STT 2200
Analyse des données

Notes de cours produites par
Thierry DUCHESNE, Ph.D., P.Stat.
Avec la collaboration de Nadia Ghazzali, Christian Genest, Anne-Sophie Charest et David Émond

Département de mathématiques et de statistique
Université Laval
thierry.duchesne@mat.ulaval.ca

Automne 2017

Avant-propos

Le cours d'analyse des données est offert par le Département de mathématiques et de statistique de l'Université Laval depuis plusieurs années. Les méthodes permettant de résumer et de visualiser des jeux de données de grande taille sans modèle qu'on y enseigne sont toujours d'actualité, et certains pourraient même dire qu'elles gagnent en popularité avec la démocratisation des données massives. Ce cours se veut une introduction à quelques-unes de ces techniques. On y couvrira à la fois les aspects mathématiques/statistiques et les aspects pratiques des méthodes.

Ce document n'est pas un manuel complet que l'étudiant peut suivre, mais plutôt un recueil des supports visuels qui seront présentés en classe dans un format plus lisible. Plusieurs enseignants ont développé ces notes de cours, soit Nadia Ghazzali, Christian Genest, Thierry Duchesne, Anne-Sophie Charest et David Émond.

Comme il s'agit de la première édition de ce document, il y a sans doute matière à amélioration. La critique constructive est bienvenue, donc n'hésitez pas à faire part de tout commentaire ou de toute suggestion à votre enseignant !

Table des matières

Première partie

Préliminaires

Chapitre 0

Introduction et révision

Ce chapitre sert à expliquer les objectifs du cours et à réviser brièvement les notions de base indispensables à la bonne compréhension des méthodes qui y seront étudiées. Bien que la grande majorité des concepts couverts devraient avoir déjà été vus par la plupart des étudiants [1], certains éléments risquent d'être nouveaux pour plusieurs personnes.

0.1 Introduction

Dans ce cours, nous visons à introduire des méthodes qui permettent aux analystes de mener une première étude d'un jeu de données de "haute dimension" ("haute" dans le sens ici où l'on ne peut faire un simple graphique de l'ensemble des observations de toutes les variables) sans avoir recours à un modèle. Les techniques qu'on y enseigne servent donc à réduire la dimension des données, identifier certains liens entre les variables, visualiser les données ou à diviser le jeu de données en groupes/classes. Même si la théorie des probabilités y joue un rôle plus effacé que dans les autres cours de statistique, plusieurs notions demeurent indispensables (fonctions de probabilité/densité marginale, conjointe et conditionnelle, indépendance, corrélation, espérance, variance, loi normale multidimensionnelle). En revanche, on y exploite de façon intensive plusieurs résultats d'algèbre linéaire, en particulier certaines décompositions de matrices. Notez que ceci justifie la présence d'un cours d'introduction à

1. Le masculin est employé dans ces notes à seule fin d'alléger le texte.

la théorie des probabilités et d'un cours d'algèbre linéaire comme préalables (indispensables) à ce cours.

Les techniques étudiées en STT-2200 ne s'appliquent, en principe, qu'à des données structurées (c'est à dire à des cas où nous avons n observations de p variables numériques). Ceci étant dit, avec un peu d'ingéniosité il est possible de recoder sous forme structurée certains ensembles de données non structurées (p.ex. textes, images) de sorte qu'il sera possible d'y appliquer les méthodes vues ici. Les évaluations ne porteront pas sur le passage d'une forme non structurée à une forme structurée, mais sur l'application des méthodes couvertes en classe à la version structurée des données ; une façon de faire le passage des données non structurées aux données structurées sera toujours fournie dans les questions de devoirs, labos ou examens.

L'enseignant s'efforcera, dans la mesure du possible, de montrer comment les méthodes vues en classe peuvent être mises en oeuvre en pratique à l'aide des logiciels R et SAS. Les étudiants pourront utiliser un logiciel **de leur choix** pour effectuer les devoirs, mais ils doivent comprendre que l'enseignant ne sera peut-être pas en mesure de les dépanner s'ils se servent d'un autre logiciel que R ou SAS. Il ne sera pas demandé aux étudiants de programmer ou de donner du code informatique dans les examens, mais ils pourraient être appelés à interpréter des sorties R et SAS.

Avertissement important

Les méthodes vues dans ce cours permettent de détecter des corrélations et **non pas des liens de causalité**. Elles permettent également de ne tirer des conclusions que sur le jeu de données recueilli, et **non pas de les extrapoler à la population de laquelle il provient**. Quelques exemples de telles situations :

— Données du service à la clientèle d'une compagnie : elles donnent un portrait de la minorité de clients qui appellent à ce service, pas de la grande majorité des gens qui n'appellent pas ! Les distributions, sous-groupes, corrélations, etc. détectées dans un tel jeu de données ne s'appliquent (probablement) pas à l'ensemble de la population des clients.

— On se rend compte que bien des gens en surplus de poids ont une forte consommation

de liqueur diète. Il ne faut pas en déduire que la liqueur diète cause le surplus de poids.

— Autre exemple similaire : on analyse une base de données sur les feux de forêts et on se rend compte que l'utilisation de bombardiers d'eau (p.ex. avion CL-415) est corrélée avec la taille que les feux atteignent. Il ne faut pas déduire que l'utilisation des bombardiers d'eau fait croître les feux !

— En analysant les données sur un site web de vente de maisons, on conclut qu'avoir plus de salles de bain augmente le temps requis pour vendre sa maison. Deux phénomènes possibles ici : (i) on infère de la causalité alors qu'il ne s'agit que d'une corrélation ; (ii) si on recueille les données à partir du site web, alors les maisons qui prennent plus de temps à vendre y apparaissent plus longtemps et sont sur-représentées dans l'échantillon.

0.2 Concepts d'algèbre linéaire

Nous ne révisons ici que les principaux résultats qui seront repris en cours de session. Une revue exhaustive des notions d'algèbre linéaire utiles en statistique est donné par Harville (2008), tandis qu'un résumé plus succint est disponible dans l'annexe C du livre sur l'apprentissage automatisé (*machine learning*) de Bishop (2006). Et bien sûr une recherche sur le web vous mènera à une grande quantité de pages où sont recensées la majorité des définitions et formules utiles dans ce cours ... ainsi que plusieurs autres.

0.2.1 Calcul matriciel

Voici un rappel de quelques identités qui nous serons utiles.

