X_1 = x_1, \dots, X_p = x_p)$$

où  $\mathbf{x} = (x_1, \dots, x_p)'$  est un vecteur de  $\mathbb{R}^p$  et on a  $\forall B \in \mathcal{B}(\mathbb{R}^n)$

$$\mathbb{P}_{\mathbf{X}}(B) = \sum_{\mathbf{x} \in B} p_{\mathbf{X}}(\mathbf{x}).$$

Si le vecteur aléatoire  $\mathbf{X}$  est absolument continu, il admet une densité  $f_{\mathbf{X}}$ , fonction de  $\mathbb{R}^p$  dans  $\mathbb{R}$ , vérifiant pour tout produit  $A$  d'intervalles réels  $I_1, \dots, I_p$

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(\mathbf{x}) dx^1 \dots dx^p.$$

On a

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \dots \int_{-\infty}^{x_1} f_{\mathbf{X}}(t_1, \dots, t_n) dt_1 \dots dt_n \quad \text{et} \quad f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p F_{\mathbf{X}}}{\partial x_1 \dots \partial x_p}(\mathbf{x}).$$

Dans la suite et s'il n'y a pas d'ambiguïté, on notera simplement  $p$ ,  $f$  et  $F$  les fonctions  $p_{\mathbf{X}}$ ,  $f_{\mathbf{X}}$  et  $F_{\mathbf{X}}$ .

### 3.3.3 Lois marginales

Tout sous-vecteur du vecteur aléatoire  $\mathbf{X}$ , c'est-à-dire tout sous-ensemble de l'ensemble des variables aléatoires  $X_1, \dots, X_p$  est lui-même un vecteur aléatoire. La loi d'un tel vecteur aléatoire est appelée loi marginale. Si  $X_{j_1}, \dots, X_{j_q}$  est ce sous-ensemble, la loi marginale sera notée  $p_{j_1, \dots, j_q}$  ou  $f_{j_1, \dots, j_q}$  suivant le cas.

Si cet ensemble se réduit à une seule variable et

- si  $\mathbf{X}$  est discret, la loi de  $X_J$  est définie par la probabilité élémentaire

$$p_j(x_j) = \sum_{x^1 \in V_1, \dots, x^{j-1} \in V_{j-1}, x^{j+1} \in V_{j+1}, \dots, x^p \in V_p} p(\mathbf{x})$$

- et si  $\mathbf{X}$  est continu, la loi de  $X_J$  est définie par la densité

$$f_j(x_j) = \int_{\mathbb{R}^{p-1}} f(\mathbf{x}) dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_p.$$

### 3.3.4 Espérance

#### Définition

L'espérance du vecteur aléatoire  $\mathbf{X}$  est le vecteur des espérances des variables aléatoires  $X_j$  :

$$\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_p))'.$$

#### Linéarité de l'espérance

Si  $\mathbf{X}$  et  $\mathbf{Y}$  sont 2 vecteurs aléatoires de dimension  $p$  et si  $A$  et  $B$  sont 2 matrices de dimensions  $(q, p)$ , alors

$$\mathbb{E}(A\mathbf{X} + B\mathbf{Y}) = A\mathbb{E}(\mathbf{X}) + B\mathbb{E}(\mathbf{Y}).$$

En particulier, on retrouve  $\mathbb{E}(u'\mathbf{X}) = u'\mathbb{E}(\mathbf{X})$  : l'espérance d'une combinaison linéaire de variables aléatoires est la combinaison linéaire des espérances.

#### Espérance d'une fonction réelle d'un vecteur aléatoire

Si  $\varphi$  est une fonction de  $\mathbb{R}^p$  dans  $\mathbb{R}$ , on a

$$\mathbb{E}(\varphi(\mathbf{X})) = \int_{\mathbb{R}^p} \varphi(\mathbf{x})f(\mathbf{x})d\mathbf{x}$$

pour un vecteur aléatoire continu et

$$\mathbb{E}(\varphi(\mathbf{X})) = \sum_{x^1 \in V_1} \dots \sum_{x^p \in V_p} \varphi(\mathbf{x})p(\mathbf{x})$$

pour un vecteur aléatoire discret.

Comme pour les variables aléatoires, ce résultat permet de calculer l'espérance d'une v. a.  $\varphi(\mathbf{X})$  sans avoir besoin de calculer sa loi.

### 3.3.5 Matrice de variance

#### Covariance

Étant données deux variables aléatoires  $X$  et  $Y$ , on appelle covariance entre  $X$  et  $Y$  la quantité

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

En appliquant le résultat du paragraphe précédent, on obtient

$$\text{Cov}(X, Y) = \begin{cases} \int_{\mathbb{R}^2} (x - \mathbb{E}(X))(y - \mathbb{E}(Y))f_{X,Y}(x, y)dx dy & \text{si } X \text{ est continu,} \\ \sum_{x \in V_X} \sum_{y \in V_Y} (x - \mathbb{E}(X))(y - \mathbb{E}(Y))p(x, y) & \text{si } X \text{ est discret.} \end{cases}$$

Nous donnons ci-dessous quelques propriétés de la covariance.

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$ .
- Inégalité de Cauchy-Schwarz :  $[\text{Cov}(X, Y)]^2 \leq \text{Var}(X)\text{Var}(Y)$  (égalité ssi  $X - \mathbb{E}(X) = k(Y - \mathbb{E}(Y))$ ).

**Matrice de variance**

La variance du vecteur aléatoire  $\mathbf{X}$ , souvent appelée *matrice de variance*, est la matrice

$$\Sigma = \text{Var}(\mathbf{X}) = (\sigma_{jj'})_{j,j'=1,\dots,p}$$

où  $\sigma_{jj} = \sigma_j^2$  est la variance de  $X_j$  et  $\sigma_{jj'}$  est la covariance du couple  $(X_j, X_{j'})$ . Il s'agit donc d'une matrice carrée, symétrique positive de dimension  $(p, p)$  dont la diagonale est formée des variances des variables aléatoires  $X_j$ .

On obtient matriciellement

$$\Sigma = \text{Var}(\mathbf{X}) = \mathbb{E}([\mathbf{X} - \mathbb{E}(\mathbf{X})][\mathbf{X} - \mathbb{E}(\mathbf{X})]')$$

et on peut montrer

$$\text{Var}(A\mathbf{X}) = A\text{Var}(\mathbf{X})A' = A\Sigma A'$$

En particulier

$$\text{Var}(u'\mathbf{X}) = u'\Sigma u \quad \text{et} \quad \text{Cov}(u'\mathbf{X}, v'\mathbf{X}) = u'\Sigma v.$$

**Matrice de corrélation**

La matrice de corrélation du vecteur aléatoire  $\mathbf{X}$ , définie de manière analogue à la matrice de variance, est la matrice dont le terme général est le coefficient de corrélation linéaire

$$\rho_{jj'} = \frac{\text{Cov}(X_j, X_{j'})}{\sqrt{\text{Var}(X_j)\text{Var}(X_{j'})}}.$$

Toutes les valeurs sont donc comprises entre -1 et +1 et les termes de la diagonale sont égaux à 1.

**3.3.6 Indépendance de variables aléatoires**

Les  $p$  v. a. réelles  $X_1, \dots, X_p$  sont *indépendantes* si la loi jointe du vecteur aléatoire  $\mathbf{X} = (X_1, \dots, X_p)'$  s'exprime comme le produit des lois marginales, c'est-à-dire si et seulement si :

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= \prod_{j=1}^p p_{X_j}(x_j) & \text{si les v. a. sont discrètes} \\ \text{et } f_{\mathbf{X}}(\mathbf{x}) &= \prod_{j=1}^p f_{X_j}(x_j) & \text{si les v. a. sont continues.} \end{aligned}$$

Intuitivement, la notion d'indépendance entre v.a. correspond à l'absence de relation entre ces variables.

On a les propriétés suivantes :

- $X_1, \dots, X_p$  indépendantes  $\Rightarrow$  tout sous-ensemble des v. a. est indépendant ; en particulier les v. a.  $X_1, \dots, X_p$  sont indépendantes 2 à 2
- Attention, la réciproque est fautive : l'indépendance 2 à 2 n'entraîne pas l'indépendance
- $X^1, \dots, X^p$  indépendantes  $\Rightarrow \mathbb{E}(X^1 \dots X^p) = \mathbb{E}(X^1) \dots \mathbb{E}(X^p)$
- $X_j$  et  $X_{j'}$  indépendantes  $\Rightarrow \text{Cov}(X_j, X_{j'}) = 0$
- $X_1, \dots, X_p$  indépendantes  $\Rightarrow \text{Var}(\sum_{j=1}^p X_j) = \sum_{j=1}^p \text{Var}(X_j)$
- La matrice de variance sera diagonale si les variables  $X_j$  sont indépendantes 2 à 2. La réciproque est fautive.

**3.4 Statistiques associées à un vecteur aléatoire**

On considère ici disposer de la réalisation d'un échantillon de taille  $n$  du vecteur aléatoire  $\mathbf{X}$ , c'est-à-dire d'une matrice  $X$  de dimension  $(n, p)$  qui peut être interprétée de la façon suivante : chaque vecteur  $\mathbf{x}_i$  correspond à une réalisation du vecteur aléatoire  $\mathbf{X}$  ; chaque vecteur  $\mathbf{x}_j$  correspond à une réalisation d'un échantillon de taille  $n$  de la variable aléatoire  $X_j$ .

Voici les principales statistiques définies à partir de cette matrice  $X$ .

– Vecteur moyenne empirique

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_p)' \quad \text{où} \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

– Variance empirique

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

– Covariance empirique

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

On a  $s_{jj} = (s_j)^2$ .

– Coefficient de corrélation linéaire empirique

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

– Matrice de variance empirique

$$S = (s_{jj'})_{j,j'=1,\dots,p} = \frac{1}{n} (X - 1_n \bar{\mathbf{x}})' (X - 1_n \bar{\mathbf{x}}) = \frac{1}{n} Y' Y$$

où  $1_n$  est la matrice de dimension  $(n, 1)$  remplie de 1 et  $Y$  est la matrice centrée associée à  $X$ .

– Matrice de corrélation empirique

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

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

## 3.5 Loi normale multidimensionnelle

### 3.5.1 Loi normale monodimensionnelle

La loi normale  $\mathcal{N}(\mu, \sigma^2)$  (encore appelée loi de Gauss) est définie par la fonction de densité de probabilité :

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right],$$

La fonction de répartition correspondante n'a pas d'expression analytique. On l'exprime communément à l'aide de la fonction  $\phi$  suivante :  $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$ , qui est la fonction de répartition de la loi  $\mathcal{N}(0, 1)$  (appelée loi normale centrée-réduite). On a en effet la propriété suivante :

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow F_X(x) = \phi \left( \frac{x - \mu}{\sigma} \right), \forall x \in \mathbb{R}.$$

### 3.5.2 Loi normale bidimensionnelle

Soit  $X = (X_1, X_2)$  le vecteur aléatoire bidimensionnel de fonction de densité

$$f_X(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left[ -\frac{1}{2(1-\rho^2)} \left( \left( \frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right) \right]$$

avec  $\sigma_1, \sigma_2 > 0$  et  $\rho \in [-1, 1]$ .

Par définition,  $X$  suit une *loi normale bidimensionnelle*. On montre que  $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$  et  $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ .

### 3.5.3 Généralisation

La densité d'un vecteur aléatoire normal  $\mathbf{X}$  de moyenne  $\boldsymbol{\mu}$  et de matrice de variance  $\Sigma$  est définie par

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

où  $\mathbf{x} = (x_1, \dots, x_p)'$ .

### 3.5.4 Propriétés

1. Tout sous-vecteur de ce vecteur aléatoire suit encore une loi normale ; en particulier, les variables  $X_1, \dots, X_p$  sont toutes gaussiennes.
2. Pour un vecteur aléatoire gaussien, les variables  $X_1, \dots, X_p$  sont indépendantes si et seulement si la matrice de variance est diagonale.
3. Si  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$  alors  $A\mathbf{X} + \mathbf{b} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A')$

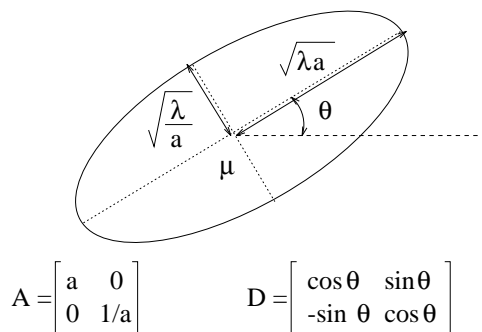
### 3.5.5 Caractérisation de la distribution

La densité est une fonction de  $\mathbb{R}^p$  dans  $\mathbb{R}$  ; il est possible de caractériser une telle densité à l'aide des ensembles d'équidensité  $f(\mathbf{x}) = cste$ . Pour la loi normale, ces ensembles sont des ellipsoïdes de  $\mathbb{R}^p$ . Le vecteur moyenne  $\boldsymbol{\mu}$  et la matrice de variance correspondent alors au centre et à la forme de ces ellipsoïdes. Il est possible d'être plus précis en s'appuyant sur la décomposition spectrale

$$\Sigma = \lambda D A D'$$

où  $\lambda$  est un réel positif,  $D$  est une matrice orthogonale et  $A$  une matrice diagonale de déterminant 1. Cette décomposition peut être interprétée de la manière suivante :  $\lambda$  caractérise le volume,  $D$  la direction et  $A$  la forme de la distribution. On peut ainsi définir une matrice de variance à partir de ces trois caractéristiques.

Par exemple, dans le cas de  $\mathbb{R}^2$ ,  $D$  est une matrice de rotation définie par un angle  $\theta$  et  $A$  est une matrice diagonale de termes diagonaux  $a$  et  $1/a$ .



### 3.5.6 Simulation d'un échantillon gaussien

On cherche à simuler un échantillon issu d'un vecteur aléatoire de moyenne  $\boldsymbol{\mu}$  et de variance  $\Sigma$ .

**Vecteur aléatoire  $\mathcal{N}(0, I)$**  Si les variables aléatoires  $X_1, \dots, X_p$  sont indépendantes, normales de moyenne 0 et de variance 1, alors le vecteur  $\mathbf{X} = (X_1, \dots, X_p)'$  est normal,

de moyenne 0 et de variance  $I$ . En effet

$$\begin{aligned} f(\mathbf{x}) &= f(x_1) \dots f(x_p) && \text{(indépendance)} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_1^2\right) \dots \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x_p^2\right) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}(x_1^2 + \dots + x_p^2)\right) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\mathbf{x}'I\mathbf{x}\right). \end{aligned}$$

**Vecteur aléatoire  $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$**  Si la matrice  $\Sigma$  peut être décomposée en  $\Sigma = TT'$  où  $T$  est une matrice de dimension  $(p^p)$ , alors la transformation  $\mathbf{Y} = T\mathbf{X} + \boldsymbol{\mu}$  transforme le vecteur aléatoire  $\mathbf{X}$  de loi  $\mathcal{N}(0, I)$  en un vecteur aléatoire de loi  $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ . Il suffit d'appliquer la propriété (3) : on obtient une loi normale de moyenne  $T \cdot 0 + \boldsymbol{\mu} = \boldsymbol{\mu}$  et de variance  $TTT' = TT' = \Sigma$ .

**Détermination de  $T$**  Tout d'abord, remarquons que s'il existe une solution, elle n'est pas unique : en effet, si  $T$  vérifie  $TT' = \Sigma$  et si  $U$  est une matrice orthogonale, alors  $S = TU$  vérifie la même propriété.

La décomposition de Cholesky  $\Sigma = TT'$  où  $T$  est une matrice triangulaire inférieure est une solution. On peut aussi facilement montrer que la matrice

$$T = \sqrt{\lambda}D\sqrt{A}D'$$

définie à partir de la décomposition précédente  $\Sigma = \lambda DAD'$  est une autre solution.

**Application en matlab** Matlab dispose de plusieurs fonctions permettant de simuler des échantillons de variables aléatoires. Par exemple, les commandes `binornd`, `normrnd`, `poissrnd`, `unidrnd` et `unifrnd` permettent de simuler des échantillons de variables aléatoires binomiales, normales, de Poisson, uniformes discrètes et uniformes. Ainsi, la commande `x=normrnd(mu, sigma2, n, 1)` permet de simuler un échantillon de taille  $n$  issu d'une loi normale  $\mathcal{N}(\boldsymbol{\mu}, \sigma^2)$ . La figure suivante donne ainsi la densité et l'histogramme d'un échantillon correspondant à une loi normale de moyenne 3 et de variance 10.



FIG. 3.1 – densité et histogramme d'une loi normale

En outre, les différentes simulations sont indépendantes : la commande `normrnd(0, 1, 100, 2)` permet ainsi de simuler un échantillon de taille 100 d'un vecteur aléatoire du plan de loi normale de moyenne  $(0, 0)'$  et de matrice de variance  $I$  et la séquence

```
mu=(mu1, mu2)';
D=[cos(theta), -sin(theta); sin(theta), cos(theta)];
T=sqrt(lambda)*D*diag([sqrt(a), 1/sqrt(a)])*D';
X=normrnd(0, 1, n, 2)*T'+ ones(n, 1)* mu';
```

permet de simuler un échantillon de taille  $n$  d'un vecteur aléatoire du plan de loi normale de moyenne  $\boldsymbol{\mu}$  et de matrice de variance  $\Sigma = \lambda DAD'$ .

La figure suivante donne ainsi la densité, les ellipses d'isodensité et l'histogramme d'un échantillon correspondant à une loi normale de moyenne  $(0, 7)'$  et de variance définie par les valeurs  $\lambda = 1$ ,  $\theta = \pi/4$  et  $a = 5$ .

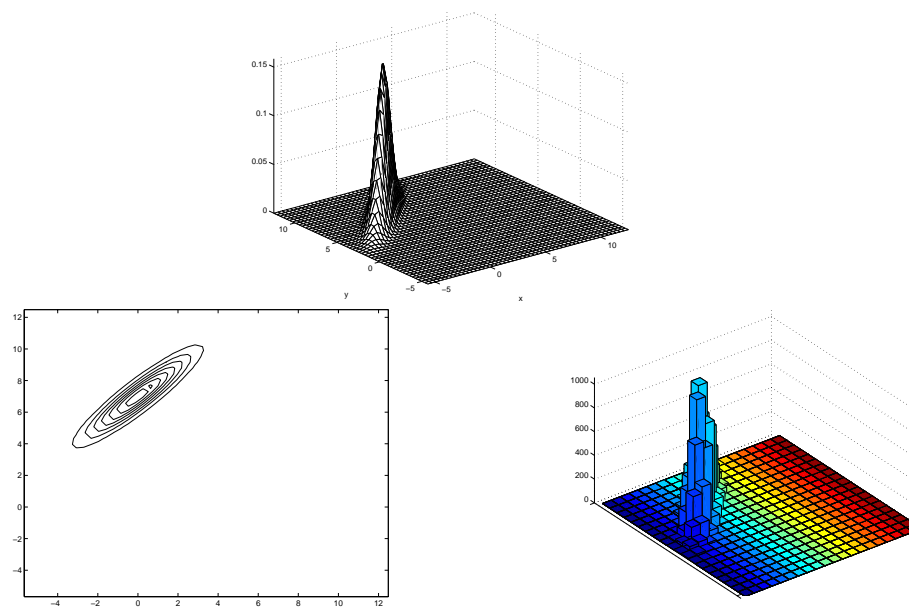


FIG. 3.2 – Densité, ellipses d'équidensité et histogramme d'un vecteur aléatoire gaussien du plan



## Chapitre 4

# Distance et représentation euclidienne

### 4.1 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  $\Omega$ . 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.

#### 4.1.1 Types de proximités

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

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \quad d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y} \quad (\text{séparation})$$

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \quad (\text{symétrie})$$

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \Omega \quad d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \quad (\text{inégalité triangulaire})$$

En analyse des données, il n'est pas toujours nécessaire d'avoir toutes ces propriétés et les mesures suivantes sont souvent suffisantes.

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

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$$

$$\forall \mathbf{x} \in \Omega \quad d(\mathbf{x}, \mathbf{x}) = 0.$$

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

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \quad s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$$

$$\forall \mathbf{x}, \mathbf{y} \in \Omega, \mathbf{x} \neq \mathbf{y} \quad s(\mathbf{x}, \mathbf{x}) = s_{max} \quad \text{avec} \quad s_{max} \geq s(\mathbf{x}, \mathbf{y}).$$

Remarquons qu'il est facile de transformer un indice de similarité  $s$  en un indice de dissimilarité en posant  $d(\mathbf{x}, \mathbf{y}) = s_{max} - s(\mathbf{x}, \mathbf{y})$ .

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

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

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \quad d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y} \quad (\text{séparation})$$

$$\forall \mathbf{x}, \mathbf{y} \in \Omega \quad d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \quad (\text{symétrie})$$

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \Omega \quad d(\mathbf{x}, \mathbf{z}) \leq \max(d(\mathbf{x}, \mathbf{y}), d(\mathbf{y}, \mathbf{z})) \quad (\text{inégalité ultramétrique})$$

Il est facile de montrer que la propriété d'inégalité ultramétrique entraîne la propriété d'inégalité triangulaire. Une ultramétrique est donc une distance.

### 4.1.2 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 peuvent être obtenus à partir d'autres tableaux. À partir de variables quantitatives, il est possible d'utiliser toutes les distances définies sur  $\mathbb{R}^p$ .

- Distance euclidienne :  $d^2(\mathbf{x}, \mathbf{y}) = \sum_j (\mathbf{x}_j - \mathbf{y}_j)^2 = (\mathbf{x} - \mathbf{y})'I(\mathbf{x} - \mathbf{y})$
- Distance euclidienne pondérée :  $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})'D(\mathbf{x} - \mathbf{y})$
- distance de Mahalanobis :  $d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})'S^{-1}(\mathbf{x} - \mathbf{y})$  où  $S$  est la matrice de variance empirique
- distance « city-block » ou distance  $L^1$  :  $d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^p |x_j - y_j|$
- distance de Chebychev ou distance  $L^\infty$  :  $d(\mathbf{x}, \mathbf{y}) = \max_j |x_j - y_j|$ .
- distance de Minkowski  $L_p$  :  $d(\mathbf{x}, \mathbf{y}) = \left( \sum_{j=1}^p (x_j - y_j)^p \right)^{1/p}$  (il s'agit de la généralisation des distances précédentes :  $L_1$ =city-block,  $L_2$ =euclidienne,  $L_\infty$ =chebychev).

Une distance entre variables peut être définie à partir de la corrélation :  $d = 1 - r^2$ .

À partir de variables qualitatives nominales, il est possible d'utilisée la distance du  $\chi^2$  ou, plus simplement, la distance  $d = 1 -$  proportion de modalités communes. Cette dernière peut être généralisée en en utilisant une table de ressemblance entre modalités.

À partir de variables qualitatives ordinales, il est possible d'utiliser la distance euclidienne sur les rangs renormalisés entre 0 et 1.

Pour les variables binaires, si  $a, b, c$  et  $d$  représente le nombre de fois où les individus  $\mathbf{x}$  et  $\mathbf{y}$  ont répondu respectivement (1, 1), (1, 0), (0, 1) et (0, 0) aux variables binaires, alors toute une série de mesures de proximité ont été proposées, par exemple :

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

### 4.1.3 Transformation

Il existe de nombreux moyens de passer d'un type de proximités à un autre. Par exemple, la relation  $s_{ii'} = (r_{ii'} + r_{i'i})/2$  permet de symétriser une proximité qui ne l'était pas ; la relation  $d_{ii'}^2 = s_{ii} - 2s_{ii'} + s_{i'i'}$  (Mardia et al. (1979)) permet de transformer une mesure de similarité en distance euclidienne ; la transformation  $d_{ii'} + c$  où  $c$  est le maximum des valeurs  $|d_{ij} + d_{jk} - d_{ik}|$  permet de transformer une dissimilarité en distance.

### 4.1.4 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 toujours possible de définir une proximité cohérente avec les données et d'appliquer ce dernier type de méthodes.