PROPOSITION **0.1** *Soit* $\mathbf{M}$, $\mathbf{N}$ *et* $\mathbf{P}$ *des matrices,* $\mathbf{A}$ *et* $\mathbf{B}$ *des matrices carrées,* $\mathbf{v}$ *et* $\mathbf{w}$ *des vecteurs colonne et* $\mathbf{I}_n$ *la matrice identité de dimension* $n \times n$.

1. $(\mathbf{M}^\top)^\top = \mathbf{M}$, $(\mathbf{MN})^\top = \mathbf{N}^\top\mathbf{M}^\top$, $(\mathbf{N}+\mathbf{M})^\top = \mathbf{N}^\top + \mathbf{M}^\top$;
2. $\mathbf{A}$ *est symétrique* $\Leftrightarrow \mathbf{A}^\top = \mathbf{A}$;

3. $\mathbf{A}\mathbf{A}^\top$ *et* $\mathbf{A}^\top\mathbf{A}$ *sont symétriques ;*
4. $\mathbf{P}(\mathbf{M}+\mathbf{N}) = \mathbf{P}\mathbf{M}+\mathbf{P}\mathbf{N}$ *;*
5. $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$ *;*
6. $\mathbf{A}$ *est symétrique* $\Leftrightarrow$ $\mathbf{A}^{-1}$ *est symétrique ;*
7. $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ *;*
8. $(\mathbf{A}+\mathbf{M}\mathbf{B}^{-1}\mathbf{N})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{M}(\mathbf{B}+\mathbf{N}\mathbf{A}^{-1}\mathbf{M})^{-1}\mathbf{N}\mathbf{A}^{-1}$ *(identité de Woodbury, très utile pour réduire le temps de calcul si* $\mathbf{A}$ *est diagonale mais de grande dimension)*
9. $|\mathbf{A}^\top| = |\mathbf{A}|$, $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$, $|\mathbf{A}\mathbf{B}| = |\mathbf{A}||\mathbf{B}|$ *;*
10. *Si* $\mathbf{M}$ *et* $\mathbf{N}$ *sont* $n \times m$*, alors* $|\mathbf{I}_n + \mathbf{M}\mathbf{N}^\top| = |\mathbf{I}_m + \mathbf{M}^\top\mathbf{N}|$ *(un cas particulier utile :* $|\mathbf{I}_n + \mathbf{v}\mathbf{w}^\top| = |1 + \mathbf{v}^\top\mathbf{w}|$*) ;*
11. $tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$ *;*
12. $tr(\mathbf{M}\mathbf{N}) = tr(\mathbf{N}\mathbf{M})$, $tr(\mathbf{M}\mathbf{N}\mathbf{P}) = tr(\mathbf{P}\mathbf{M}\mathbf{N}) = tr(\mathbf{N}\mathbf{P}\mathbf{M})$.

0.2.2 Décompositions d'une matrice

Il existe plusieurs façons de décomposer une matrice que nous exploiterons tout au long du cours. Avant de procéder, quelques définitions seront utiles.

DÉFINITION **0.1** *Soit* $\mathbf{A}$ *une matrice* $n \times n$*. Alors*

- $\mathbf{A}$ *est* définie non-négative *si* $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$ *pour tout vecteur* $\mathbf{x} \in \mathbb{R}^n$ *;*
- $\mathbf{A}$ *est* définie positive *si elle est définie non-négative et* $\mathbf{x}^\top\mathbf{A}\mathbf{x} = 0$ *seulement quand* $\mathbf{x} = \mathbf{0} \Leftrightarrow$ *si* $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$ *pour tout vecteur* $\mathbf{x} \neq 0$ *;*
- $\mathbf{A}$ *est* définie semi-positive *si elle est définie non-négative mais elle n'est* **pas** *définie positive (donc* $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$ *pour tout vecteur* $\mathbf{x} \in \mathbb{R}^n$ *et il existe* $\mathbf{x} \neq \mathbf{0}$ *tel que* $\mathbf{x}^\top\mathbf{A}\mathbf{x} = 0$*) ;*
- $\mathbf{A}$ *est* orthogonale *si elle est non-singulière et* $\mathbf{A}^{-1} = \mathbf{A}^\top$.

Une grande proportion des matrices avec lesquelles nous travaillerons seront symétriques et définies non-négatives (voire même définies positives). Dans ces situations, des décompositions utiles de ces matrices sont possibles.

THÉORÈME **0.2** *Soit* $\mathbf{A}$ *une matrice carrée symétrique définie positive. Alors il existe une unique décomposition telle que* $\mathbf{A} = \mathbf{LDU}$ *où* $\mathbf{L}$ *est une matrice triangulaire inférieure unitaire*[2]*,* $\mathbf{U}$ *est une matrice triangulaire supérieure unitaire et* $\mathbf{D}$ *est une matrice diagonale. De plus, les éléments sur la diagonale principale de* $\mathbf{D}$ *sont tous positifs.*

Un corollaire du théorème 0.2 est que dans ce cas il existe une décomposition $\mathbf{A} = \mathbf{U}^\top\mathbf{DU}$ où $\mathbf{D}$ et $\mathbf{U}$ sont telles que définies dans le théorème. Si on prend $\mathbf{T} = \mathbf{D}^{1/2}\mathbf{U}$, alors on peut écrire $\mathbf{A} = \mathbf{T}^\top\mathbf{T}$, et cette dernière décomposition s'appelle **décomposition de Choleski**. Ces résultats tiennent aussi quand $\mathbf{A}$ est définie non-négative, mais dans ce cas tout ce que l'on peut dire des éléments sur la diagonale de $\mathbf{D}$ est qu'ils sont non-négatifs.

Les valeurs et vecteurs propres d'une matrice vont jouer un rôle majeur dans ce cours.

DÉFINITION **0.2** *Soit* $\mathbf{A}$ *une matrice carrée. Alors on dit que* λ *est une* valeur propre *de* $\mathbf{A}$ *s'il existe un vecteur* $\mathbf{x} \neq \mathbf{0}$ *tel que*

$$\mathbf{Ax} = \lambda\mathbf{x}. \tag{0.1}$$

Le vecteur $\mathbf{x}$ *est appelé* vecteur propre *correspondant à la valeur propre* λ *et l'ensemble des nombres réels* λ *satisfaisant (0.1) est appelé* spectre *de la matrice* $\mathbf{A}$*.*

La proposition suivante donne quelques propriétés intéressantes des valeurs et vecteurs propres.

PROPOSITION **0.3** *Soit une matrice carré* $\mathbf{A}$ *de dimension* $n \times n$*.*

1. *Si* $\mathbf{x}$ *est un vecteur propre de* $\mathbf{A}$ *correspondant à une valeur propre* λ*, alors* $c\mathbf{x}$ *sera également un vecteur propre de* $\mathbf{A}$ *correspondant à* λ*.*
2. *Si* $\mathbf{A}$ *est symétrique et* $\mathbf{x}_1$ *et* $\mathbf{x}_2$ *sont des vecteurs propres correspondant à des valeurs propres différentes de* $\mathbf{A}$*, alors* $\mathbf{x}_1$ *et* $\mathbf{x}_2$ *sont orthogonaux, i.e.,* $\mathbf{x}_1^\top\mathbf{x}_2 = 0$*.*
3. *Si* $\mathbf{A}$ *a comme valeurs propres (réelles, mais pas nécessairement distinctes)* $\lambda_1, \ldots, \lambda_n$*, alors* $|\mathbf{A}| = \prod_{i=1}^n \lambda_i$ *et* $tr(\mathbf{A}) = \sum_{i=1}^n \lambda_i$*.*
4. *Si* $\mathbf{A}$ *est symétrique, toutes ses valeurs propres sont réelles.*

2. Une matrice triangulaire est dite *unitaire* lorsque les éléments sur sa diagonale principale sont tous égaux à 1.

5. *Si* $\mathbf{A}$ *est définie non-négative [définie positive] alors toutes ses valeurs propres sont non-négatives [positives].*
6. *Une matrice symétrique* $\mathbf{A}$ *est définie non-négative [définie positive] si et seulement si toutes ses valeurs propres sont non-négatives [positives].*

Avant de définir une décomposition d'une matrice basée sur ses valeurs et vecteurs propres, la définition suivante est utile.

DÉFINITION **0.3** *Une matrice* $n \times n$ *carrée* $\mathbf{A}$ *est dite* diagonalisable *(par une matrice* $\mathbf{Q}$*) s'il existe une matrice carrée* $n \times n$ *non-singulière* $\mathbf{Q}$ *et une matrice* $n \times n$ *diagonale* $\mathbf{D}$ *telles que*

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D} \Leftrightarrow \mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}. \tag{0.2}$$

Deux résultats particulièrement d'intérêt pour nous suivent.

THÉORÈME **0.4** *Toute matrice carrée symétrique est diagonalisable par une matrice orthogonale* $\mathbf{Q}$.

THÉORÈME **0.5** *Soit une matrice* $n \times n$ *symétrique* $\mathbf{A}$ *et ses* n *valeurs propres* $\lambda_1, \ldots, \lambda_n$. *Alors il existe une matrice orthogonale* $\mathbf{Q}$ *telle que*

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\top}, \tag{0.3}$$

où $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$.

On appelle parfois cette décomposition de la matrice $\mathbf{A}$ *décomposition spectrale* ou *décomposition en valeurs propres*. S'en suit du théorème 0.5 que si $\mathbf{A}$ admet n valeurs propres positives distinctes (par exemple si $\mathbf{A}$ est définie positive et de plein rang) et que $\mathbf{\Lambda}$ est telle que définie dans le théorème, alors on peut prendre $\mathbf{Q}$ comme la matrice dont la k-ème colonne est le vecteur propre normé correspondant à la k-ème valeur propre λ_k.

Dans certaines situations, nous ne travaillerons pas avec des matrices carrées ou des matrices non-sinuglières. Dans ces cas, la décomposition spectrale du théorème 0.5 n'est pas définie. Une décomposition plus générale appelée *décomposition en valeurs singulières* peut être une alternative intéressante.

THÉORÈME **0.6** *Soit* $\mathbf{M}$ *une matrice* **quelconque** *de dimension* $m \times n$ *et de rang* r*. Alors il existe des matrices orthogonales carrées* $\mathbf{U}$ *de dimension* $m \times m$ *et* $\mathbf{V}$ *de dimension* $n \times n$ *telles que*

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top, \tag{0.4}$$

où $\mathbf{D}$ *est une matrice* $m \times n$ *telle que* $[\mathbf{D}]_{i,i} = s_i > 0$*,* $i = 1, \ldots, r$ *et dont tous les autres éléments sont égaux à zéro.*

Les valeurs $s_1, \ldots, s_r$ du théorème 0.6 sont appelées *valeurs singulières* de $\mathbf{M}$. Parfois on appelle les colonnes de la matrice $\mathbf{U}$ *vecteurs singuliers à gauche*, alors que les colonnes de la matrice $\mathbf{V}$ sont les *vecteurs singuliers à droite*. Quelques faits intéressants ...

- Les éléments non nuls de la matrice $\mathbf{D}$ (i.e., les valeurs singulères non nulles de $\mathbf{M}$) sont les racines carrées des valeurs propres non nulles des matrices $\mathbf{M}\mathbf{M}^\top$ et $\mathbf{M}^\top\mathbf{M}$.
- Les colonnes de $\mathbf{U}$ sont des vecteurs propres de $\mathbf{M}\mathbf{M}^\top$. *Ce résultat nous sera* **très utile** *au chapitre 1 ...*
- Les colonnes de $\mathbf{V}$ sont des vecteurs propres de $\mathbf{M}^\top\mathbf{M}$.
- Si $m = n$ et $\mathbf{M}$ est définie non-négative, alors $\mathbf{V} = \mathbf{U}$ et la décomposition en valeurs singulières est également une décomposition spectrale (ou décomposition en valeurs propres).

0.2.3 Autres résultats divers

Nous devrons parfois calculer des dérivées qui impliquent des vecteurs ou matrices.

PROPOSITION **0.7** *Quelques résultats en vrac sur les dérivées impliquant des vecteurs ou matrices.*

1. $\partial \mathbf{w}^\top \mathbf{v} / \partial \mathbf{v} = \mathbf{w}$*, où l'élément en position* i *ici représente la dérivée de* $\mathbf{w}^\top\mathbf{v}$ *par rapport à* $[\mathbf{v}]_i$*.*
2. $\partial \mathbf{v}^\top \mathbf{A}\mathbf{v} / \partial \mathbf{v} = (\mathbf{A} + \mathbf{A}^\top)\mathbf{v}$*.*
3. $\partial(\mathbf{A}\mathbf{B})/\partial u = (\partial \mathbf{A}/\partial u)\mathbf{B} + \mathbf{A}(\partial \mathbf{B}/\partial u)$*, où l'élément en position* (i, j) *ici représente la dérivée de* $[\mathbf{A}\mathbf{B}]_{ij}$ *par rapport à* u*.*
4. $\partial(\mathbf{A}^{-1})/\partial u = -\mathbf{A}^{-1}(\partial \mathbf{A}/\partial u)\mathbf{A}^{-1}$*.*

5. $\partial \ln |\mathbf{A}|/\partial x = tr(\mathbf{A}^{-1}\partial \mathbf{A}/\partial x)$.