## 4.2 Rappels de géométrie et de mécanique

Dans ce paragraphe, on se placera dans l'espace vectoriel  $\mathbb{R}^p$  muni d'une distance euclidienne  $d$  définie par la matrice définie symétrique positive  $M$ .

### 4.2.1 Nuage de points

Si  $\Omega$  est un ensemble fini de  $n$  points  $\mathbf{x}$  de  $\mathbb{R}^p$  auxquels sont associés les poids  $\mu_{\mathbf{x}}$ , l'ensemble  $\mathcal{N}(\Omega) = \{(\mathbf{x}, \mu_{\mathbf{x}})/\mathbf{x} \in \Omega\}$  est appelé *nuage de points* de  $\mathbb{R}^p$ . Son *centre de*

gravité est défini par

$$\mathbf{g} = \frac{1}{\mu} \sum_{\mathbf{x} \in \Omega} \mu_{\mathbf{x}} \mathbf{x}$$

où  $\mu$  est la somme des pondérations  $\sum_{\mathbf{x}} \mu_{\mathbf{x}}$ .

### 4.2.2 Inertie

L'inertie de  $\mathcal{N}(\Omega)$  par rapport à un point  $\mathbf{a}$  est définie par

$$I_{\mathbf{a}} = \sum_{\mathbf{x} \in \Omega} \mu_{\mathbf{x}} d^2(\mathbf{a}, \mathbf{x})$$

et l'inertie du nuage  $\mathcal{N}(\Omega)$  par rapport à une variété linéaire  $F$  par

$$I_F = \sum_{\mathbf{x} \in \Omega} \mu_{\mathbf{x}} d^2(\mathbf{x}, F).$$

L'inertie  $I_{\mathbf{g}}$  du nuage  $\Omega$  par rapport à son centre de gravité est appelée simplement *Inertie du nuage* et notée  $I$ .

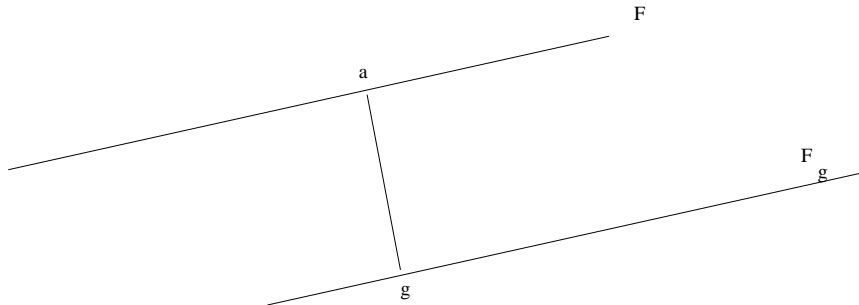
### 4.2.3 Théorèmes de Huygens

$$I_{\mathbf{a}} = I_{\mathbf{g}} + \mu d^2(\mathbf{a}, \mathbf{g}) \quad \forall \mathbf{a} \in \mathbb{R}^p$$

Le centre de gravité est donc le point d'inertie minimum.

$$I_F = I_{F_{\mathbf{g}}} + \mu d^2(\mathbf{a}, \mathbf{g}) \quad \forall \text{variété linéaire } F$$

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  $F_{\mathbf{g}}$ .

### 4.2.4 Nuage centré

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 la suite de ce paragraphe.

### 4.2.5 Inertie expliquée

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  $I$  se décompose suivant la relation  $I = I_F + I_{F^\perp}$ .

En outre, l'inertie  $I_{F^\perp}$ , 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  $F$ .

L'inertie expliquée se décompose alors de la manière suivante :

$$A = B \oplus C \quad \text{et} \quad B \perp C \Rightarrow I_{A^\perp} = I_{B^\perp} + I_{C^\perp}.$$

### 4.2.6 Expressions matricielles des inerties

Si on note  $D_p$  la matrice diagonale  $\text{diag}(\mu_1, \dots, \mu_n)$ , on peut exprimer l'inertie totale sous la forme suivante

$$I = \text{trace}(X'D_pXM)$$

et l'inertie portée par un axe

$$I_{\Delta_u^\perp} = u'MX'D_pXM u$$

où  $\Delta_u$  est un axe défini par le vecteur unitaire  $u$ .

La matrice  $X'D_pX$ , qui intervient dans ces 2 expressions, est appelée matrice d'inertie.

## 4.3 Représentation euclidienne des 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  $\mathbf{x}_i$  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 associée la pondération  $p_i = 1/n$  et à chaque variable la pondération  $q_j = 1$ . On notera  $D_p$  la matrice diagonale  $\text{diag}(p_1, \dots, p_n) = \frac{1}{n}I_n$  et  $M$  la matrice diagonale  $\text{diag}(q_1, \dots, q_j) = I_p$ . L'utilisation de pondérations générales  $p_i$  et  $q_j$  permettra, par exemple, d'étendre sans difficulté l'analyse en composantes principales à l'analyse des correspondances.

On peut alors définir le nuage

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

inclus dans  $\mathbb{R}^p$  muni de la métrique euclidienne  $M$ . Cette représentation généralise ce qui avait été utilisé lorsqu'il n'y avait que deux variables.

De façon symétrique, les variables peuvent être représentées par le nuage :

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

inclus dans  $\mathbb{R}^n$  muni de la métrique euclidienne  $D_p$  souvent appelée « métrique des poids ».

## 4.4 Interprétation statistique

Il est possible de montrer que certaines caractéristiques géométriques de ces deux nuages ont une interprétation statistique. Par exemple, le centre de gravité du nuage des individus a pour coordonnées les moyennes des  $p$  variables. Cette interprétation peut être poursuivie si des hypothèses supplémentaires sont ajoutées.

### 4.4.1 Tableau centré en colonne

On suppose souvent que le nuage des individus est centré, c'est-à-dire que son centre de gravité est à l'origine ; la moyenne de chaque variable est alors nulle. On dit 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.

Dans l'espace des variables, le produit scalaire et la norme s'expriment respectivement alors comme la covariance et la variance :

$$\begin{aligned} - \text{Cov}(\mathbf{x}^j, \mathbf{x}^{j'}) &= \sum_{i=1}^n p_i \cdot x_i^j \cdot x_i^{j'} = (\mathbf{x}^j)' D_p \mathbf{x}^{j'} = \langle \mathbf{x}^j, \mathbf{x}^{j'} \rangle_{D_p} \\ - \text{Var}(\mathbf{x}^j) &= \|\mathbf{x}^j\|_{D_p}^2 \end{aligned}$$

Dans ce cas, la matrice de variance  $S$  s'écrit  $X'D_pX$  et n'est autre que la matrice d'inertie. On a donc  $I(\mathcal{N}(\Omega)) = \text{trace}(SM)$ . Lorsque  $M = I$ , l'inertie du nuage des individus est la trace de la matrice de variance.

On a aussi

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

Cette corrélation s'interprète comme le cosinus de l'angle des deux vecteurs  $\mathbf{x}^j$  et  $\mathbf{x}^{j'}$  dans l'espace des variables  $\mathbb{R}^n$ . On peut aussi exprimer les relations précédentes en terme de métrique :

- $d^2(\mathbf{x}^j, \mathbf{x}^{j'}) = \text{Var}(\mathbf{x}^j) + \text{Var}(\mathbf{x}^{j'}) - 2\text{Cov}(\mathbf{x}^j, \mathbf{x}^{j'})$
- $d^2(0, \mathbf{x}^j) = V(\mathbf{x}^j)$ .

#### 4.4.2 Variables normées

Si on représente les variables normées, la variance de chaque variable étant alors par définition égale à 1, on obtient

- $\text{Cor}(\mathbf{x}^j, \mathbf{x}^{j'}) = \langle \mathbf{x}^j, \mathbf{x}^{j'} \rangle_{D_p}$
- $d^2(0, \mathbf{x}^j) = \|\mathbf{x}^j\|_{D_p}^2 = 1$
- $d^2(\mathbf{x}^j, \mathbf{x}^{j'}) = 2(1 - \text{Cor}(\mathbf{x}^j, \mathbf{x}^{j'}))$ .

Dans l'espace des variables  $\mathbb{R}^n$ , les variables normées sont donc toutes situées sur une hypersphère de centre 0 et de rayon 1. Cette hypersphère est appelée « cercle des corrélations ».

Si le tableau  $X$  est centré-réduit, les variables initiales sont normées et les matrices de covariance et de corrélation sont les mêmes. L'expression  $X'D_p X$  représente donc aussi dans ce cas la matrice de corrélation associée aux  $p$  variables initiales.



## Chapitre 5

# L'analyse en composantes principales

### 5.1 Introduction

Les *méthodes factorielles* ont pour objectif de visualiser, et plus généralement, de traiter 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 est une démarche habituelle (moyenne à l'école, QI, répartition des hommes politiques sur un axe droite-gauche) qui consiste à résumer plusieurs variables par une seule. Il y a mieux à faire. C'est ce qu'ont proposé les psychologues américains Spearman, Burt et Thurstone en caractérisant les résultats à 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.

## 5.2 Axes principaux d'inertie

### 5.2.1 Formulation mathématique

L'objectif est d'obtenir une représentation fidèle du nuage  $N(\Omega)$  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  $N(\Omega)$  et leurs projections. Les espaces de représentation choisis sont les espaces affines (droite, plan,...). La formulation mathématique de l'ACP est alors la suivante : *Trouver le sous-espace affine  $E_k$  de dimension  $k$  ( $k < p$ ) tel que  $I_{E_k}$ , l'inertie du nuage  $N(\Omega)$  par rapport à  $E_k$ , soit minimum.*

Rappelons que l'on a

$$I_{E_k} = \frac{1}{n} \sum_i d^2(\mathbf{x}_i, E_k).$$

En utilisant le théorème de Huygens (2), on peut en déduire que l'espace  $E_k$  minimisant  $I_{E_k}$  contient nécessairement le centre de gravité du nuage  $N(\Omega)$ , 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  $I$  se décompose en une somme  $I_{E_k} + I_{E_k^\perp}$  où  $I_{E_k^\perp}$  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  $I_{E_k^\perp}$  par  $E_k$  soit maximum.*

### 5.2.2 Résultats préalables

**Théorème 5.1 ([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  $\Delta v$  l'axe correspondant et  $G$  l'espace supplémentaire  $M$ -orthogonal à  $\Delta v$  dans  $F_k$  (on a donc  $F_k = G \oplus \Delta v$ ).

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

$$I_{F_k^\perp} = I_{G^\perp} + I_{\Delta v^\perp} \text{ car } G \perp \Delta v$$

$$I_{H^\perp} = I_{E_{k-1}^\perp} + I_{\Delta v^\perp} \text{ car } E_{k-1} \perp \Delta v.$$

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

$$I_{E_{k-1}^\perp} \geq I_{G^\perp} \Rightarrow I_{H^\perp} \geq I_{F_k^\perp}$$

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

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

**Théorème 5.2** *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  $\Delta v$   $M$ -orthogonal à  $F$  et maximisant  $I_{\Delta v^\perp}$ .*

*Preuve :* On a une décomposition  $E = F \oplus \Delta v$  avec  $\Delta v \perp F$ . On a donc  $I_{E^\perp} = I_{F^\perp} + I_{\Delta v^\perp}$ . Maximiser  $I_{E^\perp}$  est donc équivalent à maximiser  $I_{\Delta v^\perp}$ .  $\square$

### 5.2.3 Résolution du problème

On suppose dans la suite que les vecteurs  $\mathbf{u}_j$  sont unitaires. En outre, on sait que pour tout vecteur unitaire  $\mathbf{u}$ ,  $I_{\Delta \mathbf{u}^\perp}$  est égale à  $\langle \mathbf{u}, S M \mathbf{u} \rangle_M (= \mathbf{u}^T M S M \mathbf{u})$ .

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  $\Delta u_1$  maximisant l'inertie  $I_{\Delta u_1^\perp} = \langle \mathbf{u}_1, SM\mathbf{u}_1 \rangle_M$ , on note  $E_1 = \Delta u_1$  ;
- rechercher un axe  $\Delta u_2$ ,  $M$ -orthogonal à  $E_1$  maximisant l'inertie  $I_{\Delta u_2^\perp} = \langle \mathbf{u}_2, SM\mathbf{u}_2 \rangle_M$ , on note  $E_2 = E_1 \oplus \Delta u_2$  ;
- ...
- Rechercher un axe  $\Delta u_k$ ,  $M$ -orthogonal à  $E_{k-1}$  maximisant l'inertie  $I_{\Delta u_k^\perp} = \langle \mathbf{u}_k, SM\mathbf{u}_k \rangle_M$ , on note  $E_k = E_{k-1} \oplus \Delta u_k$ .

En posant  $B = SM$  et  $Q = M$ , le théorème de décomposition d'une matrice énoncé dans l'annexe B fournit une réponse à notre problème : les vecteurs propres de la matrice  $SM$  ordonnés suivant les valeurs propres décroissantes fournissent les axes  $\Delta u_1, \dots, \Delta u_k$ , appelés *axes factoriels* ou encore *axes principaux d'inertie* et les inerties  $I_{\Delta u_k^\perp}$  portées ou expliquées par ces axes sont égales aux valeurs propres  $\lambda_k$ .

Les espaces  $E_k$  sont donc solutions du problème et on obtient du même coup toutes les solutions pour les dimensions inférieures à  $k$ .

### 5.2.4 Résultats pratiques

Si  $\mathbf{u}_1, \dots, \mathbf{u}_p$  sont les vecteurs propres normés ordonnés suivant les valeurs propres décroissantes de la matrice  $SM$ , la solution pour les différentes valeurs de  $k$  est la suivante :

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

On a en outre  $I_{\Delta \mathbf{u}_k^\perp} = \lambda_k$ .

### 5.2.5 Inerties expliquées

**Proposition 5.3**  $I_{E_k^\perp} = \lambda_1 + \dots + \lambda_k$

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

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

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

**Remarque** : En prenant  $k = p$ , on retrouve  $I = \text{trace}(S)$ . De plus, si  $r$  est le rang de la matrice  $\mathbf{x}$  ( $r \leq \min(p, n)$ ), on a

$$\lambda_1, \dots, \lambda_r > 0 \text{ et } \lambda_{r+1}, \dots, \lambda_p = 0.$$

et par suite

$$I_{E_r^\perp} = I.$$

Finalement, le nuage est exactement dans le sous-espace vectoriel  $E_r$  engendré par les  $r$  premiers axes factoriels.

### 5.2.6 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  $E_\alpha$  :

- % d'inertie expliquée par  $E_1 = \frac{\lambda_1}{\sum_{\alpha=1}^p \lambda_\alpha} * 100 = \frac{\lambda_1}{\text{trace}(VM)} * 100$  ;
- % d'inertie expliquée par  $E_2 = \frac{\lambda_1 + \lambda_2}{\sum_{\alpha=1}^p \lambda_\alpha} * 100 = \frac{\lambda_1 + \lambda_2}{\text{trace}(VM)} * 100$  ;
- ... ;
- % d'inertie expliquée par  $E_k = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\sum_{\alpha=1}^p \lambda_\alpha} * 100 = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\text{trace}(VM)} * 100$ .

## 5.3 Composantes principales

### 5.3.1 Définition

Rappelons que le problème de départ était d'obtenir une représentation du nuage  $N(\Omega)$  dans des espaces de dimension réduite. On connaît 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  $c_{1\alpha}, \dots, c_{n\alpha}$  les  $n$  coordonnées ainsi obtenues avec l'axe  $\alpha$ ,  $\mathbf{c}_\alpha$  le vecteur  $(c_{1\alpha}, \dots, c_{n\alpha})'$ , appelé  $\alpha^e$  composante principale et  $C$  la matrice obtenue en rangeant en colonne les vecteurs  $\mathbf{c}_\alpha$ . On peut alors obtenir la projection du nuage  $N(\Omega)$  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  $\alpha > r$ ,  $\lambda_\alpha$ , et donc  $I_{\Delta\mathbf{u}_\alpha^\perp}$ , est nul ; ce qui entraîne que  $\mathbf{c}_\alpha = 0$ . Enfin, en exprimant l'inertie expliquée par l'axe  $\alpha$  dans la relation  $\lambda_\alpha = I_{\Delta\mathbf{u}_\alpha^\perp}$ , on obtient la relation

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

### 5.3.2 Calcul des composantes principales

**Proposition 5.4** *les composantes principales vérifient*

$$\mathbf{c}_\alpha = XM\mathbf{u}_\alpha$$

qui s'exprime matriciellement

$$C = XMU.$$

*Preuve :* La nouvelle base est orthonormée : il suffit donc de projeter les  $\mathbf{x}_i$  sur les vecteurs de base :

$$\begin{aligned} c_{i\alpha} &= \langle \mathbf{x}_i, \mathbf{u}_\alpha \rangle_M = \mathbf{x}_i' M \mathbf{u}_\alpha \\ \mathbf{c}_\alpha &= XM\mathbf{u}_\alpha \\ C &= XMU \end{aligned}$$

où  $U$  est la matrice des vecteurs propres normalisés. □

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

$$\begin{aligned} \mathbf{x}_i &= U\mathbf{c}_i \\ U'M\mathbf{x}_i &= U'MU\mathbf{c}_i = \mathbf{c}_i \text{ car } U'MU = I \\ \mathbf{c}_i' &= \mathbf{x}_i' MU \end{aligned}$$

et donc

$$C = XMU$$

□

### 5.3.3 Composantes principales : nouvelles variables

Une composante principale associée à chaque individu  $\mathbf{x}_i$  de  $\Omega$  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 5.5** *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  $M\mathbf{u}_\alpha$ .  $\square$

**Proposition 5.6** *Les composantes principales  $\mathbf{c}_\alpha$  sont centrées, de variance  $\lambda_\alpha$  et non corrélées 2 à 2.*

*Preuve* : Une combinaison linéaire de variables centrées est centrée.

$$\begin{aligned} \text{Cov}(\mathbf{c}_\alpha, \mathbf{c}_\beta) &= \langle \mathbf{c}_\alpha, \mathbf{c}_\beta \rangle = \mathbf{c}'_\alpha D_p \mathbf{c}_\beta \\ &= \mathbf{u}'_\alpha M X' D_p M X \mathbf{u}_\beta = \mathbf{u}'_\alpha M (X' D_p X) M \mathbf{u}_\beta \\ &= \mathbf{u}'_\alpha M S M \mathbf{u}_\beta = \mathbf{u}'_\alpha M (S \mathbf{u}_\beta) = \lambda_\beta \mathbf{u}'_\alpha M \mathbf{u}_\beta = \lambda_\beta \langle \mathbf{u}_\alpha, \mathbf{u}_\beta \rangle. \end{aligned}$$

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

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 normés des  $p$  variables centrées initiales, non corrélées deux à deux et de variance maximum.

**Proposition 5.7** *Les composantes principales  $\mathbf{c}_\alpha$  sont vecteurs propres de la matrice  $WD_p$  associées aux valeurs propres  $\lambda_\alpha$  où*

$$W = XMX'$$

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

*Preuve* : les  $\mathbf{u}_\alpha$  et les  $\lambda_\alpha$  étant vecteurs propres de la matrice  $SM$ , on peut en déduire

$$\begin{aligned} SM\mathbf{u}_\alpha &= \lambda_\alpha \mathbf{u}_\alpha, \\ (X' D_p X)M\mathbf{u}_\alpha &= \lambda_\alpha \mathbf{u}_\alpha. \end{aligned}$$

On multiplie à gauche par  $XM$  :

$$\begin{aligned} (XM)X' D_p X M \mathbf{u}_\alpha &= \lambda_\alpha (XM)\mathbf{u}_\alpha, \\ XMX' D_p (XM\mathbf{u}_\alpha) &= \lambda_\alpha (XM\mathbf{u}_\alpha), \\ XMX' D_p \mathbf{c}_\alpha &= \lambda_\alpha \mathbf{c}_\alpha. \end{aligned}$$

$\square$

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

## 5.4 Formule de reconstitution

À partir de la décomposition des vecteurs  $\mathbf{x}_i$  sur la base des vecteurs propres

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

on peut facilement en déduire l'égalité matricielle

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

qui représente une décomposition de la matrice  $X$  en une somme de matrices de rang 1. La dernière relation montre que l'on peut « reconstituer » le tableau initial avec les composantes principales et les axes principaux. Cette relation est appelée formule de reconstitution. Si on se limite aux  $k$  ( $k < r$ ) premiers termes, on obtient une approximation du tableau initial :

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

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

Matriciellement, on obtient

$$\begin{aligned} C &= XMU \\ CU' &= XMUU' = X \\ X &= CU'. \end{aligned}$$

## 5.5 Qualité de la représentation

### 5.5.1 Qualité globale

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

$$\frac{\lambda_1 + \dots + \lambda_k}{\text{trace}(S)} 100.$$

### 5.5.2 Contribution relative d'un axe à un individu

Sachant que l'inertie totale du nuage  $N(\Omega)$  est  $\frac{1}{n} \sum_{i=1}^p \|\mathbf{x}_i\|^2$ , la quantité  $p_i \|\mathbf{x}_i\|^2$  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 initial  $p_i \|\mathbf{x}_i\|^2$  qu'apportait l'individu  $i$ , conservée par l'axe  $\alpha$ . Le rapport de ces deux quantités est appelée *contribution relative* du  $\alpha^e$  axe factoriel à l'individu  $i$  et elle est notée  $COR(i, \alpha)$  :

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

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 :

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

### 5.5.3 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  $\lambda_\alpha$ , 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  $\alpha$ , notée  $CTR(i, \alpha)$  : c'est la part d'inertie du  $\alpha^e$  axe pris en compte (ou expliquée) par l'individu  $i$ . Nous avons :

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

## 5.6 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 donner une « interprétation » aux axes. Le calcul des ces coordonnées vérifie donc

$$\text{cor}(\alpha, j) = \text{Cov} \left( \frac{1}{\sigma_j} \mathbf{x}_j, \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{c}_\alpha \right) = \frac{1}{\sigma_j} \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{x}_j' D_p \mathbf{c}_\alpha.$$

## 5.7 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*.

### 5.7.1 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  $\bar{x}_j$  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  $\mathbf{y}_s = (y_{s1}, \dots, y_{sp})'$  en  $\mathbf{x}_s = (y_{s1} - \bar{x}_1, \dots, y_{sp} - \bar{x}_p)'$  et de le projeter sur les axes  $\mathbf{u}_\alpha$ . Les coordonnées sont ainsi obtenues avec la formule  $\langle \mathbf{x}_s, \mathbf{u}_\alpha \rangle = \mathbf{x}'_s M \mathbf{u}_\alpha$ .

### 5.7.2 Variable supplémentaire

Cette fois, la transformation précédente, devient une projection dans  $\mathbb{R}^n$ . Il faut centrer la nouvelle variable  $\mathbf{y}_s = (y_{1s}, \dots, y_{ns})'$ . 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  $\bar{y} = \frac{1}{n} \sum_i y_{is}$  et  $s = \sqrt{\frac{1}{n} \sum_i (y_{is} - \bar{y})^2}$  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  $\mathbf{x}_s = \frac{1}{s}(y_{1s} - \bar{y}, \dots, y_{ns} - \bar{y})'$  sur les axes  $\mathbf{v}_\alpha = \frac{\mathbf{c}_\alpha}{\sqrt{\lambda_\alpha}}$ . Les coordonnées sont ainsi obtenues par la relation suivante

$$\langle \mathbf{x}_s, \mathbf{v}_\alpha \rangle_{D_p} = \mathbf{x}'_s D_p \frac{\mathbf{c}_\alpha}{\sqrt{\lambda_\alpha}} = \frac{1}{n\sqrt{\lambda_\alpha}} \mathbf{x}'_s \mathbf{c}_\alpha.$$

obtenues avec la formule  $\langle \mathbf{x}_s, \mathbf{v}_\alpha \rangle_{D_p} = \mathbf{x}'_s D_p \frac{\mathbf{c}_\alpha}{\sqrt{\lambda_\alpha}}$ .