6. $\partial tr(\mathbf{AB})/\partial \mathbf{A} = \mathbf{B}^\top$, *où l'élément en position* (i, j) *ici représente la dérivée de* $tr(\mathbf{AB})$ *par rapport à* $[\mathbf{A}]_{ij}$.

7. $\partial tr(\mathbf{ABA}^\top)/\partial \mathbf{A} = \mathbf{A}(\mathbf{B} + \mathbf{B}^\top)$.

8. $\partial \ln |\mathbf{A}|/\partial \mathbf{A} = (\mathbf{A}^{-1})^\top$.

0.3 Vecteurs aléatoires

Nous prenons pour acquis que le lecteur est familier avec les notions de base en théorie des probabilités (variable aléatoire, fonctions de probabilité/densité marginale, conjointe et conditionnelle, indépendance de variables aléatoires, espérance, variance, covariance, corrélation) et nous ne présentons ici qu'une brève généralisation de certains de ces concepts aux vecteurs aléatoires.

0.3.1 Définitions et notation

Dans ce chapitre, nous définissons un *vecteur aléatoire* comme un ensemble fini de variables aléatoires. Nous utiliserons habituellement une lettre majuscule en caractère italique non gras pour dénoter une variable aléatoire (p.ex. X), une lettre majuscule italique en caractère gras pour dénoter un vecteur aléatoire (p.ex. $\boldsymbol{X}$), une lettre minuscule italique non grasse pour dénoter un nombre réel (p.ex. x), une lettre minuscule non italique en caractère gras pour dénoter un vecteur de nombres réels (p.ex. $\mathbf{x}$) et une lettre majuscule non italique en caractère gras ou une lettre grecque en caractère gras pour dénoter une matrice de nombres réels (p.ex. $\mathbf{X}$ ou $\mathbf{\Sigma}$).

Soit $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, un vecteur aléatoire de taille p (p.ex. p mesures prise sur un échantillon d'eau). Soit $\mu_i = E(X_i)$ et $\sigma_i^2 = var(X_i)$, $i = 1, \ldots, p$ et $\sigma_{ij} = cov(X_i, X_j)$,

$i, j = 1, \ldots, p$, $i \neq j$. On définit $E(\boldsymbol{X})$ et $var(\boldsymbol{X})$ ainsi :

$$E(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu} \tag{0.5}$$

$$\begin{aligned} var(\boldsymbol{X}) &= \begin{pmatrix} var(X_1) & cov(X_1, X_2) & \cdots & cov(X_1, X_p) \\ cov(X_2, X_1) & var(X_2) & \cdots & cov(X_2, X_p) \\ \vdots & \ddots & \ddots & \vdots \\ cov(X_p, X_1) & \cdots & cov(X_p, X_{p-1}) & var(X_p) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{p,1} & \cdots & \sigma_{p,p-1} & \sigma_p^2 \end{pmatrix} = \boldsymbol{\Sigma}. \end{aligned} \tag{0.6}$$

(0.7)

On peut aussi définir la matrice des écarts-types $\boldsymbol{\Delta} = \text{diag}(\sigma_1, \ldots, \sigma_p)$ et la matrice des corrélations

$$\begin{aligned} cor(\boldsymbol{X}) &= \begin{pmatrix} 1 & cor(X_1, X_2) & \cdots & cor(X_1, X_p) \\ cor(X_2, X_1) & 1 & \cdots & cor(X_2, X_p) \\ \vdots & \ddots & \ddots & \vdots \\ cor(X_p, X_1) & \cdots & cor(X_p, X_{p-1}) & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,p} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,p} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{p,1} & \cdots & \rho_{p,p-1} & 1 \end{pmatrix} = \mathbf{R}. \end{aligned} \tag{0.8}$$

(0.9)

De ces définitions, on peut déduire plusieurs propriétés et identités.

PROPOSITION **0.8** *Soit $\boldsymbol{X}$ un vecteur aléatoire de moyenne $E(\boldsymbol{X}) = \boldsymbol{\mu}$ et de variance $var(\boldsymbol{X}) = \boldsymbol{\Sigma}$, et soit $\mathbf{M}$, $\mathbf{v}$ et $\mathbf{w}$ des matrice et vecteurs de constantes.*

1. $\boldsymbol{\Sigma}$ *est définie non-négative et symétrique.*
2. $\boldsymbol{\Sigma} = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top] = E(\boldsymbol{X}\boldsymbol{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$, *où l'espérance d'une matrice est simplement la matrice des espérances des variables formant la matrice.*
3. $E(\mathbf{M}\boldsymbol{X} + \mathbf{v}) = \mathbf{M}\boldsymbol{\mu} + \mathbf{v}$.
4. $var(\mathbf{M}\boldsymbol{X} + \mathbf{v}) = var(\mathbf{M}\boldsymbol{X}) = \mathbf{M}\boldsymbol{\Sigma}\mathbf{M}^\top$.
5. $\boldsymbol{\Sigma} = \boldsymbol{\Delta}\mathbf{R}\boldsymbol{\Delta} \Leftrightarrow \mathbf{R} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1}$.

Plusieurs lois de probabilité existent pour les vecteurs aléatoires. Dans ce cours, nous ne considérerons que la loi normale multidimensionnelle.

DÉFINITION **0.4** *On dit qu'un vecteur aléatoire* $\boldsymbol{X}$ *de dimension* p *suit une loi normale multidimensionnelle de moyenne* $\boldsymbol{\mu}$ *et de matrice de variance* $\boldsymbol{\Sigma}$, *dénoté* $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *si sa densité est donnée par*

$$f_{\boldsymbol{X}}(\mathbf{x}) = (2\pi)^{-p/2}(|\boldsymbol{\Sigma}|)^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, x \in \mathbb{R}^p. \tag{0.10}$$

0.3.2 Estimation

En pratique, nous ne connaissons pas les valeurs de $\boldsymbol{\mu}$ ou $\boldsymbol{\Sigma}$ et nous les estimons à partir d'un échantillon. Soit $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\top$, $i = 1, \ldots, n$, n réalisations indépendantes d'un vecteur aléatoire $\boldsymbol{X}$ de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$. Alors on estime $\boldsymbol{\mu}$ par la *moyenne échantillonnale* des $\boldsymbol{X}_i$,

$$\bar{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \tag{0.11}$$

et on estime $\boldsymbol{\Sigma}$ par la *variance échantillonnale* des $\boldsymbol{X}_i$,

$$\boldsymbol{S}^2 := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top. \tag{0.12}$$

Si on pose $\boldsymbol{D} = \{\text{diag}(\boldsymbol{S}^2\}^{1/2}$, soit la matrice des écarts-types échantillonnaux, alors on peut calculer la matrice des corrélations échantillonnales :

$$\boldsymbol{R} := \boldsymbol{D}^{-1}\boldsymbol{S}^2\boldsymbol{D}^{-1}. \tag{0.13}$$

Deuxième partie

Méthodes non supervisées

Analyse descriptive/non supervisée

Dans la première partie du cours, nous nous concentrons sur les méthodes dites “descriptives” ou “non supervisées”. Ces méthodes se veulent une extension des méthodes vues dans des cours d’introduction à la statistique, comme par exemple les graphiques des paires, les histogrammes, les moyenne, variance ou percentiles échantillonnaux, etc. dans les situations où l’on est en présence d’un nombre de variables trop grand pour que ces méthodes soient utiles.

Les méthodes non supervisées peuvent servir à plusieurs fins. Quelques tâches classiques pour lesquelles ces méthodes sont employées :

— visualiser des jeux de données
— identifier des sous-groupes ou sous-populations (classification non supervisée, regroupement)
— détecter des structures
— compresser les données en perdant le moins d’information possible
— etc.

Chapitre 1

Analyse en composantes principales

1.1 Introduction

L'analyse en composantes principales (ACP) est une méthode qui permet de réduire la dimension d'un jeu de données tout en conservant le plus d'information possible. On l'utilise habituellement lorqu'on a n observations de p variables continues et que p est trop grand pour nos besoins. Quelques exemples d'application :

- visualisation d'un jeu de données ;
- réduction du nombre de variables de p à $p' << p$ pour faciliter la construction de modèles (p.ex. en génétique ou en "text mining", où $p >> n$) ;
- effectuer une "rotation d'axes" pour simplifier la structure de corrélation ;
- compression de données.

Article qui introduit la méthode :

Harold Hotelling (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520.

Idée clé

Visualiser, comprendre, modéliser, classifier, etc. des données sont toutes des tâches beaucoup plus simples à accomplir si le nombre de variables dans un jeu de données est faible. Si

un jeu de données comprend un grand nombre de variables (p.ex. comparer des équipes de hockey sur la base de 6 statistiques de fin de saison, comparer la criminalité entre états sur la base des taux de 7 types de crimes différents, compresser des images formées de 1084×1084 pixels, identifier le nombre de variantes d'un type de tumeur à partir du degré d'expression de millions de gènes), une première question qu'on peut se poser est "Est-il possible de réduire la dimension du problème sans trop perdre d'information ?"

En omettant tout simplement des variables, on risque de perdre beaucoup d'information utile. Une meilleure solution consiste à trouver des combinaisons linéaires des variables en vue de conserver le maximum d'information sur le jeu de données.

1.2 Description mathématique de la méthode

1.2.1 Première composante principale

Soit un vecteur aléatoire composé de p variables $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ ayant comme matrice de variance $var(\boldsymbol{X}) = \boldsymbol{\Sigma}$. On aimerait définir une **première composante principale**

$$Y_1 = \boldsymbol{\alpha}_1^\top \boldsymbol{X} = \sum_{i=1}^{p} \alpha_{1i} X_i, \tag{1.1}$$

de sorte que la variance de Y_1 **est maximale**. L'idée est simple : on désire combiner p variables en une seule, mais en "capturant" la plus grande partie possible de la variabilité.

Technicalité

Il faut d'abord ajouter une contrainte sur $\boldsymbol{\alpha}_1$, puisque sinon on n'aurait qu'à prendre des α_{1i} qui sont égaux à $\pm\infty$ et on aurait $var(Y_1) = +\infty$, ce qui est définitivement maximal ! On verra qu'il est pratique de contraindre $\boldsymbol{\alpha}_1$ de sorte qu'il ait une **norme égale à 1**.

Calcul de la première composante principale

De la proposition 0.8, on a que

$$var(Y_1) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1.$$

Le problème est donc de maximiser

$$F(\boldsymbol{\alpha}_1) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 \tag{1.2}$$

sous la contrainte que $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$. On peut récrire ce problème à l'aide des multiplicateurs de Lagrange [1], soit maximimser

$$F(\boldsymbol{\alpha}_1, \lambda) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 - 1), \tag{1.3}$$

où λ est un multiplicateur de Lagrange.

La solution du problème s'obtient en dérivant par rapport à α_{11}, ..., α_{1p} et λ, ce qui assure que la contrainte $||\boldsymbol{\alpha}_1|| = 1$ est prise en compte, viz.

$$\frac{\partial}{\partial \lambda} F(\boldsymbol{\alpha}_1, \lambda) = 1 - \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 0.$$

Notons

$$\frac{\partial}{\partial \boldsymbol{\alpha}_1} F = \left(\frac{\partial}{\partial \alpha_{11}} F, \ldots, \frac{\partial}{\partial \alpha_{1p}} F \right)^\top.$$

En s'aidant de la proposition 0.7 (bel exercice !), on obtient

$$\frac{\partial}{\partial \boldsymbol{\alpha}_1} F(\boldsymbol{\alpha}_1, \lambda) = 2\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 - 2\lambda\boldsymbol{\alpha}_1 = 0. \tag{1.4}$$

L'équation (1.4) est vraie si et seulement si

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1. \tag{1.5}$$

De la définition 0.2, on déduit de (1.5) que

(i) $\boldsymbol{\alpha}_1$ est un vecteur propre (normé) de $\boldsymbol{\Sigma}$;

(ii) λ est la valeur propre correspondante.

Maintenant on a que

$$var(Y_1) = \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = \lambda.$$

Puisque l'on veut maximiser cette quantité, on conclut que

(i) $\lambda = \lambda_1$, **la plus grande valeur propre de $\boldsymbol{\Sigma}$** ;

(ii) $\boldsymbol{\alpha}_1$, le vecteur propre normé correspondant.

1. Le lecteur qui n'est plus familier avec la notion de multiplicateur de Lagrange trouvera un excellent rappel sur le sujet dans l'annexe du livre de Bishop (2006).