### 5.7.3 Importance pratique des éléments supplémentaires

Les éléments supplémentaires permettent, par exemple, la représentation d'individus prenant des valeurs très différentes des autres (valeurs aberrantes) 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.

## 5.8 Un exemple d'ACP

### 5.8.1 Les données

Il s'agit du tableau de notes décrits dans le chapitre 2. 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	6.0	6.0	5.0	5.5	8.0
aline	8.0	8.0	8.0	8.0	9.0
annie	6.0	7.0	11.0	9.5	11.0
monique	14.5	14.5	15.5	15.0	8.0
didier	14.0	14.0	12.0	12.5	10.0
andré	11.0	10.0	5.5	7.0	13.0
pierre	5.5	7.0	14.0	11.5	10.0
brigitte	13.0	12.5	8.5	9.5	12.0
evelyne	9.0	9.5	12.5	12.0	18.0

### 5.8.2 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	dess
jean	-3.67	-3.83	-5.22	-4.55	-3
aline	-1.67	-1.83	-2.22	-2.05	-2
annie	-3.67	-2.83	0.78	-0.55	0
monique	4.83	4.67	5.28	4.95	-3
didier	4.33	4.17	1.78	2.45	-1
andré	1.33	0.17	-4.72	-3.05	2
pierre	-4.17	-2.83	3.78	1.45	-1
brigitte	3.33	2.67	-1.72	-0.55	1
evelyne	-0.67	-0.33	2.28	1.95	7

### 5.8.3 Matrice de variance

$$S = X'D_p X = \frac{1}{9} X'X$$

	math	scie	fran	lati	dess
math	11.389				
scie	9.917	8.944			
fran	2.657	4.120	12.062		
lati	4.824	5.481	9.293	7.914	
dess	0.111	0.056	0.389	0.667	8.667

### 5.8.4 Axes principaux d'inertie

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

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

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

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

### 5.8.5 Qualité de la représentation

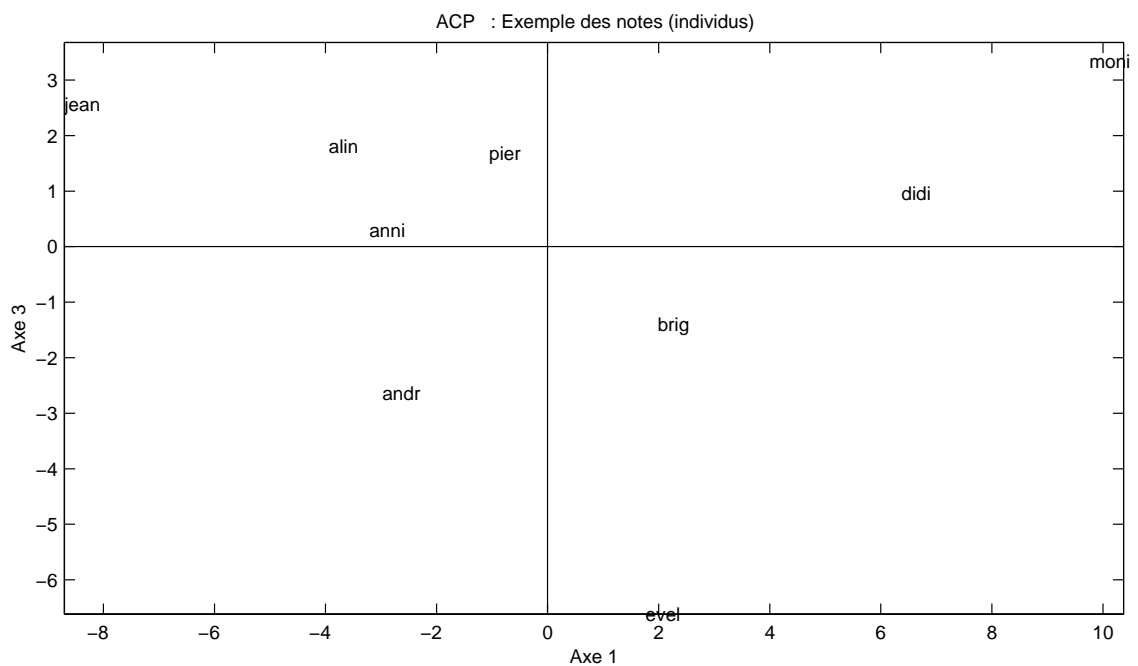
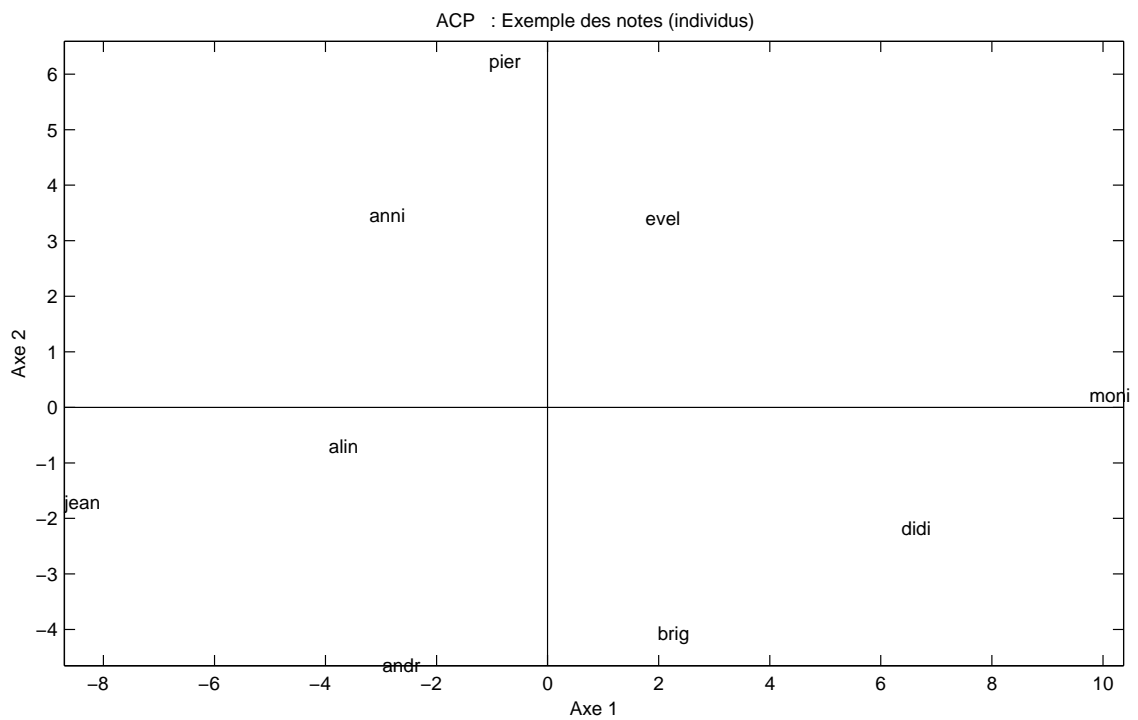
Rappelons que les inerties du nuage projeté sur les 5 axes sont égales aux valeurs propres. L'inertie du nuage est égale à  $\text{trace}(SM) = \text{trace}(S)$ , c'est-à-dire aussi à la somme des valeurs propres, ici 48.975. Les pourcentages d'inertie expliquée par chaque axe sont donc de 57.69, 24.65, 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.

### 5.8.6 Composantes principales

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

	1	2	3	4	5
jean	-8.70	-1.70	2.55	0.16	0.11
aline	-3.94	-0.72	1.81	0.09	-0.04
annie	-3.22	3.47	0.29	-0.18	-0.02
monique	9.75	0.22	3.54	0.18	-0.09
didier	6.37	-2.17	0.96	-0.07	0.18
andré	-2.97	-4.65	-2.64	0.02	-0.16
pierre	-1.05	6.21	1.67	-0.11	-0.04
brigitte	1.99	-4.07	-1.41	-0.25	0.00
evelyne	1.77	3.40	-6.62	0.15	0.07

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



## 5.8.7 Contributions relatives des axes aux individus

	1	2	3	4	5
jean	0.89	0.03	0.08	0.00	0.00
aline	0.80	0.03	0.17	0.00	0.00
annie	0.46	0.53	0.00	0.00	0.00
monique	0.89	0.00	0.11	0.00	0.00
didier	0.88	0.10	0.02	0.00	0.00
andré	0.24	0.58	0.19	0.00	0.00
pierre	0.03	0.91	0.07	0.00	0.00
brigitte	0.17	0.74	0.09	0.00	0.00
evelyne	0.05	0.20	0.75	0.00	0.00

## 5.8.8 Contributions relatives des individus aux axes

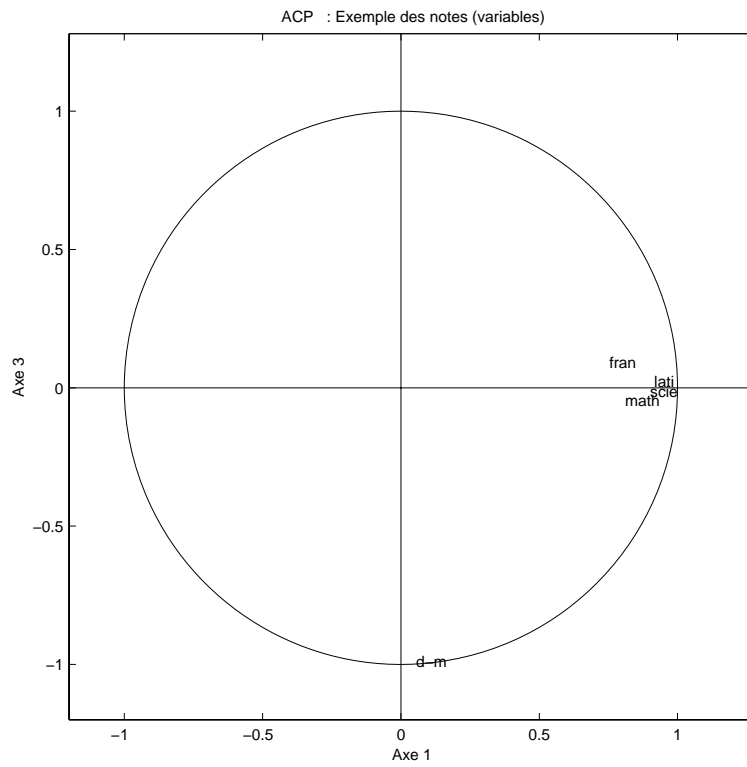
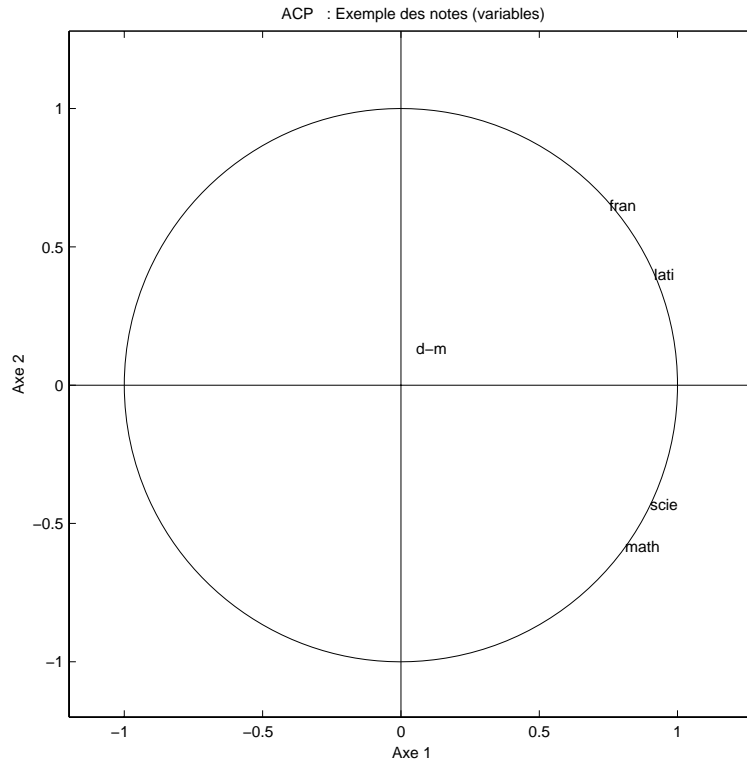
	1	2	3	4	5
jean	0.30	0.03	0.09	0.11	0.15
aline	0.06	0.00	0.04	0.04	0.02
annie	0.04	0.11	0.00	0.15	0.00
monique	0.37	0.00	0.14	0.15	0.11
didier	0.15	0.04	0.02	0.03	0.40
andré	0.03	0.20	0.09	0.00	0.25
pierre	0.00	0.36	0.04	0.07	0.02
brigitte	0.02	0.15	0.03	0.30	0.00
evelyne	0.01	0.11	0.56	0.14	0.04

5.8.9 Analyse dans  $\mathbb{R}^n$ 

Calcul des corrélations  $cor(\alpha, j)$ .

	F1	F2	F3	F4	F 5
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

### 6.1 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*) 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. 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 (1997) 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,...). Enfin, rappelons que les tableaux de proximités ont été étudiés dans le chapitre 4.

### 6.2 Le problème

Étant donnée une matrice de dissimilarités  $\Delta$  sur  $n$  individus, l'objectif est de déterminer une représentation  $X$  de dimension  $p$  donnée telle que la distance euclidienne associée  $D(X)$  à cette représentation soit proche de la dissimilarité initiale  $\Delta$ .

Avant d'étudier les méthodes proposées pour résoudre ce problème, le paragraphe suivant est consacré à l'étude de quelques propriétés des distances euclidiennes.

### 6.3 Distances euclidiennes

On notera dans la suite  $Q_n = I_n - \frac{1}{n}U_n$  la matrice associée à la projection orthogonale sur le vecteur  $(1, \dots, 1)'$  dans l'espace  $\mathbb{R}^n$ . Cette matrice correspond à l'opérateur de centrage en colonne d'une matrice de dimension  $(n, p)$ .

#### 6.3.1 Équivalence entre distances euclidiennes et produits scalaires

On considère  $X$  la matrice de données centrée associée à un ensemble de  $n$  points de  $\mathbb{R}^p$ ; on note  $D^2 = D^2(X)$  la matrice des distances au carré entre les  $n$  points,  $W = W(X)$  la

matrice  $XX'$  des produits scalaires et  $\mathbf{h} = (\|\mathbf{x}_1\|^2, \dots, \|\mathbf{x}_n\|^2)'$  = diag( $W$ ).

**Proposition 6.1**  $D^2$  est une fonction de  $W$

*Preuve* : Il suffit de développer :

$$d_{ii'}^2 = \langle \mathbf{x}_i - \mathbf{x}_{i'}, \mathbf{x}_i - \mathbf{x}_{i'} \rangle = \|\mathbf{x}_i\|^2 - 2w_{ii'} + \|\mathbf{x}_{i'}\|^2$$

que l'on peut noter matriciellement

$$D^2 = \mathbf{h}\mathbb{1}'_n - 2W + \mathbb{1}_n\mathbf{h}' = \text{diag}(W)\mathbb{1}'_n - 2W + \mathbb{1}_n\text{diag}(W)'. \quad \square$$

**Proposition 6.2**  $W$  est une fonction de  $D^2$ .

*Preuve* : Partant de la relation précédente, on obtient

$$-\frac{1}{2}Q_n D^2 Q_n = -\frac{1}{2}Q_n \mathbf{h}\mathbb{1}'_n Q_n + Q_n X X' Q_n - \frac{1}{2}Q_n \mathbb{1}_n \mathbf{h}' Q_n.$$

Sachant que  $Q\mathbb{1}_n = 0$  et que  $QX = X$  puisque la matrice  $X$  est déjà centrée, on en déduit

$$-\frac{1}{2}Q_n D^2 Q_n = X X' = W. \quad \square$$

L'expression  $Q_n D^2 Q_n$  est quelquefois appelée double-centrage.

**Conséquence** Les deux propositions précédentes montrent qu'il existe une bijection entre la matrice des distances et la matrice des produits scalaires. On notera  $D^2 = \varphi(W)$  et  $W = \varphi^{-1}(D^2)$  les fonctions associées.

### 6.3.2 Matrice de distances euclidiennes

On dira qu'une matrice de dissimilarités  $\Delta$  est une matrice de distances euclidiennes si et seulement s'il existe une représentation  $X$  des  $n$  individus dans un espace  $\mathbb{R}^p$  telle que la distance euclidienne associée soit la distance  $\Delta$ , c'est-à-dire, en notant  $D(X)$  la matrice des distances euclidiennes associées à un tableau  $X$ , si et seulement si il existe  $X$  tel que  $D(X) = \Delta$ .

### 6.3.3 CNS pour qu'une matrice de dissimilarités soit euclidienne

**Proposition 6.3** Une matrice de dissimilarités  $\Delta$  est euclidienne si et seulement si  $W = -\frac{1}{2}Q_n \Delta^2 Q_n$  est une matrice semi-définie positive (SDP). En outre, la représentation associée est contenue dans un espace de dimension  $p \leq n - 1$ .

*Preuve* : La proposition directe est immédiate car dans ce cas  $W$  est une matrice de produits scalaires et est donc semi-définie positive.

Réciproque : soient  $V$  la matrice des vecteurs propres normés (au sens de  $\frac{1}{n}I$ ) de  $\frac{1}{n}W$  et  $L$  la matrice diagonale formée des valeurs propres. Ces matrices vérifient  $\frac{1}{n}WV = VL$ ,  $VV' = \frac{1}{n}I$  et  $VLV' = W$ .  $W$  étant SDP, les valeurs propres de  $\frac{1}{n}W$  sont positives et on peut définir  $X = V\sqrt{L}$ .

La matrice  $VL = \frac{1}{n}WV$  est de la forme  $Q_n A$ ; elle est donc centrée en colonne. On peut alors en déduire que  $V$  et donc  $X$  sont aussi centrées en colonne. Le matrice des produits scalaires associée à la représentation  $X$  s'écrit donc  $XX'$  et on a

$$XX' = (V\sqrt{L})(V\sqrt{L})' = VLV' = W.$$

c'est-à-dire

$$W(X) = \varphi^{-1}(\Delta^2)$$

et donc

$$D^2(X) = \Delta^2$$

ce qui montre que  $\Delta$  est une matrice de distance euclidienne.

Le rang de la matrice  $W$  est inférieur ou égal à  $n - 1$ . La dimension de la représentation associée  $X$ , égale au nombre de valeurs propres non nulles de  $WD^{-p}$ , est donc inférieure ou égale à  $n - 1$ .  $\square$

## 6.4 Analyse factorielle d'un tableau de distances

Cette méthode est historiquement la première technique de positionnement multidimensionnel et a été développée par Togerson (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*.

### 6.4.1 $W = -\frac{1}{2}Q_n\Delta^2Q_n$ est SDP

Nous venons de voir que dans ce cas, il existait une représentation euclidienne  $X$  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  $\lambda_\alpha$  de  $\frac{1}{n}W$ , il suffit dans la construction de  $X$  d'ordonner les vecteurs propres pour que la matrice des composantes principales  $C$  ne soit rien d'autre que  $X$ .

En pratique, il faudra donc :

1. calculer la matrice  $W = \varphi^{-1}(\Delta^2) = -\frac{1}{2}Q_n\Delta^2Q_n$ ,
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  $W = -\frac{1}{2}Q_n\Delta^2Q_n$  est SDP se fait a posteriori ; il faut et il suffit que toutes les valeurs propres sont positives ou nulles.

### 6.4.2 $W = -\frac{1}{2}Q_nD^2Q_n$ 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. Par exemple, On peut en additionnant une certaine constante à la dissimilarité initiale la transformer en une distance. On peut alors appliquer sur cette distance l'AFTD. En pratique cette méthode ne donne pas toujours de très bons résultats.

### 6.4.3 L'AFTD dans R

L'AFTD peut être réalisée à l'aide de la fonction `cmdscale` qui permet, en particulier, l'introduction de la constante signalée dans le paragraphe précédent.

### 6.4.4 Un exemple

Nous avons appliqué l'AFTD aux données d'Ekman portant sur la couleur. Comme il s'agit d'un tableau de similarité, la première chose à faire est de transformer en un tableau de dissimilarités. Voilà le tableau obtenu :

	L434	L445	L465	L472	L490	L504	L537	L555	L584	L600	L610	L628	L651	L674
L434	0.00	0.14	0.6	0.6	0.82	0.94	0.9	1.0	0.98	0.93	0.9	0.88	0.87	0.84
L445	0.14	0.00	0.5	0.6	0.78	0.91	0.9	0.9	0.98	0.96	0.9	0.89	0.87	0.86
L465	0.58	0.50	0.0	0.2	0.53	0.83	0.9	0.9	0.98	0.99	1.0	0.99	0.95	0.97
L472	0.58	0.56	0.2	0.0	0.46	0.75	0.9	0.9	0.98	0.99	1.0	0.99	0.98	0.96
L490	0.82	0.78	0.5	0.5	0.00	0.39	0.7	0.7	0.93	0.98	1.0	0.99	0.98	1.00
L504	0.94	0.91	0.8	0.8	0.39	0.00	0.4	0.6	0.86	0.92	1.0	0.98	0.98	0.99
L537	0.93	0.93	0.9	0.9	0.69	0.38	0.0	0.3	0.78	0.86	0.9	0.98	0.98	1.00
L555	0.96	0.93	0.9	0.9	0.74	0.55	0.3	0.0	0.67	0.81	1.0	0.97	0.98	0.98
L584	0.98	0.98	1.0	1.0	0.93	0.86	0.8	0.7	0.00	0.42	0.6	0.73	0.80	0.77
L600	0.93	0.96	1.0	1.0	0.98	0.92	0.9	0.8	0.42	0.00	0.3	0.50	0.59	0.72
L610	0.91	0.93	1.0	1.0	0.98	0.98	0.9	1.0	0.63	0.26	0.0	0.24	0.38	0.45
L628	0.88	0.89	1.0	1.0	0.99	0.98	1.0	1.0	0.73	0.50	0.2	0.00	0.15	0.32
L651	0.87	0.87	0.9	1.0	0.98	0.98	1.0	1.0	0.80	0.59	0.4	0.15	0.00	0.24
L674	0.84	0.86	1.0	1.0	1.00	0.99	1.0	1.0	0.77	0.72	0.4	0.32	0.24	0.00

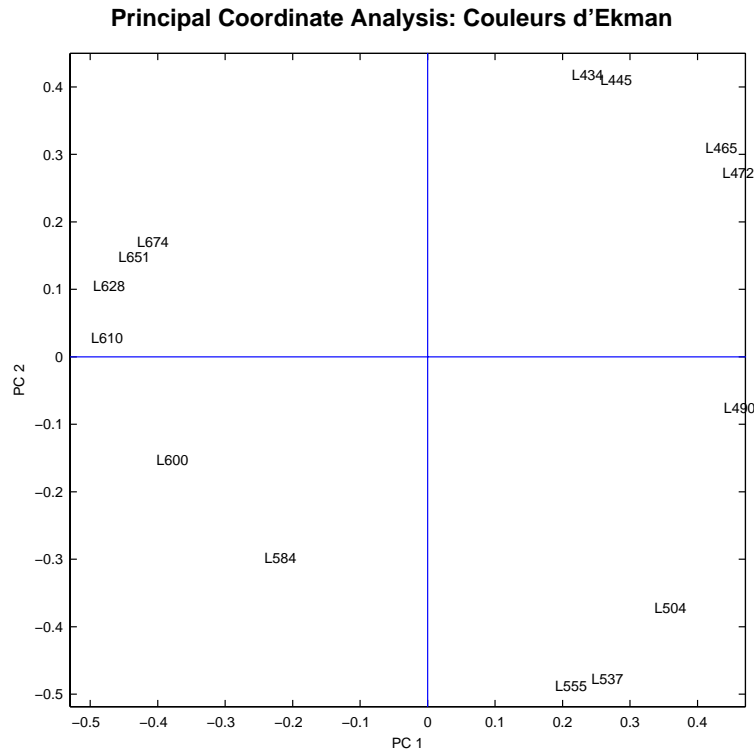
Voici les valeurs propres fournies :

$\alpha$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\lambda_\alpha$	0.14	0.09	0.03	0.027	0.011	0.007	0.003	0.0023	0.0013	0.0003	0.0001	0	-0.002	-0.003
%	45.2	29.7	10.1	8.5	3.6	2.3	0.95	0.72	0.42	0.09	0.03	0	-0.61	-1.1
	45.2	74.9	85.0	93.5	97.1	99.5	100.4	101.1	101.6	101.7	101.7	101.7	101.1	100

les coordonnées des 5 premières composantes principales :

Ind.	PC1	PC2	PC3	PC4	PC5
L434	0.21	0.42	0.23	-0.18	-0.09
L445	0.26	0.41	0.20	-0.17	-0.04
L465	0.41	0.31	-0.01	0.18	0.07
L472	0.44	0.27	-0.07	0.20	0.06
L490	0.44	-0.08	-0.27	0.16	-0.03
L504	0.34	-0.37	-0.23	-0.06	-0.09
L537	0.24	-0.48	0.03	-0.21	-0.06
L555	0.19	-0.49	0.15	-0.15	0.11
L584	-0.24	-0.30	0.28	0.22	0.17
L600	-0.40	-0.15	0.19	0.21	-0.16
L610	-0.50	0.03	-0.03	0.11	-0.13
L628	-0.50	0.10	-0.13	-0.05	-0.04
L651	-0.46	0.15	-0.19	-0.11	0.02
L674	-0.43	0.17	-0.16	-0.16	0.20

et la représentation obtenue dans le premier plan :



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. Par ailleurs, la représentation dans le premier plan fournit une très bonne représentation des données.