1.2.2 Autres composantes principales

Deuxième composante principale

On poursuit simultanément deux objectifs :
(i) conserver le **maximum de variation** présente dans $\boldsymbol{X}$;
(ii) **simplifier la structure de dépendance** pour faciliter l'interprétation et assurer la stabilité numériques d'éventuelles méthodes qui utiliseront les composantes principales obtenues.

Étant donné Y_1, la **deuxième composante principale** $Y_2 = \boldsymbol{\alpha}_2^\top \boldsymbol{X}$ est définie telle que
(i) $var(Y_2) = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2$ est maximale ;
(ii) $\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 = 1$;
(iii) $cov(Y_1, Y_2) = 0$.

On a que

$$\begin{aligned}\text{cov}(Y_1, Y_2) &= \text{cov}(\boldsymbol{\alpha}_1^\top \mathbf{X}, \boldsymbol{\alpha}_2^\top \mathbf{X}) \\ &= \boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_1 = \lambda_1 \boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1.\end{aligned}$$

On cherche donc le vecteur $\boldsymbol{\alpha}_2$ qui maximise

$$F(\boldsymbol{\alpha}_2, \lambda, \kappa) = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda(\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 - 1) - \kappa(\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1 - 0). \tag{1.6}$$

D'une part, $\boldsymbol{\alpha}_2$ est normé puisque

$$\frac{\partial}{\partial \lambda} F(\boldsymbol{\alpha}_2, \lambda, \kappa) = 1 - \boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 = 0.$$

D'autre part, $\boldsymbol{\alpha}_1$ et $\boldsymbol{\alpha}_2$ sont linéairement indépendants car

$$\frac{\partial}{\partial \kappa} F(\boldsymbol{\alpha}_2, \lambda, \kappa) = -\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_1 = -\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 = 0.$$

Un calcul simple montre que

$$\frac{\partial}{\partial \boldsymbol{\alpha}_2} F(\boldsymbol{\alpha}_2, \lambda, \kappa) = 2\boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - 2\lambda \boldsymbol{\alpha}_2 - \kappa \boldsymbol{\alpha}_1 = 0. \tag{1.7}$$

En multipliant l'équation (1.7) à gauche et à droite par $\boldsymbol{\alpha}_1^\top$, on trouve

$$2\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - 2\boldsymbol{\alpha}_1^\top \lambda \boldsymbol{\alpha}_2 - \kappa \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 0.$$

Mais

$$\boldsymbol{\alpha}_1^\top \boldsymbol{\Sigma} = \lambda_1 \boldsymbol{\alpha}_1^\top, \quad \text{et} \quad \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1,$$

d'où

$$2\boldsymbol{\alpha}_1^\top \lambda \boldsymbol{\alpha}_2 - 2\boldsymbol{\alpha}_1^\top \lambda \boldsymbol{\alpha}_2 - \kappa \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 0 \implies -\kappa = 0.$$

En substituant ce résultat dans (1.7), on obtient

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 = \lambda \boldsymbol{\alpha}_2,$$

et donc λ est une autre valeur propre de $\boldsymbol{\Sigma}$. Puisque

$$var(Y_2) = \boldsymbol{\alpha}_2^\top \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^\top \lambda \boldsymbol{\alpha}_2 = \lambda,$$

on a que cette variance est maximale si $\lambda = \lambda_2$, la deuxième plus grande valeur propre de $\boldsymbol{\Sigma}$, et conséquemment $\boldsymbol{\alpha}_2$ est le vecteur propre normé correspondant.

Généralisation

En utilisant des maximisations successives, on conclut que

$$\begin{aligned} Y_k &= k^{\text{e}} \text{ composante principale} \\ &= \boldsymbol{\alpha}_k^\top \boldsymbol{X}, \end{aligned}$$

où $\boldsymbol{\alpha}_k$ est le vecteur propre normé associé à λ_k, la k^{e} plus grande valeur propre de $\boldsymbol{\Sigma}$.

1.2.3 Écriture matricielle

Matrice des composantes principales

Pour définir **simultanément** et de façon plus compacte les composantes principales, on pose

$$\boldsymbol{Y} = \mathbf{A}^\top \boldsymbol{X}, \tag{1.8}$$

où

$$\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p) = \begin{pmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{pp} \end{pmatrix}.$$

PROPOSITION **1.1** *Quelques popriétés en vrac de* $\mathbf{A}$.

1. *Les colonnes de la matrice* $\mathbf{A}$ *sont les vecteurs propres de* $\boldsymbol{\Sigma}$;
2. $\mathbf{A}^\top \mathbf{A} = \mathbf{A}\mathbf{A}^\top = \mathbf{I}_p$;
3. $\mathbf{A}^\top = \mathbf{A}^{-1}$.
4. $\boldsymbol{\Sigma}\mathbf{A} = \mathbf{A}\boldsymbol{\Lambda}$, *où* $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.
5. $var(\boldsymbol{Y}) = \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A} = \boldsymbol{\Lambda} \Rightarrow cov(Y_i, Y_j) = 0$ *si* $i \neq j$ *et* $var(Y_i) = \lambda_i \geq var(Y_j) = \lambda_j$ *si et seulement si* $i \leq j$.

1.2.4 Variation expliquée

La **trace de Pillai** est une **mesure globale de la variation** :

$$\mathrm{tr}(\boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_{i=1}^{p} \lambda_i. \tag{1.9}$$

La **proportion de variation expliquée par** la composante principale Y_i est

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}.$$

1.3 L'analyse en composantes principales en pratique

1.3.1 Estimation de $\boldsymbol{\Sigma}$

Bien sûr en pratique, la matrice $\boldsymbol{\Sigma}$ est **inconnue**. Comme on l'a vu à la section 0.3.2, à partir d'un échantillon aléatoire $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, elle peut être estimée par

$$\boldsymbol{S}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right) \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top.$$

Règle générale, $\boldsymbol{S}^2$ est inversible ; elle l'est avec probabilité 1 si $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (sous certaines conditions mineures pour $\boldsymbol{\Sigma}$). Elle admet alors la représentation

$$\boldsymbol{S}^2 = \hat{\mathbf{A}}\mathbf{L}\hat{\mathbf{A}}^\top,$$

où $\hat{\mathbf{A}} = (a_{ij})$ et $\mathbf{L} = \mathrm{diag}(\ell_1, \ldots, \ell_p)$.

REMARQUE **1.1** *Si $\ell_1 > \cdots > \ell_p$, alors les composantes principales sont uniques, à un signe près.*

1.3.2 Points faibles et améliorations

Sensibilité à l'échelle de $X_1, \ldots, X_p$

Puisque l'ACP cherche une combinaison qui maximise la variance, une variable X_i à grande variance aura un poids démesuré dans les composantes principales. Par exemple mesure la distance en mètres au lieu dans km rendrait multiplierait la variance de cette variable par 1 million et elle aurait un poids majeur dans toutes les composantes. C'est pourquoi **on recommande très fortement d'effectuer l'ACP sur les variables standardisées**, à moins qu'elles n'aient déjà des variances plutôt similaires.

Définissons

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}; \qquad s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2;$$
$$X_{ij}^* = \frac{(X_{ij} - \bar{X}_j)}{\sqrt{s_j^2}}$$

et $\boldsymbol{X}^*$ la matrice $n \times p$ des X_{ij}^*. Alors la matrice de variance échantillonnale $\boldsymbol{S}^{*2} = n^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*\top}$ calculée à partir des données standardisées est égale à la matrice des corrélations échantillonnales $\hat{\mathbf{R}}$ qui estime $\mathbf{R}$. **Il est donc fortement recommandé d'effectuer l'ACP à partir de la matrice des corrélations, et non pas à partir de la matrice de variance**.

Si p est TRÈS grand ?

Les matrices $\boldsymbol{S}^2$ ou $\hat{\mathbf{R}} = n^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*\top}$ sont de dimension $p \times p$. Si $p \leq n$, alors ces matrices seront de rang (au plus) p avec (au plus) p valeurs propres positives. Mais dans plusieurs champs d'applications, $p >> n$ ("*fat data*"), comme par exemple en génétique ou en "text mining", où p peut facilement excéder 10 000. Dans ces cas, nous sommes aux prises avec deux problèmes : (i) les matrices de variance et de corrélation auront n valeurs propres positives et un très grand nombre $p - n$ de valeurs propres égales à 0 dont nous n'avons pas vraiment besoin ; (ii) mais le problème majeur sera de stocker une matrice $p \times p$ en mémoire et demander à notre pauvre ordinateur d'en calculer les valeurs et vecteurs propres !

La solution est d'utiliser **la décomposition en valeurs singulières**. En effet, on va passer par $\boldsymbol{X}^{*\top}\boldsymbol{X}^*$ (qui est de dimension $n \times n << p \times p$) pour aller chercher les (au plus n) valeurs propres non nulles de $n^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*\top}$ et les vecteurs propres qui lui sont associés. Du théorème 0.6 on a

$$\begin{aligned} n^{-1}\boldsymbol{X}^{*\top}\boldsymbol{X}^* &= n^{-1}(\mathbf{UDV}^\top)^\top(\mathbf{UDV}^\top) \\ &= n^{-1}\mathbf{VDU}^\top\mathbf{UDV}^\top = \mathbf{V}(n^{-1}\mathbf{D}^2)\mathbf{V}^\top \end{aligned}$$

et similairement

$$\begin{aligned} n^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*\top} &= n^{-1}(\mathbf{UDV}^\top)(\mathbf{UDV}^\top)^\top \\ &= n^{-1}\mathbf{UDV}^\top\mathbf{VDU}^\top = \mathbf{U}(n^{-1}\mathbf{D}^2)\mathbf{U}^\top, \end{aligned}$$

Les n premières valeurs propres que nous cherchons sont n^{-1} fois les carrés des valeurs singulières formant la matrice diagonale $\mathbf{D}$ de la décomposition en valeurs singulières de $\boldsymbol{X}^*$ et les vecteurs propres associés sont les colonnes de la matrice $\mathbf{U}$ de cette décomposition ; les $p - n$ valeurs propres restantes sont égales à zéro et n'expliquent donc aucune variabilité.

$\Rightarrow$ **En pratique dans les cas où p est de grande valeur, on effectue une décomposition en valeurs singulières de $\boldsymbol{X}^*$ en demandant au logiciel d'utiliser une matrice $\mathbf{U}$ de dimension $n \times r$, où r est le nombre de composantes principales désirées (valeur de votre choix de 1 à n). Les valeurs propres $\ell_1, \ldots, \ell_r$ de l'ACP seront n^{-1} fois les carrés des valeurs singulières obtenues et les vecteurs propres normés correspondants seront les colonnes de la matrice $\mathbf{U}$.**

ACP comme une première étape dans une analyse prédictive

Dans certaines applications il arrive que l'ACP soit effectuée parce que l'on va essayer de prédire la valeur d'une variable V à partir des valeurs de variables $X_1, \ldots, X_p$ mais que p est tout simplement trop grand. Dans ces cas on applique l'ACP pour obtenir les $k << p$ premières composantes principales $Y_1, \ldots, Y_k$, et ce sont ces variables qui seront utilisées pour prédire V.

- C'est une façon raisonnable de faire dans bien des situations, puisque les composantes principales vont retenir la majeure partie de l'information contenue dans les variables originales. D'ailleurs les grands gagnants de concours Kaggle (p.ex. Walmart en 2014) ont souvent recours à cette stratégie.
- Il arrive par contre parfois qu'une variable X_j à faible variabilité soit fortement associée à V et qu'elle serait très utile si le but est de prédire V. Mais comme la variabilité de X_j est faible, son poids dans les composantes principales sera négligeable. Dans ce type de situation, l'ACP n'est peut-être pas idéale.
- Si le but de l'ACP est de réduire la dimension afin de prédire une autre variable, alors il en existe des variantes intéressantes :
  - l'analyse canonique des corrélations (*canonical correlation analysis*, voir Kuhn and Johnson (2013))
  - la méthode des moindres carrés partiels (*partial least squares*, voir James et al. (2014), Kuhn and Johnson (2013))
  - la régression à rang réduit (*reduced rank regression*, voir James et al. (2014)).