## 6.5 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'imposent 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  $\Delta$  et  $D$  appelée *Stress*.

### 6.5.1 Fonctions Stress

Plusieurs fonctions ont été proposées :

$$Stress_1(X) = \frac{\sum_{i < i'} (\delta_{ii'} - d_{ii'})^2}{\sum_{i < i'} d_{ii'}^2}$$

$$Stress_2(X) = \frac{\sum_{i < i'} w_{ii'} (\delta_{ii'} - d_{ii'})^2}{\sum_{i < i'} w_{ii'} d_{ii'}^2}$$

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

$$Stress_3(X) = \frac{1}{\sum_{i < i'} \delta_{ii'}} \sum_{i < i'} \frac{(\delta_{ii'} - d_{ii'})^2}{\delta_{ii'}}$$

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.

### 6.5.2 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 (la plus efficace à ce jour) ;
- méthode de Newton.

### 6.5.3 Projection de Sammon

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

### 6.5.4 Remarques

- Le choix du nombre de dimension 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  $\delta_{ij}, d_{ij}$ .

## 6.6 Méthodes non métriques ou ordinales

### 6.6.1 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 :

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

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.

### 6.6.2 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.

## 6.7 Quelques remarques

### 6.7.1 Dissimilarités initiales

La dissimilarité initiale  $\Delta$  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.

### 6.7.2 Autres méthodes

On peut citer quelques méthodes voisines : par exemple, l'analyse procustéenne permet de comparer deux tableaux de dissimilarités et si il y a plus de deux tableaux de

dissimilarités, les méthode 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

### 7.1 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<sup>e</sup> 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 maladie ; 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 (e.g. : « Tous les hommes sont mortels ») ; On peut aussi regrouper en classe les objets qui posséderont des caractéristiques « proches ». Cette démarche est dite *polythétique*. Généralement, on utilise pour cela la notion de mesure de proximité (distance, dissimilarité). 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.

## 7.2 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*.

### 7.2.1 Partition

**Définition 7.1**  $\Omega$  étant un ensemble fini, un ensemble  $P = (P_1, P_2, \dots, P_g)$  de parties non vides de  $\Omega$  est une partition si :

1.  $\forall k \neq \ell, P_k \cap P_\ell = \emptyset,$
2.  $\cup_k P_k = \Omega.$

Dans un ensemble  $\Omega$  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  $\mathbf{x}_i$  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.

### 7.2.2 La hiérarchie indicée

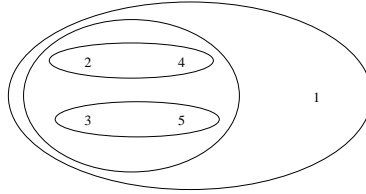
**Définition 7.2**  $\Omega$  étant un ensemble fini, un ensemble  $H$  de parties non vides de  $\Omega$  est une hiérarchie sur  $\Omega$  si

- $\Omega \in H;$
- $\forall \mathbf{x} \in \Omega \quad \{\mathbf{x}\} \in H;$
- $\forall h, h' \in H \quad h \cap h' = \emptyset \text{ ou } h \subset h' \text{ ou } h' \subset h.$

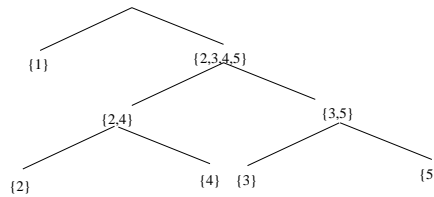
**Exemple**  $\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{2, 4\}, \{3, 5\}, \{2, 3, 4, 5\}, \{1, 2, 3, 4, 5\}\}$  est une hiérarchie définie sur  $\Omega = \{1, 2, 3, 4, 5\}$ .

**Représentations d'une hiérarchie** Deux types de représentation peuvent être utilisés

- la représentation ensembliste



- la représentation par arbre



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

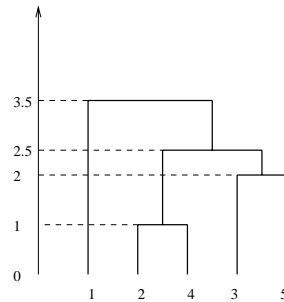
**Définition 7.3** On appelle indice sur une hiérarchie  $H$  une fonction  $i$  de  $H$  dans  $\mathbb{R}^+$  vérifiant les propriétés :

- $h \subset h'$  et  $h \neq h' \Rightarrow i(h) < i(h')$  ( $i$  est une fonction strictement croissante),
- $\forall i \in \Omega \quad i(\{i\}) = 0$ .

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

**Exemple** On peut associer aux classes  $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{2, 4\}, \{3, 5\}, \{2, 3, 4, 5\}, \{1, 2, 3, 4, 5\}$  de la hiérarchie précédente les valeurs  $0, 0, 0, 0, 1, 2, 2.5, 3.5$ .

**Représentation à l'aide d'un arbre hiérarchique : dendrogramme**



### 7.2.3 Partition et hiérarchie

Si  $P = (P_1, P_2, \dots, P_g)$  est une partition de  $\Omega$ , l'ensemble  $H$  formé des classes  $P_k$  de  $P$ , des singletons de  $\Omega$  et de l'ensemble  $\Omega$  lui-même forme une hiérarchie. Remarquons qu'inversement, il est possible d'associer à chaque niveau d'une hiérarchie indicée une partition. Une hiérarchie indicée correspond donc à un ensemble de partitions emboîtées.

### 7.2.4 Aspects combinatoires

Le nombre de hiérarchies et de partitions qu'il est possible de définir sur un ensemble  $\Omega$  devient vite énorme lorsque le cardinal de  $\Omega$  augmente. Par exemple, le nombre de

partitions d'un ensemble de  $n$  éléments en  $g$  classes est donné par la formule suivante :

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

Lorsque  $n$  et  $g$  deviennent grand, on a  $S(n, g) \approx \frac{g^n}{g!}$ .

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

1	2	3	4	5	6	7	8
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

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

## 7.3 Liens avec la notion d'ultramétrie

### 7.3.1 Recherche de partitions associées à une mesure de dissimilarité

On dispose d'une mesure de dissimilarité  $d$  sur l'ensemble  $\Omega$  et à toute valeur réelle  $\alpha \geq 0$  on associe la relation binaire sur  $\Omega$  :

$$xV_\alpha y \Leftrightarrow d(x, y) \leq \alpha.$$

**Problème** Peut-on trouver une partition de  $\Omega$  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  $V_\alpha$  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

$$\forall \alpha \geq 0 \quad xV_\alpha y \text{ et } yV_\alpha z \quad \Rightarrow \quad xV_\alpha z$$

ce qui donne

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

Cette propriété n'est pas vraie en général pour une mesure de dissimilarité mais nous allons montrer maintenant que cette propriété est équivalente à l'inégalité ultramétrique. Si  $d$  est une ultramétrie alors il est clair que l'équation 7.1 est vraie.

Réciproquement : supposons que  $d$  est une mesure de dissimilarité vérifiant 7.1. Soient 3 points  $x, y$  et  $z$ . Posons  $\alpha = \max(d(x, y), d(y, z))$ . On a :

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

On peut donc en déduire

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

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

### 7.3.2 Ultramétrie associée à une hiérarchie indicée : fonction $\varphi$

( $H, i$ ) étant une hiérarchie indicée sur  $\Omega$ , on peut lui associer la fonction

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

de la façon suivante :

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

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

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

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

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

**Proposition 7.4**  $\delta$  est une ultramétrie sur  $\Omega$ .

### 7.3.3 Hiérarchie indicée associée à une ultramétrie : fonction $\psi$

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

#### Construction d'une hiérarchie à partir de l'ultramétrie $\delta$

$D_\delta$  étant l'ensemble des valeurs prises par l'ultramétrie  $\delta$  sur  $\Omega$ , on définit l'ensemble  $H$  comme l'ensemble de toutes les classes d'équivalence des relations  $V_\alpha$  lorsque  $\alpha$  parcourt  $D_\delta$ . On peut alors montrer la proposition suivante :

**Proposition 7.5**  $H$  est une hiérarchie sur  $\Omega$ .

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

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

La proposition suivante peut alors être facilement montrée

**Proposition 7.6** La fonction  $i$  définit un indice sur la hiérarchie  $H$ .

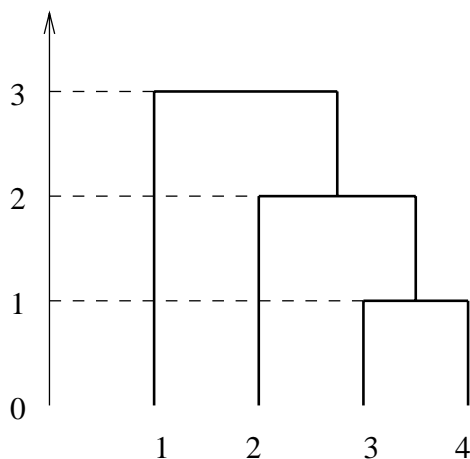
### 7.3.4 Équivalence entre hiérarchie indicée et ultramétrie

**Proposition 7.7** Les fonctions  $\varphi$  et  $\psi$  sont réciproques. C'est-à-dire :

$$\psi \circ \varphi(H, i) = (H, i) \quad \text{et} \quad \varphi \circ \psi(\delta) = \delta.$$

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

### 7.3.5 Exemples



On part de la hiérarchie  $(H_1, i_1)$  :

La distance ultramétrique  $\delta_1 = \varphi(h_1, i_1)$  obtenue est alors la suivante

	1	2	3	4
1	0			
2	3	0		
3	3	2	0	
4	3	2	1	0

On peut maintenant appliquer la fonction  $\psi$  à l'ultramétrique  $\delta_1$  pour obtenir une hiérarchie indicée  $(H_2, i_2)$  :

On a  $D_\delta = \{0, 1, 2, 3\}$ . Les classes d'équivalence des 4 relations  $R_\alpha$  sont :

$$R_0 : \{1\}, \{2\}, \{3\}, \{4\};$$

$$R_1 : \{1\}, \{2\}, \{3, 4\};$$

$$R_2 : \{1\}, \{2, 3, 4\};$$

$$R_3 : \{1, 2, 3, 4\}.$$

La hiérarchie  $H_2$  est donc

$$\{\{1\}, \{2\}, \{3\}, \{4\}, \{3, 4\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}$$

et les indices associés aux parties de cette hiérarchie sont respectivement

$$0, 0, 0, 0, 1, 2, 3.$$

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

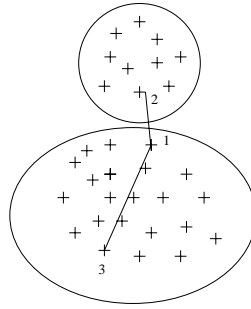
## 7.4 Objectifs de la classification

### 7.4.1 Difficultés de caractériser les objectifs

Rappelons que l'objectif de la classification automatique est l'organisation en classes homogènes des éléments d'un ensemble  $\Omega$ . Pour définir cette notion de classes homogènes, on utilise le plus souvent une mesure de similarité (ou de dissimilarité) sur  $\Omega$ . Par exemple, si  $d$  est une mesure de dissimilarité sur  $\Omega$ , on peut caractériser cette homogénéité en imposant aux classes de la partition recherchée de vérifier la propriété suivante :

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

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 7.4.1, 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.

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

### 7.4.2 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  $\Omega$  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  $\Omega$  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

$$S = \frac{1}{n}(X - \mathbb{1}_n \bar{x})'(X - \mathbb{1}_n \bar{x}) = \frac{1}{n} \sum_i (\mathbf{x}_i - \bar{x})(\mathbf{x}_i - \bar{x})'$$

et l'inertie  $I = \frac{1}{n} \sum_i d^2(\mathbf{x}_i, \bar{x})$  vérifie  $I = \text{trace}(S)$ .

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

$$S_W = \frac{1}{n} \sum_k n_k S_k$$

où  $S_k$  est la matrice de variance de chaque classe ( $S_k = \frac{1}{n}(X_k - \mathbb{1}_n \bar{x}_k)'(X_k - \mathbb{1}_n \bar{x}_k)$ ) et l'inertie intra-classe

$$I_W = \sum_k I(P_k)$$

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

$$I_W = \text{trace}(S_W).$$

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.

### 7.4.3 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 chapitre suivant.

## 7.5 La classification ascendante hiérarchique

L'objectif est de construire une hiérarchie indicée d'un ensemble  $\Omega$  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  $\Omega$  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  $\Omega$  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  $\Omega$  lui-même. C'est cette procédure, beaucoup plus utilisée que la précédente, que nous étudions dans ce chapitre.

### 7.5.1 L'algorithme

#### Construction de la hiérarchie

$\Omega$  étant l'ensemble à classifier et  $d$  une mesure de dissimilarité sur cet ensemble  $\Omega$ , on définit, à partir de  $d$ , une « distance »  $D$  entre les parties de  $\Omega$ . Cette distance est en réalité une mesure de dissimilarité qui ne vérifie pas nécessairement toutes les propriétés d'une distance sur l'ensemble des parties de  $\Omega$  définie à partir de la mesure de dissimilarité  $d$  sur  $\Omega$ . Nous verrons plus loin plusieurs façons de définir  $D$  à partir de  $d$ . En général, cette fonction  $D$  est appelé *critère d'agrégation*.

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.

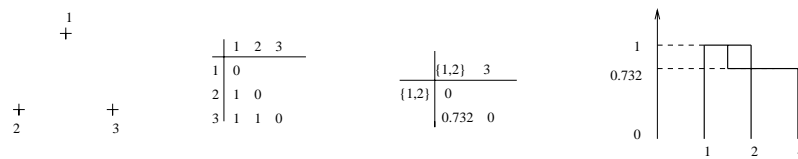
Il est facile de montrer que l'ensemble des classes définies au cours de cet algorithme forme une hiérarchie sur  $W$ .

**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 séparent 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 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. Dans l'exemple suivant, la classe  $A \cup B$  qui a le même indice que la classe  $A \cup B \cup C$  peut 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 nécessiterons donc de prévoir cette opération de filtrage.

**7.5.2 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 link)

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

- critère du lien maximum (ou saut maximum)

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

- critère de la distance moyenne

$$D(A, B) = \frac{\sum_{i \in A} \sum_{i' \in B} d(i, i')}{n_A \cdot n_B}$$

où  $n_E$  représente le cardinal de l'ensemble  $E$ . Remarquons que les hiérarchies fournies par les deux premiers critères ne dépendent que de l'ordre des distances.

### 7.5.3 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 essentielle pour la mise en place pratique de l'algorithme de CAH qui, sans cette relation, serait prohibitif en temps calcul. Ces relations appelées généralement formules de récurrence de Lance et Williams, sont les suivantes :

**Proposition 7.8** *Pour les trois critères d'agrégation du saut minimum, du saut maximum et de la moyenne, on peut calculer la distance  $D$  entre les deux classes  $A$  et  $B \cup C$  uniquement à partir des distances  $D$  entre  $A$  et  $B$  et entre  $A$  et  $C$  :*

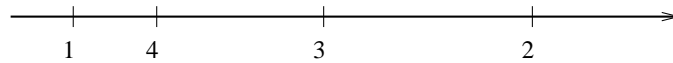
$$D_{\min} : \quad D(A, B \cup C) = \min\{D(A, B), D(A, C)\};$$

$$D_{\max} : \quad D(A, B \cup C) = \max\{D(A, B), D(A, C)\};$$

$$D_{\text{moy}} : \quad D(A, B \cup C) = \frac{n_B \cdot D(A, B) + n_C \cdot D(A, C)}{n_B + n_C}.$$

### 7.5.4 Un exemple

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



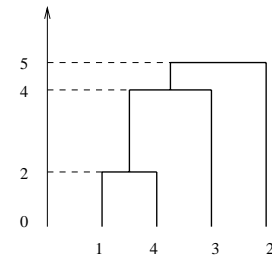
On prend comme mesure de dissimilarité entre ces points la distance euclidienne habituelle et on effectue la CAH suivant les trois critères d'agrégation :

– Critère  $D_{\min}$

	1	2	3	4
1	0			
2	11	0		
3	6	5	0	
4	2	9	4	0

	{1,4}	2	3
{1,4}	0		
2	9	0	
3	4	5	0

	{1,4,3}	2
{1,4,3}	0	
2	5	0

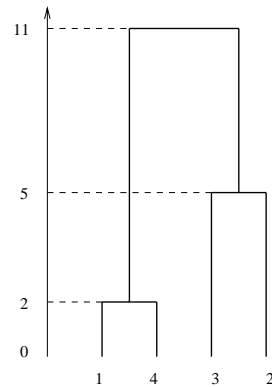


– Critère  $D_{\max}$

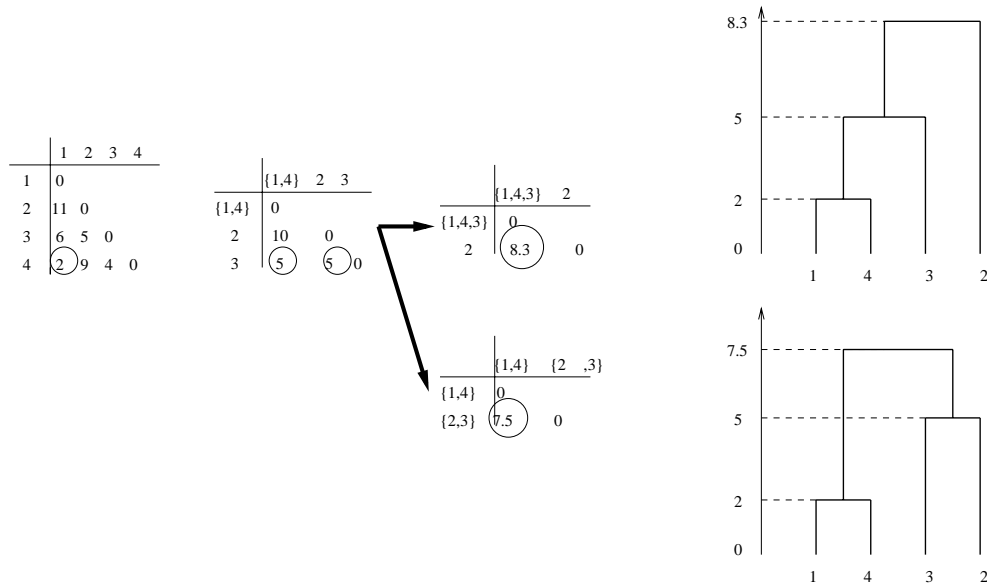
	1	2	3	4
1	0			
2	11	0		
3	6	5	0	
4	2	9	4	0

	{1,4}	2	3
{1,4}	0		
2	11	0	
3	6	5	0

	{1,4}	{2,3}
{1,4}	0	
{2,3}	11	0



– Critère  $D_{\text{moy}}$



Remarquons que dans le dernier cas, on peut obtenir deux solutions différentes suivant que l'on choisit de regrouper les classes {1,4} et {3} ou les classes {2} et {4}.

### 7.5.5 Méthode de Ward

Lorsque l'ensemble  $\Omega$  à classifier correspond à un nuage de points dans  $\mathbb{R}^p$  muni des pondérations  $\frac{1}{n}$  et 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 :

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

### 7.5.6 Propriétés d'optimalité

Nous avons vu que la notion de hiérarchie indicée est équivalente à la notion d'ultramétrie. La CAH transforme donc une mesure de dissimilarité  $d$  initiale en une nouvelle mesure de dissimilarité  $\delta$  qui possède la propriété d'être ultramétrique. La classification hiérarchique pourrait alors être posée en ces termes :

*trouver l'ultramétrie  $\delta$  la plus proche de  $d$ .*

Il reste à munir l'espace des mesures de dissimilarité sur  $\Omega$  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

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

Malheureusement, il s'agit 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.

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

Soit  $\delta_m$  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  $\delta_m$  est encore une ultramétrique. On l'appelle ultramétrique sous-dominante.

**Proposition 7.9** *Quelque soit  $\Delta$  la distance entre deux mesures de dissimilarité, l'ultramétrique sous-dominante est l'ultramétrique la plus proche, au sens de  $\Delta$ , d'une mesure de dissimilarité  $d$  parmi toutes les ultramétriques inférieures à  $d$ .*

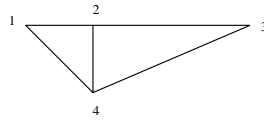
**Proposition 7.10** *L'ultramétrique associée à la hiérarchie indicée obtenue par la CAH avec le critère du saut minimum est l'ultramétrique sous-dominante.*

Cette propriété entraîne le corollaire suivant :

**Corollaire 7.11** *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  $\Omega$ . 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 :

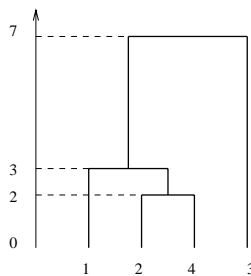


– construction de l'ultramétrique sous-dominante :

	1	2	3	4
1	0			
2	3	0		
3	10	7	0	
4	3.6	2	7.3	0

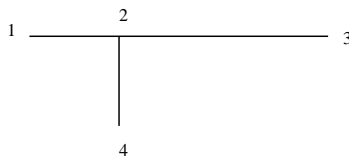
	1	{2,4}	3
1	0		
{2,4}	2	0	
3	10	7	0

	{1,2,4}	3
{1,2,4}	0	
3	7	0



	1	2	3	4
1	0			
2	3	0		
3	7	7	0	
4	3	2	7	0

En ne retenant du graphe complet initial que les 3 arêtes ayant participé à l'algorithme, c'est-à-dire l'arête (2,4) de longueur 2, l'arête (1,2) de longueur 3 et l'arête (2,3) 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'ultramétrie est supérieure à la dissimilarité  $d$ . Malheureusement, les propriétés de l'ultramétrie fournie par la CAH ne sont pas aussi intéressantes que celles de l'ultramétrie 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  $\Omega$ .

Remarquons que l'on peut construire de façon parallèle à l'ultramétrie sous-dominante, qui a été définie comme l'enveloppe supérieure des ultramétries inférieures, l'enveloppe inférieure des ultramétries supérieures à  $d$ . Malheureusement cette enveloppe n'est pas nécessairement une ultramétrie. L'exemple de la figure 7.1 en est un contre-exemple.

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

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

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

FIG. 7.1 – Distance  $d$  et ultramétries  $\delta_1$  et  $\delta_2$

On peut vérifier que  $\delta_1$  et  $\delta_2$  sont deux ultramétries supérieures à la distance  $d$  définie sur les 3 points a,b, c et que l'enveloppe inférieure de ces deux ultramétries est tout simplement  $d$ . Par conséquent, l'enveloppe inférieure de toutes les ultramétries supérieures à  $d$  est nécessairement  $d$  qui n'est pas ultramétrie.

### Hiérarchie de la moyenne

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

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

### Méthode de Ward

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

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

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 :

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

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.

### 7.5.7 Utilisation des méthodes

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

## 7.6 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.

### 7.6.1 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) lorsque l'ensemble à classifier est mesuré par  $p$  variables. Dans tout ce paragraphe, l'ensemble  $\Omega$  à classifier correspond à un ensemble de  $n$  individus mesurés par  $p$  variables quantitatives. 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

L'algorithme des centres-mobiles peut se définir ainsi :

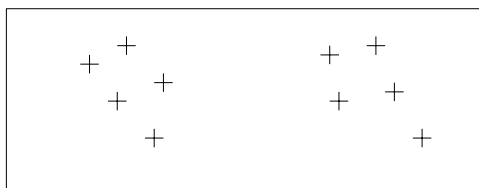
1. Tirage au hasard de  $g$  points de  $\Omega$  qui forment les centres initiaux des  $g$  classes.
2. Tant que non convergence
  - (a) construction de la partition suivante en affectant chaque point de  $\Omega$  à la classe dont il est le plus près du centre (en cas d'égalité, l'affectation se fait 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  $L = (\lambda_1, \dots, \lambda_K)$  représente un  $K$ -uplet de  $\mathbb{R}^p$  et  $P = (P_1, \dots, P_K)$  une partition de  $\Omega$  en  $K$  classes, la suite construite par l'algorithme peut être notée sous la forme :

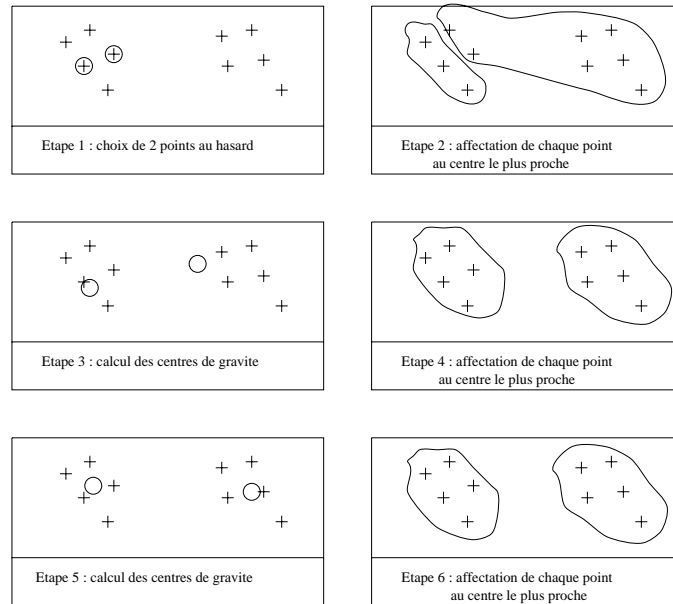
$$L^0 \rightarrow P^1 \rightarrow L^1 \rightarrow P^2 \rightarrow L^2 \rightarrow \dots \rightarrow P^n \rightarrow L^n \rightarrow \dots$$

#### Exemple

Les données sont constituées d'un ensemble  $\Omega$  de 10 points du plan.



L'algorithme des centres-mobiles peut alors se résumer de la façon suivante :



La poursuite de cet algorithme ne changera plus les résultats : l'algorithme a convergé. Remarquons que la classification obtenue correspond effectivement à la structure en deux classes observable 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 :

$$C(P, L) = \sum_k I(P_k, \lambda_k) = \frac{1}{n} \sum_k \sum_{\mathbf{x} \in P_k} d^2(\mathbf{x}, \lambda_k)$$

où  $P = (P_1, P_2, \dots, P_g)$  et  $L = (\lambda_1, \dots, \lambda_g)$ .

### 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 7.12

$$\begin{aligned} C(P^{n+1}, L^n) &\leq C(P^n, L^n) \\ C(P^{n+1}, L^{n+1}) &\leq C(P^{n+1}, L^n) \end{aligned}$$

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

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

où  $k(\mathbf{x})$  est le numéro de la classe à laquelle appartient  $\mathbf{x}$  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  $\Omega$  au meilleur centre, la relation 7.12 est vraie.

Le critère  $C(P, L)$  s'écrit aussi :

$$C(P, L) = \sum_k I(P_k, \lambda_k).$$

$L^{n+1}$  est par définition de l'algorithme des centres mobiles formé des  $g$  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

$$I(P_k^{n+1}, \lambda_k^{n+1}) \leq I(P_k^{n+1}, \lambda_k^n).$$

L'inéquation 7.12 est donc démontrée.

□

**Corollaire 7.13** *La suite numérique  $C(P^n, L^n)$  est une suite stationnaire.*

*Preuve :* Les deux inégalités de la propriété 7.12 entraîne la décroissance 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 7.14** *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 :

$$\dots, P, L, P', L', P, L, \dots, P, L, P', L', \dots$$

avec

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

*Preuve :*  $C(P^n, L^n)$  est stationnaire, c'est-à-dire :

$$\exists N \text{ t.q. } \forall n > N \quad C(P^n, L^n) = C(P^{n+1}, L^{n+1}).$$

Cette relation entraîne d'après la propriété 7.12

$$\exists N \text{ t.q. } \forall n > N \quad C(P^n, L^n) = C(P^{n+1}, L^n) = C(P^{n+1}, L^{n+1}).$$

Sachant que pour tout  $k$  on a nécessairement  $I(P_k^{n+1}, \lambda_k^n) \geq I(P_k^{n+1}, \lambda_k^{n+1})$  (propriété du centre de gravité), l'égalité précédente entraîne les égalités  $I(P_k^{n+1}, \lambda_k^n) \geq I(P_k^{n+1}, \lambda_k^{n+1})$  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  $\lambda_k^{n+1} = \lambda_k^n$  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

$$P^{n+1} = P^n.$$

□

### 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  $P^n$  :

$$C(P^n, L^n) = \sum_k I(P_k^n, \lambda_k^n) = \sum_k I(P_k^n)$$

puisque  $\lambda_k^n$  est le centre de gravité de la classe  $P_k^n$ . Et en conséquence

$$C(P^n, L^n) = I_W(P^n).$$

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  $g$  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 comme, 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  $g$  prototypes sont tirés au hasard parmi les  $n$  points.
  2. A l'itération  $q$ , un individu  $\mathbf{x}_i$  est choisi au hasard.
    - Détermination du prototype le plus proche de  $\mathbf{x}_i$  :

$$\lambda_k^q = \min_j \|\mathbf{x}_i - \lambda_j^q\|.$$

L'individu est affecté à la classe  $k$ .

- Modification du prototype  $\lambda_k^q$  :

$$\lambda_k^{q+1} = \frac{\mathbf{x}_i + n_k^q \cdot \lambda_j^q}{n_k^q + 1},$$

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.

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

**7.6.2 Généralisation : la méthode des nuées dynamiques**

L'idée de base consiste à remplacer les *centres*  $\lambda$  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,
- $D : \Omega \times L \rightarrow \mathbb{R}^+$ , une mesure de ressemblance entre éléments de  $\Omega$  et de  $\mathbb{L}$ .

L'objectif est alors de trouver la partition en  $g$  classes ( $g$  fixé a priori) de  $\Omega$  minimisant le critère

$$C(P, L) = \sum_k \sum_{\mathbf{x} \in P_k} D(\mathbf{x}, \lambda_k)$$

où  $P = (P_1, \dots, P_g)$  et  $L = (\lambda_1, \dots, \lambda_g)$  avec  $\lambda_k \in \mathbb{L}$ .

Pour ceci, on utilise l'algorithme d'optimisation alternée associée décrit dans le paragraphe suivant.

### Algorithme

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

$$L^0 \rightarrow P^1 \rightarrow L^1 \rightarrow P^2 \rightarrow L^2 \rightarrow \dots \rightarrow P^n \rightarrow L^n \rightarrow \dots$$

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

1.  $P^{n+1}$  est obtenu en minimisant  $C(\cdot, L^n)$
2.  $L^{n+1}$  est obtenu 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  $\Omega$  est inclus dans  $\mathbb{R}^p$ ,  $\mathbb{L}$  est l'espace  $\mathbb{R}^p$  et  $D(\mathbf{x}, \lambda) = d^2(\mathbf{x}, \lambda)$  où  $d$  est la distance euclidienne, on retrouve alors simplement la méthode des centres-mobiles.

### Tableau de dissimilarités

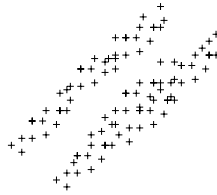
On suppose cette fois que l'on ne connaît sur  $\Omega$  qu'une mesure de dissimilarité  $d$ . On peut alors proposer la situation suivante :  $\mathbb{L} = \Omega$  et  $D(\mathbf{x}, \lambda) = d(\mathbf{x}, \lambda)$ .

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  $D(\mathbf{x}, (\mathbf{a}, d)) = d(\mathbf{x}, \mathbf{a})$ .

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, ...

### 7.6.3 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'un 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 comme 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.

## Chapitre 8

# Modèles probabilistes en classification

### 8.1 Introduction

Rappelons que l'objectif de la classification automatique est la recherche de classes « homogènes ». Nous avons vu que cet objectif conduisait à des algorithmes souvent conçus d'un point de vue heuristique et utilisant des critères métriques. Ainsi, les deux algorithmes de classification sans doute les plus utilisés, l'algorithme des centres-mobiles (ou *k-means*) pour la recherche de partitions et l'algorithme de classification ascendante hiérarchique de Ward pour la recherche de hiérarchies utilisent tous deux l'inertie intraclasse d'une partition, c'est-à-dire la somme des inerties de chaque classe. La difficulté de cette approche est la justification du choix de la métrique et du critère utilisés. Pour mettre en place de telles solutions, il est donc nécessaire de choisir d'une part, une métrique mesurant la dissimilarité entre les objets de l'ensemble à classifier et d'autre part, un critère défini à partir de cette métrique mesurant de degré de cohésion et de séparation des classes.

Tout ceci a conduit depuis quelques années à une évolution de l'approche algorithmique, heuristique et géométrique vers une approche plus statistique qui utilise des *modèles probabilistes de classification* pour formaliser l'idée intuitive de la notion de classe naturelle. Cette approche permet d'analyser de manière précise et de donner une interprétation statistique à certains critères métriques dont les différentes variantes n'étaient pas toujours bien claires (comme par exemple les critères d'inertie  $\text{trace}(S_W)$  et  $|S_W|$  définis à partir de la matrice de variance intraclasse  $S_W$ ) et permet en outre de proposer de nouvelles variantes répondant à des hypothèses précises. Elle fournit aussi un cadre formel pour proposer des solutions à des problèmes difficiles comme la détermination du nombre de classes ou encore la validation de la structure de classification obtenue.

Remarquons enfin que, très souvent l'ensemble à classifier n'est qu'un échantillon d'une population beaucoup plus grande et que les conclusions obtenues à partir de la classification de cet échantillon sont souvent étendues à toute la population. Dans ce cas, la classification n'a pas de sens sans un recours à un modèle probabiliste permettant de justifier cette inférence.

### 8.2 Approches probabilistes de la classification

L'hypothèse faite dans toute approche probabiliste de la classification automatique est de considérer que les données forment un échantillon aléatoire  $\mathbf{x}_1, \dots, \mathbf{x}_n$ , issu d'une population, et de s'appuyer sur l'analyse de la distribution de probabilité de cette population pour définir une classification. Différentes approches probabilistes de la classification ont été envisagées parmi lesquelles on peut distinguer les approches paramétriques et les approches non paramétriques.

### 8.2.1 Approches paramétriques

Une première approche consiste à faire des hypothèses sur la distribution de probabilité induisant une classification et formalisant ainsi la notion de classes « naturelles ». On peut distinguer plusieurs types de modèles paramétriques de la classification ; les plus importants étant les modèles de mélange fini de lois de probabilités, les modèles fonctionnels à effet fixe et les processus ponctuels.

#### Modèles de mélange fini

Les modèles de mélange fini, qui supposent que chaque classe est caractérisée par une distribution de probabilité, sont des modèles très souples permettant de prendre en compte des situations variées comme la présence de population hétérogène ou d'éléments aberrants. Grâce à l'algorithme d'estimation EM, particulièrement adapté à cette situation, les modèles de mélange ont fait l'objet de nombreux développements en statistique. Leur utilisation pour la classification a été considérée par de nombreux auteurs. Cette approche se justifie pour plusieurs raisons : elle correspond souvent à l'idée intuitive que l'on peut se faire d'une population composée de plusieurs classes ; elle possède des liens forts avec des méthodes de références comme l'algorithme des *k-means* ; enfin, elle est capable de prendre en compte de manière assez naturelle de nombreuses situations particulières. C'est l'approche qui va être développée dans ce chapitre.

#### Les modèles fonctionnels à effet fixe

Les modèles fonctionnels à effet fixe caractérisés par l'équation :

$$\text{Données} = \text{Structure} + \text{Erreur}$$

où la structure est inconnue mais fixe et l'erreur est aléatoire, peuvent être appliqués à la classification en choisissant une structure adaptée. Si les données sont des vecteurs  $\mathbf{x}_1, \dots, \mathbf{x}_n$  de  $\mathbb{R}^p$ , le modèle  $\mathbf{x}_i = \mathbf{y}_i + \varepsilon_i$ , où on impose aux  $\mathbf{y}_i$  d'appartenir à un ensemble de  $g$  centres  $\{\mathbf{a}_1, \dots, \mathbf{a}_g\}$  et aux erreurs  $\varepsilon_i$  de suivre une loi normale centrée de même variance, en est l'exemple le plus simple. On peut aussi appliquer ce type de modèles à des données de similarité et supposer, par exemple, que la dissimilarité  $d$  entre deux objets de l'ensemble à classifier s'écrit sous la forme  $d(a, b) = \delta(a, b) + \varepsilon(a, b)$  où  $\delta$  est une distance ultramétrique. Degens montre ainsi que la classification ascendante hiérarchique du lien moyen est un maximum local de la vraisemblance de ce modèle lorsque l'erreur est gaussienne.

#### Les processus ponctuels

En statistique spatiale, les données qui peuvent être, par exemple, la répartition des arbres dans une forêt ou des étoiles dans l'espace, sont considérées comme des semis de point issus de processus ponctuels. Certains de ces processus correspondent à une organisation en agrégats et peuvent être considérés comme des modèles probabilistes associés à une classification. Le plus utilisé est le processus de Neyman-Scott qui peut être interprété comme une génération des données en trois étapes : (a)  $g$  points  $\mathbf{a}_1, \dots, \mathbf{a}_g$  sont tirés au hasard suivant une distribution uniforme sur une région convexe ; (b) les tailles  $n_1, \dots, n_g$  des classes sont tirées au hasard, par exemple à l'aide d'une distribution de Poisson ; (c) pour chaque classe  $k$ ,  $n_k$  points sont tirés au hasard en utilisant une distribution sphérique centrée en  $\mathbf{a}_k$ , par exemple une distribution gaussienne de moyenne  $\mathbf{a}_k$ .

### 8.2.2 Approches non paramétriques

Cette seconde approche regroupe les méthodes probabilistes de classification qui ne supposent aucune hypothèse sur la distribution de probabilités. Ces méthodes peuvent être très variées mais vont toutes s'appuyer sur la forme de cette distribution. Lorsque les données sont continues, cette distribution est caractérisée par sa fonction de densité et ces méthodes utilisent alors cette fonction de densité pour définir la notion de classes,

par exemple des classes de forte densité ou des classes modales. Hartigan définit ainsi une classe de forte densité comme un sous-ensemble connexe de points de densité supérieure à un certain seuil et, en faisant varier ce seuil, il obtient un arbre hiérarchique de classes. La présence de plusieurs maxima de la densité peut être interprétée comme la présence de données hétérogènes et donc de classes. La recherche de ces maxima et l'affectation des points de l'espace de référence à chacun d'entre eux permet alors de définir les classes modales.

L'application de ces méthodes nécessite évidemment l'estimation de la distribution inconnue à partir des données. Les méthodes les plus courantes s'appuient sur une estimation non paramétrique de la densité comme l'estimation par la méthode des plus proches voisins, par la méthode des noyaux ou même simplement à l'aide d'un histogramme. Cette démarche a donné lieu à de nombreux algorithmes et des liens avec des algorithmes classiques comme l'algorithme de classification hiérarchique du lien minimal ont pu être établis. Nous ne détaillerons pas plus ces méthodes dans ce chapitre.

On pourrait aussi ranger dans ces approches non paramétriques les algorithmes de classification hiérarchique de Lerman définis à partir de la notion de vraisemblance du lien.

### 8.2.3 Validation

Une autre utilisation importante des outils probabilistes en classification concerne la validation des résultats. En effet, tout algorithme de classification fournissant toujours un résultat, il est nécessaire de savoir si ce résultat correspond à une véritable structure ou s'il est simplement le fait du hasard ; pour ceci, des outils de validation statistique ont été développés. Lorsque les résultats ont été obtenus à l'aide des modèles paramétriques précédents, une approche naturelle consiste à s'appuyer sur ces modèles pour définir de tels outils en vérifiant, par exemple, la normalité d'une classe dans le cas des modèles de mélanges gaussiens. Il existe aussi d'autres outils statistiques de validation qui sont indépendants des algorithmes de classification. On peut citer, par exemple, l'utilisation du modèle des graphes aléatoires de Ling pour tester la significativité d'une classe ; B. Van Cutsem et B. Ycart ont aussi étudié les structures de classification produites par différents algorithmes sous des hypothèses de non-classifiabilité aléatoire et ont proposé une batterie de tests explicites de non-classifiabilité. Pour avoir plus d'information, on pourra se reporter aux travaux de Bock qui donne une synthèse bibliographique détaillée de ces problèmes de validation.

### 8.2.4 Notations

Dans tout ce chapitre, nous supposons que les données se présentent sous la forme d'un échantillon  $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  où chaque individu est mesuré par un vecteur  $\mathbf{x}_i = (x_i^1, \dots, x_i^p)$ . Les données sont ainsi caractérisées par une matrice de dimension  $(n, p)$  que nous noterons aussi  $\mathbf{x}$  définie par les nombres  $x_i^j$  où  $i$  décrit un ensemble  $I$  de  $n$  individus et  $j$  décrit un ensemble de  $p$  variables qui pourront être continues ou qualitatives.

L'objectif sera donc la recherche d'une partition  $\mathbf{z}$  en  $g$  classes de l'ensemble  $I$ . Le nombre de classes  $g$  sera supposé connu.

Nous utiliserons les notations  $\mathbf{z} = (z_1, \dots, z_n)$ , où  $z_i \in \{1, \dots, g\}$  indique la classe de l'objet  $i$  et  $\mathbf{z} = (z_{11}, \dots, z_{ng})$  avec  $z_{ik} = 1$  si  $i$  appartient à la classe  $k$  et  $z_{ik} = 0$  sinon. Dans ce dernier cas, la classification sera représentée par une matrice  $\mathbf{z}$  vérifiant  $z_{ik} \in \{0, 1\}$  et  $\sum_{k=1}^g z_{ik} = 1$ . Ainsi, la classe  $k$  correspond à l'ensemble des objets  $i$  tel que  $z_i = k$  ou encore  $z_{ik} = 1$  et  $z_{i\ell} = 0 \forall \ell \neq k$ . Enfin, on notera  $n_k$  le cardinal de la classe  $k$ .

Par exemple, si l'ensemble  $I$  est constitué de 5 éléments, pour la partition  $\mathbf{z}$  constituée des 2 classes  $\{1, 3, 4\}$  et  $\{2, 5\}$ , on aura :

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 1 \\ 2 \end{pmatrix} \quad \text{notée aussi} \quad \mathbf{z} = \begin{pmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ z_{31} & z_{32} \\ z_{41} & z_{42} \\ z_{51} & z_{52} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

et  $n_1 = 3$  et  $n_2 = 2$ .

## 8.3 Le modèle de mélange

### 8.3.1 Introduction

Depuis leur utilisation par Newcomb en 1886 pour la détection de points aberrants, puis par Pearson en 1894 pour l'identification de deux populations de crabes, les mélanges finis de distributions ont permis de modéliser une grande variété de phénomènes aléatoires. Ces modèles supposent que les mesures sont effectuées sur un ensemble d'individus provenant de différentes classes dont l'origine est inconnue. Une étude portant sur la migration des passereaux permet d'illustrer cette situation : des mesures rapides, pour ne pas perturber les oiseaux, sont effectuées ; par exemple, nous disposons de la longueur de l'aile mais pas du sexe, plus difficile à identifier. Or l'étude statistique doit tenir compte de l'origine, mâle ou femelle, des oiseaux. Les données sont reportées dans le tableau 8.1 et la figure 8.1 représente graphiquement les fréquences.

Longueur	82	83	84	85	86	87	88	89	90
Fréquence	5	3	12	36	55	45	21	13	15
Longueur	91	92	93	94	95	96	97	98	
Fréquence	34	59	48	16	12	6	0	1	

TAB. 8.1 – Données passereau

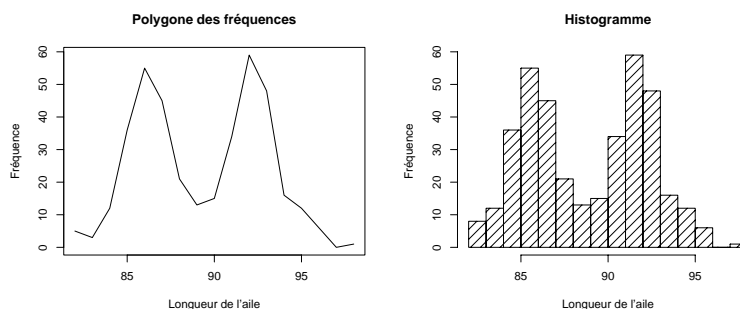


FIG. 8.1 – Longueurs des ailes en mm de 381 passereaux

Ce type de modèle permet donc de représenter l'hétérogénéité d'une population et sera tout particulièrement adapté au problème de la classification. Ce domaine a fait l'objet de très nombreux travaux ; le livre récent de McLachlan and Peel (2000) constitue une référence très détaillée de ce domaine qui s'est beaucoup développé ces dernières années. Dans ce paragraphe, nous rappelons succinctement le modèle et les problèmes posés par l'estimation de ses paramètres.

### 8.3.2 Le modèle

Si on reprend l'exemple précédent, les données peuvent être modélisées à l'aide d'un couple de variables aléatoires  $(X, Z)$  où  $X$  est la variable aléatoire continue associée à la longueur de l'aile et  $Z$  la variable aléatoire discrète associée au sexe. La distribution de probabilité d'un tel couple est définie par une fonction  $f(x, z)$  vérifiant

$$P(X \in I, Z \in A) = \sum_{z \in A} \int_I f(x, z) dx$$

où  $I$  est un intervalle réel et  $A$  est un sous-ensemble de l'ensemble {mâle, femelle}. Remarquons que la fonction  $f(x, z)$  n'est ni une densité ni une probabilité ; mais  $f(\cdot, z)$

pour  $z$  fixée est une densité et  $f(x, \cdot)$  pour  $x$  fixée est une probabilité. Ces deux fonctions définissent les distributions conditionnelles.

Si l'on cherche à ajuster ce modèle, comme on ne dispose que d'un échantillon de la variable aléatoire  $X$ , les valeurs de la variable  $Z$  étant manquantes, l'estimation des paramètres du modèle devra se faire à partir de la loi de  $X$ , c'est-à-dire d'une loi marginale qui peut être obtenue à partir de la loi du couple  $(X, Z)$  par la relation :

$$f(x) = \sum_{z=1}^2 f(x, z) = \sum_{z=1}^2 p(z)f(x/z) = \sum_{z=1}^2 \pi_z f_z(x)$$

où  $\pi_z = P(Z = z)$ ,  $f_z$  est la densité de  $X$  conditionnellement à  $Z = z$  et les valeurs de la variable  $Z$  ont été codées 1 et 2. La loi marginale de  $X$  est appelée *distribution de mélange fini à 2 composants*.

De manière plus générale, dans la suite les données disponibles pourront être des vecteurs de  $p$  mesures et la variable  $Z$  pourra prendre un nombre fini quelconque  $g$  de valeurs codées  $1, \dots, g$ ; les données  $\mathbf{x}_1, \dots, \mathbf{x}_n$  constitueront donc un échantillon de  $n$  réalisations indépendantes d'un vecteur aléatoire  $\mathbf{X}$  à valeur dans  $\mathbb{R}^p$  dont la fonction de densité peut s'écrire sous la forme :

$$f(\mathbf{x}) = \sum_{z=1}^g \pi_z f_z(\mathbf{x})$$

où  $g$  est le nombre de composants, les  $f_z$  sont les densités de chacun des composants et les  $\pi_z$  sont les proportions du mélange ( $\pi_z \in ]0, 1[ \forall z$  et  $\sum_z \pi_z = 1$ ).

Une interprétation de ce modèle de mélange consiste à considérer que, connaissant les proportions  $\pi_1, \dots, \pi_g$  et les distributions  $f_k$  de chaque classe, les données sont générées suivant le mécanisme suivant :

- $z$  : chaque individu est rangé dans une classe suivant les probabilités  $\pi_1, \dots, \pi_g$  ;
- $\mathbf{x}$  : chaque  $\mathbf{x}_i$  suit la loi de probabilité associée à la classe à laquelle il appartient.

Généralement, on suppose en plus que les densités  $f_k$  des composants appartiennent à une famille paramétrée  $f(\cdot, \alpha)$  ; la densité du mélange peut alors s'écrire :

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_z \pi_z f(\mathbf{x}; \alpha_z), \quad \forall i \in I,$$

plus souvent notée

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_k \pi_k f(\mathbf{x}, \alpha_k),$$

où  $\boldsymbol{\theta} = (\pi_1, \dots, \pi_g, \alpha_1, \dots, \alpha_g)$  est le paramètre du modèle.

Dans l'exemple des passereaux, il est certain qu'il existe une variable sexe ; la seule inconnue est la valeur de cette variable pour les individus de l'échantillon. En pratique, on utilisera le modèle de mélange dans des situations où l'existence même d'une telle variable n'est pas sûre. Par exemple, dans une étude portant sur les programmes de vaccination contre les oreillons, une enquête a fourni la log-concentration d'anticorps de 385 enfants non vaccinés contre les oreillons, tous âgés de 14 ans. L'histogramme de cette distribution est fourni dans la figure 8.2. Un mode important apparaît autour de la valeur 3 et un second mode, moins net, semble aussi apparaître autour de la valeur 0. Pour ce type de données, il est connu qu'une population homogène aurait du conduire à une distribution sensiblement gaussienne. Comme l'immunisation peut être obtenue par vaccination, ce qui n'est pas le cas dans cet échantillon, ou, naturellement, par contact avec le virus, une explication raisonnable des deux modes serait que la population est un mélange de deux groupes : les enfants immunisés naturellement et les enfants non immunisés. Contrairement à l'exemple précédent, cette fois les deux groupes sont moins séparés mais surtout, l'existence de deux groupes, qui avait une signification physique incontestable dans le premier cas, n'est maintenant qu'une hypothèse de travail suggérée par les données et qui n'est pas directement observable.

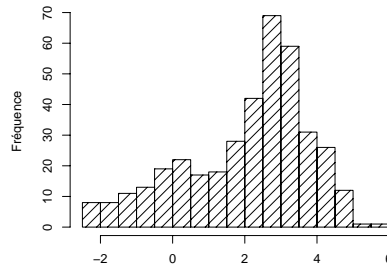


FIG. 8.2 – Histogramme de la log-concentration d'anticorps contre les oreillons de 385 enfants de 14 ans

### 8.3.3 Exemples

La densité d'un modèle de mélange de deux densités normales dans  $\mathbb{R}$  s'écrit

$$f(x; \pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \pi \varphi(x; \mu_1, \sigma_1^2) + (1 - \pi) \varphi(x; \mu_2, \sigma_2^2)$$

où  $\varphi(\cdot; \mu, \sigma^2)$  est la densité de la loi normale univariée de moyenne  $\mu$  et de variance  $\sigma^2$ . La figure 8.3 (a) représente les fonctions de densité de trois lois normales monodimensionnelles et figure 8.3(b) représente la fonction de densité mélange ainsi obtenue. Dans

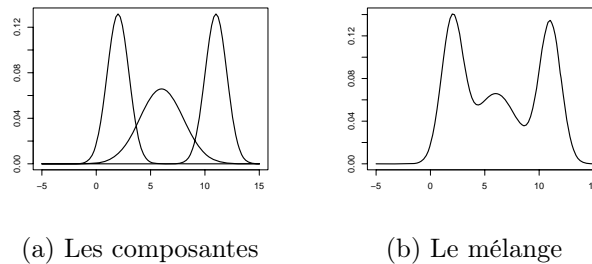


FIG. 8.3 – Mélanges gaussiens dans  $\mathbb{R}$

la figure 8.4, nous avons reporté la loi mélange correspondant aux paramètres suivants :

- (a) :  $p_1 = 0.5, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 1$
- (b) :  $p_1 = 0.25, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 3, \sigma_2 = 1$
- (c) :  $p_1 = 0.8, \mu_1 = 1, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 4$
- (d) :  $p_1 = 0.6, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2, \sigma_2 = 2$
- (e) :  $p_1 = 0.9, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2.5, \sigma_2 = 0.2$
- (f) :  $p_1 = 0.6, \mu_1 = 0, \sigma_1 = 1, \mu_2 = 2.5, \sigma_2 = 1$ .

Ces exemples permettent d'illustrer la richesse des situations qui peuvent être modélisées par une loi mélange. Un dernier exemple de mélange gaussien, cette fois dans  $\mathbb{R}^2$  est donné dans la figure 8.5.

### 8.3.4 Estimation des paramètres

L'estimation des paramètres du modèle de mélange a fait l'objet de nombreuses approches depuis les travaux de Pearson qui, pour estimer les 5 paramètres d'un modèle de mélanges gaussiens unidimensionnel à 2 composants  $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi)$  par la méthode des moments, fut conduit à résoudre des équations polynomiales de degré 9. Depuis, ce problème a fait l'objet de nombreux travaux et différentes méthodes d'estimation ont été envisagées : en dehors de la méthode des moments déjà citée, on retrouve des méthodes graphiques, la méthode du maximum de vraisemblance et des approches bayésiennes. La méthode la plus utilisée aujourd'hui est sans doute celle du maximum de vraisemblance à l'aide

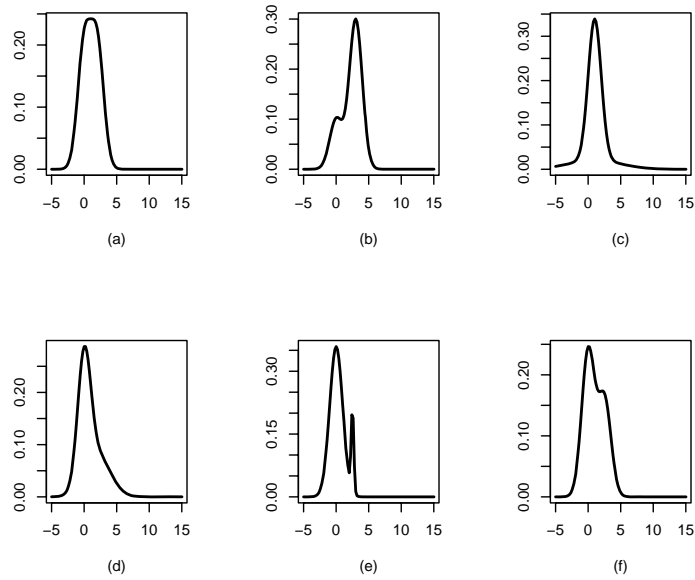


FIG. 8.4 – Mélanges gaussiens dans  $\mathbb{R}$

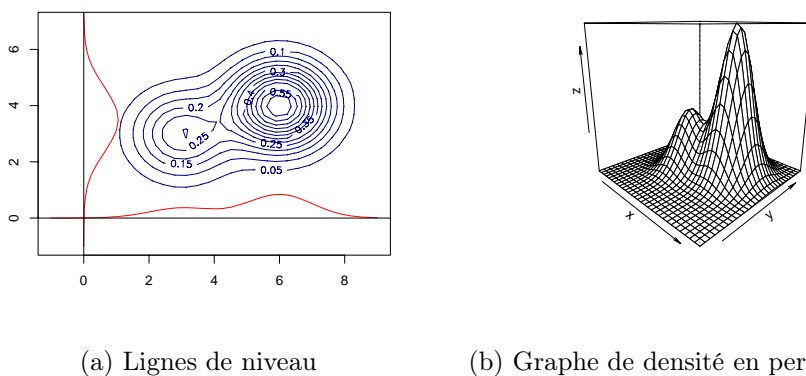


FIG. 8.5 – Mélange gaussien dans  $\mathbb{R}^2$

de l'algorithme EM. Avant de développer cette approche, nous allons évoquer quelques difficultés posées par l'estimation des paramètres d'un modèle de mélange.

### 8.3.5 Nombre de composants

Dans un certain nombre de situations, comme celle des oiseaux du paragraphe 8.3.1 où la notion de composant a une signification physique bien précise, le nombre de composants peut être parfaitement déterminé mais, le plus souvent, ce nombre inconnu est un paramètre supplémentaire qu'il faut aussi estimer.

Remarquons que, si l'on considère le nombre de composants comme un paramètre supplémentaire, le modèle de mélange peut être vu comme un compromis semi-paramétrique entre un problème d'estimation paramétrique classique quand le nombre de composant est égal à une constante fixée et un problème d'estimation non paramétrique, ici par la méthode des noyaux, quand le nombre de composants est égal à la taille de l'échantillon. Nous supposons dans la suite que le nombre  $g$  de composants est connu et nous verrons plus loin les solutions proposées pour effectuer ce choix difficile.

### 8.3.6 Identifiabilité

Pour que ce problème ait un intérêt, il est nécessaire que la densité du mélange soit identifiable, c'est-à-dire que deux mélanges ayant la même densité correspondent exactement aux mêmes paramètres. De nombreux travaux ont été menés sur ce problème. Plusieurs difficultés apparaissent. La première est due à la numérotation des classes ; par exemple, dans le cas d'un mélange de deux composants, les paramètres  $(\pi_1, \pi_2, \alpha_1, \alpha_2)$  et  $(\pi_2, \pi_1, \alpha_2, \alpha_1)$ , bien que différents, conduisent évidemment à la même densité : la densité d'un mélange n'est donc jamais identifiable ! Les difficultés entraînées par cette situation dépendront des algorithmes d'estimation. Par exemple, avec l'algorithme EM que nous utiliserons, cette difficulté n'est pas gênante – ce qui n'est pas le cas pour l'approche bayésienne où cette situation est connue sous le nom de « *switching problem* ». La seconde difficulté, celle-ci beaucoup plus gênante, peut provenir de la forme même des densités des composants. On peut ainsi facilement vérifier qu'un mélange de lois uniformes ou de lois binomiales n'est pas identifiable. Par contre, les mélanges gaussiens, exponentiels et de Poisson sont identifiables.

### 8.3.7 Estimation du maximum de vraisemblance

Rappelons que cette méthode, très largement utilisée, consiste à maximiser la log-vraisemblance

$$L(\boldsymbol{\theta}; \mathbf{x}) = \ln \left( \prod_i f(\mathbf{x}_i; \boldsymbol{\theta}) \right) = \sum_i \ln \left( \sum_k \pi_k f(\mathbf{x}_i; \boldsymbol{\alpha}_k) \right).$$

L'annulation des dérivées partielles conduit aux *équations de vraisemblance*.

Dans le cas du modèle de mélange, cette approche pose plusieurs difficultés :

- la résolution des équations de vraisemblance ainsi obtenues ne conduit pas à une solution analytique et une méthode de maximisation itérative, comme la méthode de Newton-Raphson ou l'algorithme EM de Dempster, Laird et Rubin (1977) est nécessaire ;
- il existe souvent de nombreux maxima locaux de la fonction de vraisemblance et les algorithmes précédents convergent vers un des ces maxima ce qui ne garantit pas l'obtention du maximum global de vraisemblance ;
- enfin, souvent la fonction de vraisemblance n'est pas bornée supérieurement : il n'y a donc pas de maximum de vraisemblance. Par exemple, dans le cas d'un modèle gaussien monodimensionnel, cette situation se produit si la variance d'un des composants tend vers 0. On traitera ce type de problème en empêchant l'algorithme d'atteindre ces solutions inintéressantes.

## 8.4 Algorithme EM

Lorsque la maximisation de la vraisemblance ne conduit pas à une solution analytique, l'algorithme EM est méthode de maximisation itérative souvent simple à mettre en place. L'idée fondamentale est de considérer que les données observées  $\mathbf{x}$  ne correspondent qu'à une connaissance partielle de données  $\mathbf{y}$  inconnues, appelées *données complétées*, et que la maximisation de la vraisemblance associée à ces nouvelles données conduit à une solution explicite.

### 8.4.1 Données complétées et vraisemblance complétée

Le seule hypothèse qui est faite sur le lien entre les données complétées et les données observées est qu'elles sont reliées par une fonction  $\mathbf{x} = h(\mathbf{y})$ . Ces données complétées peuvent prendre, par exemple, la forme  $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ ;  $\mathbf{z}$  est alors appelée *information manquante*. Cette notion de données complétées peut avoir une véritable signification pour le modèle comme ce sera le cas pour le modèle de mélange, ou être, au contraire, complètement artificielle. La vraisemblance  $f(\mathbf{y}; \boldsymbol{\theta})$  calculée à partir de ces données complétées est appelée *vraisemblance complétée* ou, dans le cas du modèle de mélange, *vraisemblance classifiante*. Partant de la relation

$$f(\mathbf{y}; \boldsymbol{\theta}) = f(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})f(\mathbf{x}; \boldsymbol{\theta}) \quad \text{où } \mathbf{x} = h(\mathbf{y})$$

qui s'écrit encore

$$f(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{y}; \boldsymbol{\theta})/f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) \quad \forall \mathbf{y} \in h^{-1}(\mathbf{x})$$

on obtient la relation :

$$L(\boldsymbol{\theta}; \mathbf{x}) = L(\boldsymbol{\theta}; \mathbf{y}) - \log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) \quad \forall \mathbf{y} \in h^{-1}(\mathbf{x}) \quad (8.1)$$

entre la log-vraisemblance initiale  $L(\boldsymbol{\theta}; \mathbf{x})$  et la log-vraisemblance complétée  $L(\boldsymbol{\theta}; \mathbf{y})$ .

### 8.4.2 Principe

La vraisemblance  $L(\boldsymbol{\theta}; \mathbf{y})$ , supposée simple à maximiser, n'est pas calculable puisque  $\mathbf{y}$  est inconnu. La procédure itérative va alors s'appuyer sur la maximisation de l'espérance conditionnelle de la vraisemblance  $L(\boldsymbol{\theta}; \mathbf{Y})$  pour une valeur du paramètre courant  $\boldsymbol{\theta}^{(c)}$ . En effet, en calculant l'espérance conditionnelle des 2 membres de la relation 8.1, on obtient la relation fondamentale de l'algorithme EM :

$$L(\boldsymbol{\theta}; \mathbf{x}) = \underbrace{E(L(\boldsymbol{\theta}; \mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}^c)}_{Q(\boldsymbol{\theta}, \boldsymbol{\theta}^c)} - \underbrace{E(\log f(\mathbf{Y}|\mathbf{x}, \boldsymbol{\theta})|\mathbf{x}, \boldsymbol{\theta}^c)}_{H(\boldsymbol{\theta}, \boldsymbol{\theta}^c)}$$

et en itérant à partir d'une valeur initiale  $\boldsymbol{\theta}^0$ , la relation  $\boldsymbol{\theta}^{c+1} = \text{Argmax } Q(\boldsymbol{\theta}, \boldsymbol{\theta}^c)$ , on définit un algorithme faisant croître la vraisemblance car, si on note  $\Delta L = L(\boldsymbol{\theta}^{c+1}; \mathbf{x}) - L(\boldsymbol{\theta}^c; \mathbf{x})$ , on obtient

$$\Delta L = \underbrace{(Q(\boldsymbol{\theta}^{c+1}, \boldsymbol{\theta}^c) - Q(\boldsymbol{\theta}^c, \boldsymbol{\theta}^c))}_{\geq 0 \text{ par construction}} - \underbrace{(H(\boldsymbol{\theta}^{c+1}, \boldsymbol{\theta}^c) - H(\boldsymbol{\theta}^c, \boldsymbol{\theta}^c))}_{\leq 0 \text{ Inégalité de Jensen}} \geq 0.$$

Une itération de l'algorithme EM ainsi défini se décompose en deux étapes :

- étape E (estimation) : calcul de  $Q$  à partir de  $\boldsymbol{\theta}^c$
- étape M (maximisation) : détermination de  $\boldsymbol{\theta}^{c+1}$ .

### 8.4.3 Propriétés

Sous certaines conditions de régularité, il a été établi que l'algorithme EM assure une convergence vers un maximum local de la vraisemblance. Il a un bon comportement pratique mais peut être toutefois assez lent dans certaines situations; c'est le cas, par exemple, si les classes sont très mélangées. Cet algorithme, proposé par Dempster, Laird et Rubin dans un papier célèbre, souvent simple à mettre en place, est devenu populaire et a fait l'objet de nombreux travaux que l'on pourra trouver dans l'ouvrage très complet de McLachlan and Krishnan (1997).

### 8.4.4 Application au modèle de mélange

Pour le modèle de mélange, les données complétées sont obtenues en ajoutant le composant d'origine  $\mathbf{z}_i$  de chaque individu de l'échantillon :

$$\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n) = ((\mathbf{x}_1, z_1), \dots, (\mathbf{x}_n, z_n)).$$

Si on code  $z_i = (z_{i1}, \dots, z_{ig})$  où  $z_{ik}$  est égal à 1 si  $i$  appartient au composant  $k$  et 0 sinon, on obtient les relations suivantes :

$$\begin{aligned} f(\mathbf{y}_i; \boldsymbol{\theta}) &= f(\mathbf{x}_i, z_i; \boldsymbol{\theta}) = \pi_{z_i} f(\mathbf{x}_i; \alpha_{z_i}), \\ L(\boldsymbol{\theta}; \mathbf{y}) &= \log(f(\mathbf{y}; \boldsymbol{\theta})) = \sum_i f(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_i \log(\pi_{z_i} f(\mathbf{x}_i; \alpha_{z_i})) \\ &= \sum_{i,k} z_{ik} \log(\pi_k f(\mathbf{x}_i; \alpha_k)), \\ Q(\boldsymbol{\theta}, \boldsymbol{\theta}') &= E(L(\boldsymbol{\theta}; \mathbf{Y}) | \mathbf{x}, \boldsymbol{\theta}') = \sum_{i,k} E(Z_{ik} | \mathbf{x}, \boldsymbol{\theta}') \log \pi_k f(\mathbf{x}_i; \alpha_k) \\ &= \sum_{i,k} t_{ik} \log \pi_k f(\mathbf{x}_i; \alpha_k) \quad \text{« vraisemblance pondérée »} \end{aligned}$$

où  $t_{ik} = E(Z_{ik} | \mathbf{x}, \boldsymbol{\theta}') = P(Z_{ik} = 1 | \mathbf{x}, \boldsymbol{\theta}')$  sont les probabilités d'appartenance a posteriori. L'algorithme EM prend alors la forme suivante :

- initialisation : choix arbitraire d'une solution initiale  $\boldsymbol{\theta}^{(0)}$  ;
- répétition jusqu'à la convergence des 2 étapes suivantes :
  - étape E (estimation) : calcul des probabilités d'appartenance des  $\mathbf{x}_i$  aux classes conditionnellement au paramètre courant :

$$t_{ik}^{(c)} = \pi_k^{(c)} f(\mathbf{x}_i, \alpha_k^{(c)}) / \left( \sum_{\ell} \pi_{\ell}^{(c)} f(\mathbf{x}_i, \alpha_{\ell}^{(c)}) \right);$$

- étape M (maximisation) : maximisation de la vraisemblance conditionnellement aux  $t_{ik}^{(c)}$  ; les proportions sont alors obtenues simplement par la relation  $\pi_k^{(c+1)} = 1/n \sum_i t_{ik}^{(c)}$  alors que les paramètres  $\alpha_k^{(c+1)}$  sont obtenus en résolvant des équations de vraisemblance qui dépendent du modèle de mélange retenu.

### 8.4.5 Exemple des mélanges gaussiens monodimensionnel à 2 composants

Dans cette situation, l'algorithme EM s'écrit :

- Initialisation de  $\pi_1^{(0)}, \pi_2^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, (\sigma_1^2)^{(0)}$  et  $(\sigma_2^2)^{(0)}$  ;
- étape E : calcul des  $t_{ik}^{(c)}$  pour  $i = 1 \dots, n$

$$\begin{aligned} t_{i1}^{(c)} &= \frac{\pi_1^{(c)} \varphi(x_i, \mu_1^{(c)}, (\sigma_1^2)^{(c)})}{\pi_1^{(c)} \varphi(x_i, \mu_1^{(c)}, (\sigma_1^2)^{(c)}) + \pi_2^{(c)} \varphi(x_i, \mu_2^{(c)}, (\sigma_2^2)^{(c)})} \\ t_{i2}^{(c)} &= \frac{\pi_2^{(c)} \varphi(x_i, \mu_2^{(c)}, (\sigma_2^2)^{(c)})}{\pi_1^{(c)} \varphi(x_i, \mu_1^{(c)}, (\sigma_1^2)^{(c)}) + \pi_2^{(c)} \varphi(x_i, \mu_2^{(c)}, (\sigma_2^2)^{(c)})} \end{aligned}$$

- étape M :

$$\begin{aligned} \pi_1^{(c+1)} &= \frac{\sum_i t_{i1}^{(c+1)}}{n} \quad \text{et} \quad \pi_2^{(c+1)} = 1 - \pi_1^{(c+1)} \\ \mu_1^{(c+1)} &= \frac{\sum_i t_{i1}^{(c+1)} x_i}{\sum_i t_{i1}^{(c+1)}} \quad \text{et} \quad \mu_2^{(c+1)} = \frac{\sum_i t_{i2}^{(c+1)} x_i}{\sum_i t_{i2}^{(c+1)}} \\ (\sigma_1^2)^{(c+1)} &= \frac{\sum_i t_{i1}^{(c+1)} (x_i - \mu_1^{(c+1)})^2}{\sum_i t_{i1}^{(c+1)}} \quad \text{et} \quad (\sigma_2^2)^{(c+1)} = \frac{\sum_i t_{i2}^{(c+1)} (x_i - \mu_2^{(c+1)})^2}{\sum_i t_{i2}^{(c+1)}}. \end{aligned}$$

L'application de cet algorithme aux exemples précédents fournit les résultats reportés dans le tableau 8.2 et la figure 8.6.

	Paramètres	$\pi_1$	$\pi_2$	$\mu_1$	$\mu_2$	$\sigma_1^2$	$\sigma_2^2$	nb d'itér.
Passereaux	initiaux	0.50	0.50	85	95	1	1	
	obtenus	0.49	0.51	86.1	92.3	2.2	2.5	54
Oreillons	initiaux	0.50	0.50	-3	5	1	1	
	obtenus	0.30	0.70	-0.07	2.98	1.35	0.79	221

TAB. 8.2 – Résultats numériques de l'algorithme EM

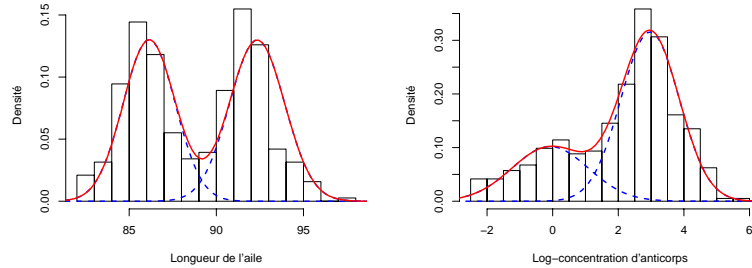


FIG. 8.6 – Densités mélange obtenues

## 8.5 Classification et modèle de mélange

### 8.5.1 Les deux approches

L'utilisation des modèles de mélange pour obtenir une partition des données initiales peut se faire de deux manières :

- la première, appelée *approche mélange*, estime les paramètres du modèle puis détermine la partition en rangeant chaque individu dans la classe maximisant la probabilité *a posteriori*  $t_{ik}$  calculée à partir des paramètres estimés ; cette affectation est connue sous le nom de la méthode du MAP ou maximum *a posteriori* ;
- la seconde, appelée *approche classification*, consiste à rechercher une partition de l'échantillon de telle sorte que chaque classe  $k$  soit assimilable à un sous-échantillon issue de la loi  $f(\cdot, \alpha_k)$ . Il s'agit donc d'estimer simultanément les paramètres du modèle et la partition recherchée.

Dans la suite de cette section, nous précisons le critère optimisé par cette dernière approche et l'algorithme d'optimisation généralement utilisé dans cette situation. Nous faisons ensuite une rapide comparaison des deux approches et étudions les liens que peut avoir ce type de méthodes avec les approches métriques plus classiques de la classification. Nous terminons cette section sur l'interprétation que l'on peut faire du modèle de mélange en terme de classification floue.

### 8.5.2 La vraisemblance classifiante

L'introduction de la partition  $\mathbf{z}$  dans le critère de vraisemblance n'est pas immédiate et plusieurs propositions ont été faites : Scott et Symons définissent le critère :

$$L_{CR}(\boldsymbol{\theta}, \mathbf{z}) = \sum_k \sum_{i/z_i=k} \log f(\mathbf{x}_i, \alpha_k)$$

dans lequel les proportions n'apparaissent pas. Symons, remarquant que ce critère a tendance à donner des classes de mêmes proportions, le modifie pour finalement utiliser la log-vraisemblance complétée (ou classifiante) définie précédemment :

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = \sum_{k,i} z_{ik} \log \pi_k f(\mathbf{x}_i, \alpha_k)$$

liée au critère précédent par la relation :

$$L_C(\boldsymbol{\theta}, \mathbf{z}) = L_{CR}(\boldsymbol{\theta}, \mathbf{z}) + \sum_k \#z_k \log \pi_k$$

où  $\#z_k$  est le cardinal de la classe  $k$ . La quantité  $\sum_k \#z_k \log \pi_k$  représente un terme de pénalité qui disparaît si on impose aux proportions d'être toutes identiques. Le critère  $L_{CR}(\boldsymbol{\theta}, \mathbf{z})$  apparaît donc comme une variante de la vraisemblance classifiante restreinte à un modèle de mélange où les classes ont toutes la même proportion.

### 8.5.3 L'algorithme CEM

Pour maximiser la vraisemblance classifiante, il est possible d'utiliser une version classifiante de l'algorithme EM obtenue en lui ajoutant une étape de classification. On obtient ainsi l'algorithme de classification très général appelé CEM (classification EM) (Celeux and Govaert, 1992) défini de la manière suivante :

- étape 0 : choix arbitraire d'une solution initiale  $\boldsymbol{\theta}^{(0)}$  ;
- étape E : calcul des  $t_{ik}^{(c)}$  comme dans l'algorithme EM ;
- étape C : la partition  $\mathbf{z}^{(c+1)}$  est obtenue en rangeant chaque  $\mathbf{x}_i$  dans la classe maximisant  $t_{ik}^{(c)}$  (MAP) ;
- étape M : maximisation de la vraisemblance conditionnellement aux  $z_{ik}^{(c+1)}$  : les estimations du maximum de vraisemblance des  $\pi_k$  et des  $\alpha_k$  sont obtenues en utilisant les classes de la partition  $\mathbf{z}^{(c+1)}$  comme sous-échantillons. Les proportions sont alors fournies par la formule  $\pi_k^{(c+1)} = \frac{1}{n} \#z_k^{(c+1)}$ , le calcul des  $\alpha_k^{(c+1)}$  dépendant du modèle de mélange retenu.

On retrouve ici un algorithme d'optimisation alternée de type nuées dynamiques (Diday, E. et Collaborateurs, 1979) où les étapes *E* et *C* correspondent à l'étape d'affectation et l'étape *M* à l'étape de représentation.

On peut montrer que cet algorithme itératif est stationnaire et fait croître à chaque itération la vraisemblance complétée sous des conditions très générales.

### 8.5.4 Comparaison des deux approches

L'approche classification, déterminant à chaque itération les paramètres à l'aide d'échantillons tronqués du modèle de mélange, fournit une estimation biaisée et inconsistante car le nombre de paramètres à estimer croît avec la taille de l'échantillon. Différents auteurs ont étudié ce problème et ont montré qu'il est généralement préférable d'utiliser l'approche mélange.

Toutefois, lorsque les classes sont bien séparées et les effectifs relativement petits, l'approche classification peut fournir de meilleurs résultats. D'autre part, l'algorithme CEM est beaucoup plus rapide que l'algorithme EM et son utilisation peut être nécessaire lorsque des contraintes de temps de calcul sont imposées, pour un fonctionnement en temps réel par exemple, ou sur de données de très grande taille.

Enfin, l'approche classification a l'avantage de pouvoir présenter de nombreux algorithmes de classification comme des cas particuliers de l'algorithme CEM et de pouvoir les englober dans une approche probabiliste de la classification. Ainsi, par exemple, nous verrons dans le paragraphe 8.6.3 que l'algorithme des centres-mobiles correspond à un cas particulier simple de l'algorithme CEM. Nous montrerons en particulier que les critères optimisés, l'inertie intraclasse pour les données continues et le critère d'information pour les données qualitatives, correspondent à la vraisemblance classifiante associée à un modèle de mélange particulier.

### 8.5.5 Classification floue

Dans les méthodes de classification floue, l'appartenance, vraie ou fausse, d'un objet à une classe est remplacée par un degré d'appartenance. Formellement, une classification floue sera caractérisée par une matrice  $\mathbf{c}$  de terme général  $c_{ik}$  vérifiant  $c_{ik} \in [0, 1]$  et

$\sum_k c_{ik} = 1$ . La méthode des « k-moyennes floues » de Bezdek (1981), l'une des plus répandues, consiste à minimiser le critère :

$$W(\mathbf{c}) = \sum_{i,k} c_{ik}^\gamma d^2(\mathbf{x}_i, \mathbf{g}_k)$$

où  $\gamma > 1$  est un coefficient permettant de régler le degré de flou,  $\mathbf{g}_k$  est le centre de gravité de la classe et  $d$  est la distance euclidienne. Il est nécessaire d'imposer à  $\gamma$  d'être différent de 1, sinon la fonction  $W$  est minimale pour des valeurs de  $c_{ik} = 0$  ou 1 et on retrouve le critère habituel d'inertie intraclasse. Les valeurs généralement conseillées se situent entre 1 et 2. La minimisation de ce critère se fait à l'aide d'un algorithme qui alterne les deux étapes suivantes :

1. calcul des centres :  $\mathbf{g}_k = \sum_i c_{ik}^\gamma \mathbf{x}_i / \sum_i c_{ik}^\gamma$  ;
2. calcul de la partition floue :  $c_{ik} = \frac{D_i}{\|\mathbf{x}_i - \mathbf{g}_k\|^{\frac{2}{\gamma-1}}}$  où  $D_i = \sum_\ell \frac{1}{\|\mathbf{x}_i - \mathbf{g}_\ell\|^{\frac{2}{\gamma-1}}}$ .

La validation d'une telle approche avec, en particulier le choix du coefficient  $\gamma$ , est assez délicate.

L'estimation des paramètres d'un modèle de mélange est une autre façon d'aborder, et de manière plus naturelle, ce problème. En effet, l'estimation des probabilités *a posteriori*  $t_{ik}$  d'appartenance des objets à chaque classe fournit directement une classification floue et l'algorithme *EM*, appliqué au modèle de mélange, peut être considéré comme un algorithme de classification floue.

Hathaway (1986) a montré que la recherche d'une partition floue et du paramètre  $\theta$ , effectuée à l'aide d'une optimisation alternée d'un critère de classification floue, conduit exactement aux deux étapes de l'algorithme *EM* qui peut donc être considéré comme un algorithme de classification floue. Il montre en particulier que l'algorithme *EM* maximise de manière alternée le critère :

$$W(\mathbf{c}, \theta) = L_C(\theta, \mathbf{c}) + H(\mathbf{c})$$

où  $L_C$  est la fonction de log-vraisemblance complétée dans laquelle on a remplacé la partition  $\mathbf{z}$  par la partition floue  $\mathbf{c}$  :

$$L_C(\theta, \mathbf{c}) = \sum_{i,k} c_{ik} \log(\pi_k f(\mathbf{x}_i; \alpha_k))$$

et  $H$  est la fonction d'entropie :

$$H(\mathbf{c}) = - \sum_i \sum_k c_{ik} \log c_{ik}.$$

Il est facile de vérifier que, si l'on supprime le terme d'entropie du critère  $W$ , on obtient à chaque étape des partitions « dures ». L'algorithme ainsi obtenu est alors simplement l'algorithme *CEM* : la différence entre l'algorithme *EM* et l'algorithme *CEM* est la présence du terme d'entropie. Si, à la convergence de *EM*, les composants sont très séparés, la partition floue  $\mathbf{z}(\theta)$  est proche d'une partition et on a  $H(\mathbf{z}(\theta)) \approx 0$  et  $L(\theta) = W(\mathbf{z}(\theta), \theta) = L_C(\theta, \mathbf{z}(\theta)) + H(\mathbf{z}(\theta)) \approx L_C(\theta, \mathbf{z}(\theta))$ .

## 8.6 Modèle de mélange gaussien

Nous allons maintenant étudier ce que devient cette approche lorsque chaque classe est modélisée par une distribution normale.

### 8.6.1 Le modèle

La densité du mélange s'écrit  $f(\mathbf{x}; \theta) = \sum_k \pi_k \varphi(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k)$  où  $\varphi$  est la densité de la loi normale multidimensionnelle :

$$\varphi(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right\}$$

et  $\theta$  est le vecteur  $(\pi_1, \dots, \pi_g, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_g, \Sigma_1, \dots, \Sigma_g)$  formé des proportions  $\pi_k$  et des paramètres  $\boldsymbol{\mu}_k$  et  $\Sigma_k$  qui sont respectivement le vecteur moyenne et la matrice de variance de la classe  $k$ . Si on note  $d_{\Sigma_k}^2(\mathbf{x}_i, \boldsymbol{\mu}_k)$  la distance quadratique définie par  $\Sigma_k^{-1}$ ,  $\varphi$  s'écrit aussi

$$\varphi(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp -\frac{1}{2} d_{\Sigma_k}^2(\mathbf{x}_i, \boldsymbol{\mu}_k).$$

Lorsque la taille de l'échantillon est faible ou lorsque la dimension de l'espace est grande, il devient nécessaire de diminuer le nombre de paramètres afin d'obtenir des modèles plus parcimonieux. Pour ceci, il est possible de décomposer les matrices de variances en utilisant leur décomposition en valeurs propres et vecteurs propres :

$$\Sigma_k = D_k B_k D_k'$$

où  $D_k$  est la matrice des vecteurs propres et  $B_k$  la matrice diagonale composée des valeurs propres. Pour obtenir une décomposition unique, les valeurs propres sont ordonnées suivant leurs valeurs décroissantes. Ensuite, chaque matrice  $B_k$  peut être elle-même décomposée en un nombre réel  $\lambda_k$  et une matrice  $A_k$  tel que

$$B_k = \lambda_k A_k \text{ avec } |A_k| = 1.$$

Chaque matrice de variance est donc décomposée sous la forme  $\Sigma_k = \lambda_k D_k A_k D_k'$  où  $A_k$ , matrice diagonale de déterminant 1, avec des valeurs allant en décroissant, caractérise la *forme* de la classe,  $D_k$ , matrice orthogonale, caractérise l'*orientation* de la classe et  $\lambda_k$ , nombre réel positif, représente le *volume* de la classe.

Le modèle de mélange est finalement paramétré par les centres  $\mu_1, \dots, \mu_K$ , les proportions  $\pi_1, \dots, \pi_K$ , les volumes  $\lambda_1, \dots, \lambda_K$ , les formes  $A_1, \dots, A_K$  et les orientations  $D_1, \dots, D_K$  de chaque classe.

Par exemple, lorsque les données sont dans un plan,  $D$  est une matrice de rotation définie par un angle  $\alpha$  et  $A$  est une matrice diagonale de termes diagonaux  $a$  et  $1/a$ . La figure 8.7 représente alors l'ellipse d'équidensité de cette distribution en fonction des valeurs  $\alpha$ ,  $\lambda$  et  $a$ .

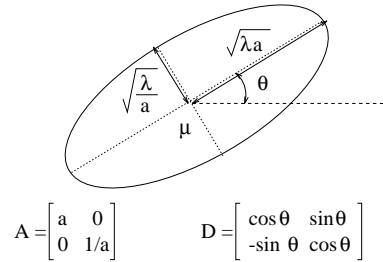


FIG. 8.7 – Paramétrage d'une classe gaussienne dans le plan

En utilisant cette paramétrisation, il est alors possible de proposer des situations intermédiaires entre des hypothèses restrictives (matrices de variance proportionnelles à la matrice identité ou matrices de variance identiques pour toutes les classes) et les hypothèses très générales (aucune contrainte) (Celeux and Govaert, 1995).

Cette paramétrisation met aussi en évidence deux notions souvent confondues sous l'appellation un peu floue de taille : la proportion des individus présents dans une classe et le volume que représente la place occupée par une classe dans l'espace. En particulier, il est possible d'avoir des classes de faible volume et de grande proportion et, vice-versa, des classes de grand volume et de faible proportion.

Nous étudions maintenant ce que devient l'algorithme CEM et le critère de vraisemblance classifiante pour le modèle de mélange gaussien. Remarquons que ce travail pourrait se faire de manière similaire avec l'algorithme EM.

### 8.6.2 L'algorithme CEM

#### Étape de classification

Chaque  $\mathbf{x}_i$  est rangé dans la classe qui maximise la probabilité d'appartenance  $t_{ik} = \pi_k \varphi(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k) / (\sum_{\ell} \pi_{\ell} \varphi(\mathbf{x}_i; \boldsymbol{\mu}_{\ell}, \Sigma_{\ell}))$ , c'est-à-dire  $\pi_k \varphi(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)$  ou, de manière équivalente, qui minimise  $-\log(\pi_k \varphi(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k))$  ou encore :

$$d_{\Sigma_k^{-1}}^2(\mathbf{x}_i, \boldsymbol{\mu}_k) + \log |\Sigma_k| - 2 \log(\pi_k). \quad (8.2)$$

#### Étape M

Cette fois, pour une partition  $\mathbf{z}$  donnée, il s'agit de déterminer le paramètre  $\theta$  maximisant  $L_C(\theta, \mathbf{z})$  égal à

$$-\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_k \left( \sum_{i/z_i=k} (\mathbf{x}_i - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) + n_k \log |\Sigma_k| - 2n_k \log \pi_k \right)$$

Le paramètre  $\boldsymbol{\mu}_k$  est alors nécessairement le centre de gravité  $\bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i \in z_k} \mathbf{x}_i$  et les proportions, si elles sont libres, vérifient  $\pi_k = n_k/n$ . Les paramètres  $\Sigma_k$  doivent alors minimiser la fonction :

$$F(\Sigma_1, \dots, \Sigma_g) = \sum_k n_k (\text{trace}(S_k \Sigma_k^{-1}) + \log |\Sigma_k|) \quad (8.3)$$

où  $S_k = \frac{1}{n_k} \sum_{i \in z_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)'$  est la matrice de variance de  $z_k$  et on a

$$L_C(\mathbf{z}, \theta) = -\frac{1}{2} F(\Sigma_1, \dots, \Sigma_g) + \sum_k n_k \log \pi_k - \frac{np}{2} \log 2\pi.$$

Nous allons maintenant examiner trois situations particulières.

### 8.6.3 Forme sphérique, proportions et volumes identiques

Nous étudions la situation la plus simple correspondant à des classes ayant toutes une distribution normale sphérique de même volume et ayant les mêmes proportions. Les matrices de variance s'écrivent  $\Sigma_k = \lambda D_k I D_k' = \lambda I \quad \forall k$  et la formule [8.2] montre que l'affectation des individus aux classes se fait simplement en utilisant la distance euclidienne habituelle  $d^2(\mathbf{x}_i, \boldsymbol{\mu}_k)$ . La fonction  $F$  devient alors :

$$F(\lambda) = n \left( \frac{1}{\lambda} \text{trace}(S_W) + p \log(\lambda) \right)$$

où  $S_W = \frac{1}{n} \sum_k n_k S_k$  est la matrice de variance intraclasse. On obtient alors  $\lambda = \frac{\text{trace}(S_W)}{p}$  et la vraisemblance classifiante s'écrit :

$$L_C(\mathbf{z}, \theta) = -\frac{1}{2} \left( np + np \log \frac{\text{trace}(S_W)}{p} \right) - n \log g - \frac{np}{2} \log 2\pi$$

La maximisation de la vraisemblance classifiante est donc équivalente à la minimisation du critère de variance intraclasse  $\text{trace}(S_W)$ . En outre, l'algorithme CEM est alors simplement l'algorithme des centres mobiles (*k-means*). En conclusion, utiliser le critère d'inertie revient à supposer que les classes sont sphériques, de même proportion et de même volume.

### 8.6.4 Forme sphérique, proportions identiques, volumes différents

On reprend le modèle précédent en le modifiant légèrement pour prendre en compte l'existence de classes pouvant avoir des volumes différents. Cette fois, les matrices de variance s'écrivent  $\Sigma_k = \lambda_k I$  et la formule [8.2] montre que l'affectation des individus aux classes se fait en utilisant la distance :

$$\frac{1}{\lambda_k} d^2(\mathbf{x}_i, \mu_k) + p \log(\lambda_k).$$

La distance d'un point au centre d'une classe a été modifiée par une quantité qui dépend du volume de la classe. Cette modification a des répercussions importantes ; par exemple, les surfaces séparatrices, qui étaient dans le cas précédent des hyperplans, deviennent des hypersphères. On peut montrer que l'on obtient

$$F(\lambda_1, \dots, \lambda_k) = \sum_k n_k \left( \frac{1}{\lambda_k} \text{trace}(S_k) + p \log \lambda_k \right)$$

$$\lambda_k = \frac{\text{trace}(S_k)}{p}.$$

La vraisemblance classifiante à maximiser s'écrit alors

$$L_C(\mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2} \sum_k p \left( n_k + n_k \log \frac{\text{trace}(S_k)}{p} \right) - n \log g - \frac{np}{2} \log 2\pi$$

ce qui revient à minimiser

$$\sum_k n_k \log \text{trace}(S_k).$$

Ce modèle permet de reconnaître des situations comme celle de la figure 8.8 sans aucune difficulté. Dans cet exemple, les 2 classes ont été simulées suivant deux lois normales sphériques avec les mêmes proportions mais avec des volumes très différents. Le résultat obtenu avec le critère d'inertie intraclasse classique correspond à la séparation de la population par la droite et n'a donc aucun rapport avec la partition simulée. Avec le modèle à volume variable, la partition obtenue, indiquée par le cercle est très proche de la classification initiale.

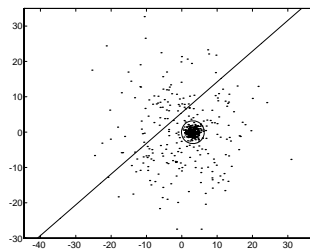


FIG. 8.8 – Exemple de classes de volumes différents

Remarquons que, sans l'aide du modèle de mélange, il aurait été difficile de proposer la distance et le critère utilisés dans cette approche à partir d'une simple interprétation métrique.

### 8.6.5 Formes diagonales identiques, proportions identiques

La matrice de variance de chaque classe a maintenant la forme

$$\Sigma_k = \lambda B$$

où  $B$  est une matrice diagonale de déterminant 1. L'affectation des individus aux classes se fait donc en utilisant la distance

$$d_{B^{-1}}^2(\mathbf{x}_i, \mu_k)$$

qui correspond à une distance euclidienne avec des pondérations sur les variables. La fonction  $F$  devient

$$F(\lambda, B) = \frac{n}{\lambda} \text{tr}(S_W B^{-1}) + np \ln(\lambda).$$

On peut alors montrer que la matrice  $B$  et la valeur  $\lambda$  minimisant  $F$  sont

$$B = \frac{\text{diag}(S_W)}{|\text{diag}(S_W)|^{1/p}}$$

et

$$\lambda = |\text{diag}(S_W)|^{1/p}$$

où  $\text{diag}(S_W)$  est la matrice diagonale obtenue en ne conservant que la diagonale de  $S_W$ . La vraisemblance classifiante à maximiser s'écrit alors

$$L_C(\mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2} (np + n \log |\text{diag}(S_W)|) - n \log g - \frac{np}{2} \log 2\pi$$

Cette méthode minimise donc le critère  $|\text{diag}(S_W)|$  et l'objectif d'une telle méthode peut être interprétée comme la recherche simultanée d'une partition et d'une pondération des variables.

### 8.6.6 Formes identiques, proportions identiques

La matrice de variance de chaque classe a maintenant la forme  $\Sigma_k = \Sigma$ . L'affectation des individus aux classes se fait alors en utilisant la distance

$$d_{\Sigma^{-1}}^2(\mathbf{x}_i, \mu_k)$$

et la fonction  $F$  devient  $F(\Sigma) = n (\text{trace}(S_W \Sigma^{-1}) + \log |\Sigma|)$ . La matrice optimale  $\Sigma$  vérifie donc

$$\Sigma = S_W.$$

La vraisemblance classifiante à maximiser s'écrit alors

$$L_C(\mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2} (np + n \ln |S_W|) - n \log g - \frac{np}{2} \log 2\pi$$

On retrouve ainsi le critère  $|S_W|$  quelquefois proposé justement pour permettre d'obtenir des classes non sphériques mais toutes de même forme.

### 8.6.7 Cas général, proportion identique

La matrice de variance de chaque classe a maintenant la forme  $\Sigma_k$ . L'affectation des individus aux classes se fait alors en utilisant la distance  $d_{\Sigma_k^{-1}}^2(\mathbf{x}_i, \mu_k)$  et la fonction  $F$  devient

$$F(\Sigma_1, \dots, \Sigma_g) = \sum_k n_k (\text{trace}(S_k \Sigma_k^{-1}) + \log |\Sigma_k|).$$

En utilisant le corollaire A.9, on peut montrer que les matrices optimales  $\Sigma_k$  sont alors simplement les matrices de variance de chaque classe  $\Sigma_k = S_k$  et la vraisemblance classifiante maximisée s'écrit

$$L_C(\mathbf{z}, \boldsymbol{\theta}) = -\frac{1}{2} \sum_k n_k (p + \log |S_k|) - n \log g - \frac{np}{2} \log 2\pi.$$

Le critère optimisé est donc le suivant

$$\sum_k n_k \log |S_k|.$$

## 8.7 Mise en œuvre

Il existe plusieurs logiciels permettant d'utiliser les méthodes étudiées dans ce chapitre ; on peut citer en particulier le logiciel MIXMOD<sup>1</sup>. Dans cette section, on passe en revue rapidement les problèmes que peut poser leur mise en œuvre.

### 8.7.1 Choix du modèle et du nombre de classes

Les méthodes de classification automatique sont souvent justifiées de façon heuristique et le choix de la « bonne » méthode ou du « bon » nombre de classes est alors un problème difficile et souvent mal posé. L'utilisation de méthodes de classification s'appuyant sur les modèles de mélange permet de placer ce problème dans le cadre plus général de la sélection de modèles probabilistes.

Dans le cadre bayésien, la recherche du modèle le plus probable conduit à des critères de sélection de modèles très utilisés comme le critère BIC de Schwarz constitués de deux termes : le premier est la vraisemblance qui a tendance à choisir le modèle le plus complexe et le second est une terme de pénalisation, fonction croissante du nombre de paramètres du modèle. On peut citer en particulier le critère ICL qui, en prenant en compte l'objectif de classification, fournit généralement de bonnes solutions.

### 8.7.2 Stratégies d'utilisation

La maximisation du critère de vraisemblance par l'algorithme EM ou de la vraisemblance classifiante par l'algorithme CEM conduit à chaque fois à la construction d'une suite de solutions faisant croître le critère vers un maximum local qui dépend donc de la position initiale choisie par l'algorithme. La stratégie généralement retenue pour obtenir une « bonne » solution consiste à répéter l'algorithme à partir de plusieurs positions initiales et à retenir la meilleure. On pourra se reporter, par exemple, au texte de Biernacki et al. (2003), où des stratégies plus fines et assez performantes, incluant une première phase consistant à lancer de nombreuses fois l'algorithme sans attendre la convergence complète, ont été étudiées.

---

<sup>1</sup><http://www-math.univ-fcomte.fr/mixmod>

# Annexe A

## Quelques résultats

### A.1 Trois minimisations classiques

**Proposition A.1** *La fonction*

$$f(x) = \frac{a}{x} + b \log x, \quad a, x > 0, \quad b > 0$$

atteint son minimum pour  $x = \frac{a}{b}$

*Preuve :*

$$f'(x) = -\frac{a}{x^2} + \frac{b}{x} = \frac{-a + bx}{x^2}$$
$$f''(x) = \frac{2a}{x^3} - \frac{b}{x^2} = \frac{2a - bx}{x^3}$$

En  $x = \frac{a}{b}$ , la dérivée s'annule et la dérivée seconde est positive ; la proposition est donc démontrée.  $\square$

**Proposition A.2** *La fonction*

$$f(x_1, \dots, x_p) = \sum_{i=1}^p a_i \log x_i$$

sous les contraintes  $x_i > 0, a_i > 0 \forall i$  et  $\sum_i x_i = 1$  atteint son maximum pour  $x_i = \frac{a_i}{\sum_j a_j}$

*Preuve :* En utilisant les multiplicateurs de Lagrange, on se ramène à la maximisation de

$$g(x_1, \dots, x_p) = \sum_i a_i \log x_i + \lambda(\sum x_i - 1).$$

L'annulation des dérivées partielles

$$g'_{x_j}(x_1, \dots, x_p) = \frac{a_j}{x_j} + \lambda = 0 \quad \forall j$$

entraîne

$$x_j = -\frac{a_j}{\lambda} \quad \forall j.$$

En utilisant la contrainte  $\sum_j x_j = 1$ , on en déduit  $\lambda = -\sum_j a_j$  et donc le résultat attendu.  $\square$

**Proposition A.3** *La fonction*

$$f(x_1, \dots, x_p) = \sum_{i=1}^p x_i$$

sous les contraintes  $x_i > 0 \forall i$  et  $\prod_i x_i = 1$  atteint son minimum pour  $x_i = 1 \quad \forall i$

*Preuve* : En utilisant les multiplicateurs de Lagrange, on se ramène à la minimisation de

$$g(x_1, \dots, x_p) = \sum_i x_i + \lambda(\prod x_i - 1).$$

L'annulation des dérivées partielles

$$g'_{x_j}(x_1, \dots, x_p) = x_j + \lambda \frac{\prod x_i}{x_j} = x_j + \frac{\lambda}{x_j} = 0 \quad \forall j$$

entraîne

$$x_j^2 = -\lambda \quad \forall j$$

et en utilisant la contrainte  $\prod_j x_j = 1$ , on en déduit  $\lambda = -1$  et donc le résultat attendu.  $\square$

## A.2 Minimisations matricielles

**Proposition A.4** *La matrice symétrique définie positive  $M$  de dimension  $p \times p$  et de déterminant  $|M| = 1$  minimisant  $\text{trace}(M)$  est la matrice identité  $I$  et la valeur minimisée est égale à  $p$ .*

*Preuve* : Si on note  $\lambda_1, \dots, \lambda_p$  les  $p$  valeurs propres de la matrice symétrique  $M$ , le problème se ramène à la minimisation de  $\sum_i \lambda_i$  sous la contrainte  $\prod \lambda_i = 1$ . Sachant que toutes les valeurs propres sont  $> 0$ , la proposition précédente entraîne donc que les valeurs propres sont toutes égales à 1 et le résultat en découle alors directement.  $\square$

**Corollaire A.5** *La matrice symétrique définie positive  $M$  de dimension  $p \times p$  et de déterminant  $|M| = 1$  minimisant  $\text{trace}(M^{-1})$  est la matrice identité  $I$  et la valeur minimisée est égale à  $p$ .*

*Preuve* : Il suffit de poser  $N = M^{-1}$ . La proposition précédente permet d'en déduire que  $N = I$  et donc le résultat.  $\square$

**Corollaire A.6** *La matrice symétrique  $M$  de dimension  $p \times p$  de déterminant  $|M| = 1$  minimisant  $\text{tr}(QM^{-1})$  où  $Q$  est une matrice symétrique définie positive est*

$$M = \frac{Q}{|Q|^{\frac{1}{p}}},$$

*et la valeur minimisée est égale à  $p|Q|^{\frac{1}{d}}$ .*

*Preuve* : Il suffit de poser  $N = |Q|^{-\frac{1}{p}}QM^{-1}$  et d'utiliser la proposition A.4  $\square$

**Corollaire A.7** *La matrice diagonale  $M$  de dimension  $p \times p$  de déterminant  $|M| = 1$  minimisant  $\text{tr}(QM^{-1})$  où  $Q$  est une matrice symétrique positive est*

$$M = \frac{\text{diag}(Q)}{|\text{diag}(Q)|^{\frac{1}{d}}},$$

*et la valeur minimisée est  $p|\text{diag}(Q)|^{\frac{1}{d}}$ .*

*Preuve* : Si  $M$  est une matrice diagonale,  $M^{-1}$  est aussi une matrice diagonale et nous avons  $\text{tr}(QM^{-1}) = \text{tr}(\text{diag}(Q)M^{-1})$ . Le corollaire précédent permet alors de conclure.  $\square$

**Proposition A.8** *La matrice symétrique  $M$  de dimension  $p \times p$  minimisant  $\text{tr}(M^{-1}) + \alpha \ln |M|$  où  $\alpha$  est un réel positif est  $M = \frac{1}{\alpha}I$ .*

*Preuve* : On peut écrire  $M = d.N$  avec  $d = |M|^p$  et  $|N| = 1$ . nous obtenons alors

$$\text{trace}(M^{-1}) + \alpha \ln |M| = \frac{1}{d} \text{trace}(N^{-1}) + \alpha p \ln d.$$

On peut donc d'abord chercher la matrice  $N$  de déterminant 1 minimisant  $\text{trace}(N^{-1})$ . En utilisant le corollaire A.5, on obtient  $N = I$ . Il suffit alors de minimiser

$$\frac{p}{d} + \alpha p \ln d = p \left( \frac{1}{d} + \alpha \ln d \right).$$

La proposition A.1 permet alors d'en déduire  $d = \frac{1}{\alpha}$  ce qui conduit au résultat attendu.  $\square$

**Corollaire A.9** *La matrice symétrique  $M$  de dimension  $p \times p$  minimisant  $\text{trace}(QM^{-1}) + \alpha \ln |M|$  où  $Q$  est une matrice symétrique positive et  $\alpha$  est un réel positif est  $M = \frac{1}{\alpha}Q$ .*

*Preuve* : En posant  $N = Q^{-1}M$ , on obtient

$$\text{trace}(QM^{-1}) + \alpha \ln |M| = \text{trace}(N^{-1}) + \alpha \ln |N| + cste.$$

En utilisant la proposition précédente, on obtient  $N = \frac{1}{\alpha}I$  ce qui conduit au résultat attendu.  $\square$

**Corollaire A.10** *La matrice diagonale  $M$  de dimension  $p \times p$  minimisant  $\text{trace}(QM^{-1}) + \alpha \ln |M|$  où  $Q$  est une matrice symétrique positive et  $\alpha$  est un réel positif est  $M = \frac{1}{\alpha} \text{diag}(Q)$ .*

*Preuve* : Si  $M$  est une matrice diagonale,  $M^{-1}$  est aussi une matrice diagonale et nous avons  $\text{trace}(QM^{-1}) = \text{tr}(\text{diag}(Q)M^{-1})$ . Le corollaire précédent permet alors de conclure.  $\square$



## Annexe B

# Outils d'algèbre linéaire

### B.1 Espace vectoriel

#### Définitions

Dans cet 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 réels.

#### L'espace $\mathbb{R}^p$

Le produit cartésien  $\mathbb{R}^p$ , ensemble de tous les  $p$ -uplets de réels, est un exemple d'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  $\mathbb{R}^p$ . Les éléments de  $\mathbb{R}^p$  sont notés dans la suite sous la forme de « vecteurs colonnes » ou matrice de dimension  $(p, 1)$  :

$$x = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{pmatrix}$$

et la base canonique est formée des vecteurs

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ 0 \end{pmatrix}, \dots, e_p = \begin{pmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ 1 \end{pmatrix}.$$

La décomposition de  $x$  s'écrit donc  $x = \sum_{i=1}^p x_i e_i$ . Il y a donc identité entre les coordonnées de  $x$  dans la base canonique et les composantes de  $x$ , élément du produit cartésien  $\mathbb{R}^p$ . Ceci n'est vrai que pour la base canonique.

#### Décomposition en somme directe

**Définition B.1** *Un espace vectoriel  $E$  est somme directe des sous-espaces vectoriels  $E_1, \dots, E_k$  si et seulement si  $\forall x \in E$ ,  $x$  s'écrit de manière unique  $x = x_1 + \dots + x_k$  avec  $x_i \in E_i$ .*

On note  $E = E_1 \oplus \dots \oplus E_k$ . Lorsque le nombre de sous-espaces de la somme directe se réduit à deux, on parle de sous-espaces supplémentaires.

## B.2 Applications linéaires et matrices

### Application linéaire

**Définition B.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 :

$$\begin{aligned} \forall x, y \in E & \quad f(x + y) = f(x) + f(y), \\ \forall x \in E \text{ et } a \in \mathbb{R} & \quad f(ax) = af(x). \end{aligned}$$

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  $(e_1, \dots, e_p)$  et  $(f_1, \dots, f_n)$  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  $i$ ème coordonnée de  $f(e_j)$ .

$y = f(x)$  s'écrit alors matriciellement  $y = Ax$ . 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  $y = Ax$  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.

## B.3 Changement de base

### Matrice de changement de base

Si  $(e_j)_{j=1,p}$  et  $(f_j)_{j=1,p}$  sont deux bases d'un espace vectoriel  $E$ , et si  $x$  et  $x'$  sont les vecteurs des coordonnées d'un élément de  $E$  dans ces deux bases, on a la relation :

$$x = Px' \text{ et } x' = P^{-1}x$$

où  $P$  est une matrice carrée de dimension  $p$ , appelée matrice de changement de base. Pour obtenir cette matrice de changement de base, il suffit de ranger en colonne les coordonnées des nouveaux vecteurs de base sur l'ancienne base.

### 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  $(e_j)$ ,  $B$  la matrice associée à  $f$  dans la base  $(f_j)$ , alors on a la relation  $B = P^{-1}AP$ . *Preuve* : Soit  $a$  un élément de  $E$ . Notons  $x$  et  $y$  les coordonnées de  $a$  et de  $f(a)$  dans la première base et  $x'$  et  $y'$  les coordonnées des mêmes éléments dans la seconde base, nous avons :

$$y = Py' \quad x = Px' \quad y = Ax$$

d'où

$$Py' = APx \quad Py' = P^{-1}APx' = Bx \quad \text{avec } B = P^{-1}AP.$$

□

## B.4 Vecteurs et valeurs propres d'un endomorphisme

### Définition et propriétés

**Définition B.3** On appelle vecteur propre d'un endomorphisme  $f$  sur  $E$  tout élément  $x$  de  $E$  non nul tel qu'il existe un réel  $\lambda$  vérifiant  $f(x) = \lambda x$ . Ce réel  $\lambda$  est appelé valeur propre associée au vecteur propre  $x$ .

**Proposition B.4** Si  $x$  est un vecteur propre, les vecteurs  $ax$  où  $a$  est un réel non nul sont aussi des vecteurs propres et ont même valeur propre.

**Proposition B.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  $\lambda$  et noté  $E_\lambda$ .

### Recherche des valeurs propres et vecteurs propres

On suppose dans ce paragraphe que  $E$  est de dimension finie. Soient  $e_i$  une base de  $E$  et  $A$  la matrice carrée associée à un endomorphisme  $f$  dans cette base, on a alors :

$$x \text{ vecteur propre de } f \iff x \neq 0 \text{ et } Ax = \lambda x \iff x \neq 0 \text{ et } (A - \lambda I)x = 0$$

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  $\lambda$  doivent donc annuler le déterminant de la matrice  $(A - \lambda I)$ . Il suffit ensuite pour chaque valeur  $\lambda$  réalisant cette condition de trouver les vecteurs  $x$  vérifiant le système  $Ax = \lambda x$ .

### Application : diagonalisation d'une matrice

#### Le problème

Si  $E$  est un espace vectoriel de dimension finie muni d'une base  $(e_j)$  et  $f$  un endomorphisme sur  $E$  dont la matrice associée dans cette base est  $A$ , on cherche une nouvelle base  $(f_j)$  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 B.6).

## B.5 Produit scalaire, norme, distance et orthogonalité

**Définition B.6** On appelle produit scalaire sur un espace vectoriel  $E$  une application de  $E \times E$  dans  $\mathbb{R}$  :

- bilinéaire,
- symétrique :  $\forall x, y \in E \quad \langle x, y \rangle = \langle y, x \rangle,$
- définie :  $\forall x, y \in E \quad \langle x, x \rangle = 0 \Rightarrow x = 0,$
- positive :  $\forall x, y \in E \quad \langle x, x \rangle \geq 0.$

**Expression matricielle** On se place dans l'espace  $\mathbb{R}^p$  muni de sa base canonique. On montre facilement que tout produit scalaire  $\langle x, y \rangle$  s'écrit sous la forme  $x'My$  où  $M$  est une matrice :

$$\begin{array}{ll} \text{- symétrique :} & M' = M, \\ \text{- définie :} & \forall x \in \mathbb{R}^p \quad x'Mx = 0 \Rightarrow x = 0, \\ \text{- positive :} & \forall x \in \mathbb{R}^p \quad x'Mx \geq 0. \end{array}$$

On note souvent  $\langle x, y \rangle_M$  ce produit scalaire. Le produit scalaire habituel correspond à la matrice identité.

**Définition B.7** On appelle norme sur un espace vectoriel  $E$  une application de  $E$  dans  $\mathbb{R}^+$  vérifiant :

$$\begin{array}{ll} \forall x \in E, \forall \lambda \in \mathbb{R} & \|\lambda x\| = |\lambda| \|x\|, \\ \forall x \in E & \|x\| = 0 \Rightarrow x = 0, \\ \forall x, y \in E & \|x + y\| \leq \|x\| + \|y\|. \end{array}$$

**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  $\|x\|_M = \sqrt{\langle x, x \rangle_M}$ .

**Vecteur normé** Un vecteur est normé si sa norme est égale à 1.

**Définition B.8** On appelle distance sur un ensemble quelconque  $A$  une application  $d$  de  $A \times A$  dans  $\mathbb{R}^+$  vérifiant :

$$\begin{array}{ll} \forall x, y \in A & d(x, y) = d(y, x), \\ \forall x, y \in A & d(x, y) = 0 \Leftrightarrow x = y, \\ \forall x, y, z \in A & d(x, y) \leq d(x, z) + d(z, y). \end{array}$$

**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  $d(x, y) = \|x - y\|$  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 :

$$d_M(x, y) = \|x - y\|_M = \sqrt{\langle x - y, x - y \rangle_M} = \sqrt{(x - y)'M(x - y)}$$

## 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  $x$  et  $y$  de  $E$  sont orthogonaux si leur produit scalaire est nul :

$$x \perp y \iff \langle x, y \rangle = 0.$$

**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 :

$$F \perp G \iff (\forall x \in F, \forall y \in G \quad x \perp y)$$

**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  $F$  :

$$F^\perp = \{x \in E / \forall y \in F \quad x \perp y\}$$

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 B.9** *Théorème de Pythagore* Si un élément  $x$  de  $E$  se décompose suivant deux sous-espaces supplémentaires orthogonaux en  $y$  et  $z$ , on a

$$\|x\|^2 = \|y\|^2 + \|z\|^2$$

**Preuve** soit  $E = F \oplus G$ . On a  $x = y + z$  avec  $x \in E$  et  $y \in F$  et  $z \in G$

$$\|x\|^2 = \langle x, x \rangle = \langle y + z, y + z \rangle = \langle y, y \rangle + 2\langle y, z \rangle + \langle z, z \rangle = \|y\|^2 + \|z\|^2.$$

QED

**Généralisation du théorème de Pythagore** Si  $x = x_1 + \dots + x_k$  est la décomposition d'un élément suivant une somme directe de sous-espaces orthogonaux, on a

$$\|x\|^2 = \sum_{i=1}^k \|x^i\|^2.$$

**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  $x_1, \dots, x_p$  sont les coordonnées d'un vecteur  $x$  dans une base orthonormée, on montre facilement, en utilisant le théorème de Pythagore, la relation :

$$\|x\|^2 = \sum_{i=1}^k (x^i)^2.$$

La matrice de passage entre deux bases orthonormées est orthogonale, c'est-à-dire vérifie la relation  $P' = P^{-1}$  (sa transposée est aussi son inverse).

### Définition

Soit  $E = F \oplus G$  une décomposition de  $E$  en deux sous-espaces supplémentaires, la décomposition unique  $x = y + z$  avec  $y \in F$  et  $z \in G$  permet alors de définir deux applications :

- la première, qui associe au vecteur  $x$  de  $E$  le vecteur  $y$  de  $F$ , est appelée projection sur  $F$  parallèlement à  $G$  ;
- la seconde, qui associe au vecteur  $x$  de  $E$  le vecteur  $z$  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 B.10** On appelle projection orthogonale sur un sous-espace vectoriel  $F$  la projection sur  $F$  parallèlement à  $F^\perp$ .

**Proposition B.11**  $F$  étant un sous-espace vectoriel,  $x$  un point quelconque de  $E$ ,  $y$  sa projection orthogonale sur  $F$  et  $t$  un point quelconque de  $F$ , alors on a

$$(x - t) = (x - y) + (y - t).$$

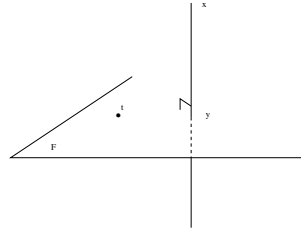


FIG. B.1 – Projection sur un plan

Cette relation représente la décomposition de  $(x - t)$  sur  $F$  et  $F^\perp$  (figure B.1) et le théorème de Pythagore peut donc s'appliquer :

$$\|x - t\|^2 = \|x - y\|^2 + \|y - t\|^2$$

$$d^2(x, t) = d^2(x, y) + d^2(y, t)$$

La quantité  $d^2(y, t)$  étant toujours positive, cette relation permet d'affirmer que  $y$  est l'élément de  $F$  le plus proche de  $x$ . Finalement, les trois relations suivantes sont équivalentes :

$y$  est la projection orthogonale de  $x$  sur  $F$

$$\forall t \in F \quad (x - y) \perp t \text{ ou } (x - y)'Mt = 0$$

$$d(x, y) = \inf\{d(x, t)/t \in F\}$$

La quantité  $d(x, y)$  est souvent appelée "distance" de  $x$  au sous-espace  $F$  et notée  $d(x, F)$ .

### Projection orthogonale sur un sous-espace affine

**Définition B.12 (sous-espace affine)** Si  $G$  est un sous-espace vectoriel de  $E$  et  $a$  un élément de  $E$ , on appelle sous-espace affine l'ensemble  $F$  des éléments  $x$  de  $E$  tels que  $x = a + y$  avec  $y$  élément de  $G$ . On note  $F = a + G$ . 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 B.13 (Sous-espaces affines orthogonaux)** Deux espaces affines sont orthogonaux si les sous-espaces vectoriels qui les définissent le sont.

### Projection orthogonale sur un sous-espace affine

Soit  $F = a + G$  un sous-espace affine,  $x$  un point quelconque de  $E$  et  $H = x + G^\perp$  (fig B.2). On peut montrer que l'intersection de  $H$  et  $F$  est réduite à un seul élément noté ici  $y$ . Cet élément est appelé projection orthogonale de  $x$  sur  $F$ . On peut alors étendre les résultats obtenus précédemment :

1.  $y$  est la projection orthogonale de  $x$  sur  $F$
2.  $\forall t, u \in F \quad (x - y) \perp (t - u)$  ou encore  $(x - y)'M(t - u) = 0$
3.  $d(x, y) = \inf\{d(x, t)/t \in F\}$

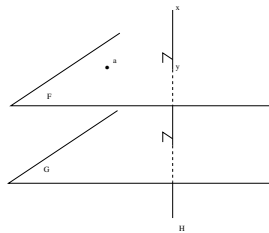


FIG. B.2 – Projection sur un espace affine

La quantité  $d(x, y)$  est appelée "distance" de  $x$  au sous-espace  $F$  et notée  $d(x, F)$ .

## B.6 Matrices symétriques et matrices Q-symétriques

**Proposition B.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 ( $P' = P^{-1}$  ou  $P'P = PP' = I$ ). De plus, si  $B$  est positive alors toutes les valeurs propres sont positives ou nulles.*

**Proposition B.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  $P'QP = PP'Q = QPP' = I$  et  $P'QBP$  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 B.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  $(u_1, \dots, u_p)$  une base Q-orthonormée de vecteurs propres de la matrice  $B$  rangés suivant l'ordre décroissant des valeurs propres  $\lambda_k$  associés, alors :*

- Le vecteur de norme 1 maximisant  $\langle u, Bu \rangle$  est le vecteur  $u_1$  et la valeur maximisée est  $\lambda_1$ .
- $\forall k, 1 < k \leq p$ , le vecteur de norme 1, orthogonal au sous-espace engendré par les vecteurs  $u_1, \dots, u_{k-1}$  maximisant  $\langle u, Bu \rangle$  est le vecteur  $u_k$  et la valeur maximisée est  $\lambda_k$ .



# Bibliographie

- Ball, G. H. and Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science*, 12(2) :153–155.
- Benzecri, J.-P. (1973). *L'analyse des données tome 1 : la taxinomie*. Dunod, Paris.
- Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.
- Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41 :561–575.
- Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. Springer, New York.
- Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–332.
- Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793.
- 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.
- Diday, E. et Collaborateurs (1979). *Optimisation et classification automatique*. INRIA, Rocquencourt.
- Duda, R., Hart, P., and Stork, D. (2001). *Pattern Classification, 2nd Edition*. Wiley Interscience, New York.
- Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer, New York.
- Govaert, G. (2003). *Analyse de données*. Hermes.
- Hathaway, R. J. (1986). Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters*, 4 :53–56.
- 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.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, New York.
- McLachlan, G. J. and Krishnan, K. (1997). *The EM Algorithm*. Wiley, New York.
- McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.
- 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. (1990). *Probabilités, analyse de données et statistique*. 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, Massachusetts.
- 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.