1.3.3 Aspects géométriques

Puisque $\mathbf{A}$ est orthonormale, $\boldsymbol{Y} = \mathbf{A}^\top \boldsymbol{X}$ (ou $\boldsymbol{Y} = \mathbf{A}^\top \boldsymbol{X}^*$, tout dépendant de la matrice utilisée pour effectuer l'ACP) représente une rotation d'axes. Les nouveaux axes correspondent aux directions orthogonales de variation maximale successives, en supposant toujours $\ell_1 > \cdots > \ell_p$. La formule

$$\mathbf{Y}_i = \mathbf{A}^\top \boldsymbol{X}_i$$

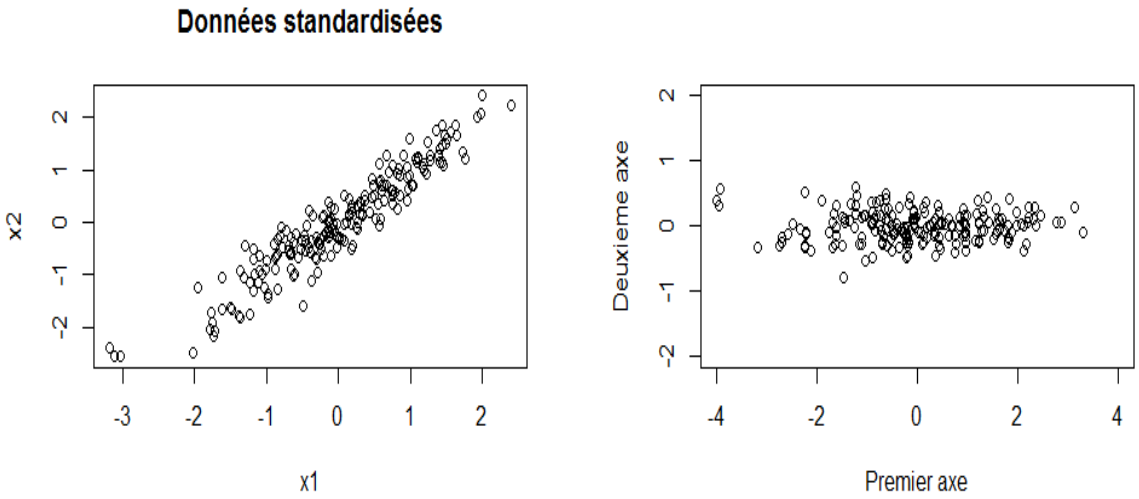

FIGURE 1.1 – Données simulées standardisées (gauche) et données montrées selon les 2 axes principaux (droite)

donne les coordonnées de l'observation $\boldsymbol{X}_i$ dans le nouveau système d'axes. On appelle

$$Y_{ij} = \mathbf{a}_j^\top \mathbf{X}_i = \sum_{k=1}^{p} a_{jk} X_{ik}$$

le **score** de $\boldsymbol{X}_i$ sur l'axe principal j.

EXEMPLE **1.1** *Le graphique de gauche de la figure 1.1 montre 200 observations standardisées de 200 paires de variables aléatoires* (X_{i1}, X_{i2}), $i = 1, \ldots, 200$, *simulées selon une loi normale bidimensionnelle.*

La décomposition en valeurs propres de $\hat{\mathbf{R}}$ *donne comme valeurs propres 1.945 et 0.055 et comme vecteurs propres correspondants* $\mathbf{v}_1^\top = (0.707, 0.707)$ *et* $\mathbf{v}_1^\top = (-0.707, 0.707)$. *La première composante principale est donc* $Y_1 = 0.707x_1^* + 0.707x_2^*$ *et elle explique* $l.395/(l.395 + 0.234) = 97\%$ *de la variabilité. La seconde est* $Y_2 = -0.707x_1^* + 0.707x_2^*$. *On calcule les valeurs de* (Y_1, Y_2) *pour chaque observation et on montre ces valeurs au graphique de droite de la figure 1.1. On constate que l'on a tout simplement effectué une rotation des axes. La première*

composante principale est l'axe le long duquel la variabilité est la plus forte. Ainsi dans cet exemple, même si on ne connaissait que la coordonnée des points le long de la première composante principale, on ne perdrait pas beaucoup d'information.

REMARQUE **1.2** *L'ACP préserve la distance entre les points, puisque* $\hat{\mathbf{A}}^\top = \hat{\mathbf{A}}^{-1}$.

$$\begin{aligned}
||\boldsymbol{Y}_i - \boldsymbol{Y}_j||^2 &= (\boldsymbol{Y}_i - \boldsymbol{Y}_j)^\top (\boldsymbol{Y}_i - \boldsymbol{Y}_j) \\
&= \{\hat{\mathbf{A}}^\top (\boldsymbol{X}_i - \boldsymbol{X}_j)\}^\top \hat{\mathbf{A}}^\top (\boldsymbol{X}_i - \boldsymbol{X}_j) \\
&= (\boldsymbol{X}_i - \boldsymbol{X}_j)^\top \hat{\mathbf{A}}\hat{\mathbf{A}}^\top (\boldsymbol{X}_i - \boldsymbol{X}_j) \\
&= (\boldsymbol{X}_i - \boldsymbol{X}_j)^\top (\boldsymbol{X}_i - \boldsymbol{X}_j) \\
&= ||\boldsymbol{X}_i - \boldsymbol{X}_j||^2,
\end{aligned}$$

EXEMPLE **1.2** *Exemple avec 5-6 variables et dont les 2 premières composantes s'interprètent bien ...*

L'exemple précédent mène à d'autres questions d'ordre pratique :

— La valeur numérique des éléments des vecteurs propres des composantes les plus importantes ont-ils des interprétations pratiques ?

— Si le but de notre ACP est la réduction de la dimension, combien de composantes principales doit-on conserver ?

1.3.4 Interprétation des poids dans les composantes principales

Quand on est chanceux, on peut trouver une signification aux scores dans les premières composantes principales en regardant les poids accordés aux variables originales dans les composantes principales. Par exemple supposons que nous avons une base de données météorologiques et que X_1, X_2 et X_3 représentent respectivement les températures moyennes en janvier, avril et juillet pour certaines villes alors que X_4, X_5 et X_6 représentent la quantité de précipitations durant ces mêmes mois. Supposons que la première composante est données par $Y_1 = -0.3X_1 + 0.02X_2 + 0.9X_3 + 0.001X_4 - 0.02X_5 - 0.002X_6$. Alors les villes ayant un score élevé pour cette composantes sont celles qui ont un climat chaud en juillet et froid en janvier. Si la deuxième composante principale est $Y_2 = 0.03X_1 - 0.02X_2 + 0.01X_3 + 0.5X_4 + 0.4X_5 + 0.3X_6$,

alors les villes ayant un score élevé pour cette composante seront celles qui reçoivent beaucoup de précipitations.

Pour aider à déterminer si un poids α_{ij} a une valeur suffisamment grande pour qu'on se donne la peine de l'interpréter, on peut tester si la variable X_j à laquelle il correspond est corrélée de façon significative avec la composante principale Y_i. (Exercice : Calculez la corrélation théorique entre X_j et Y_i et voyez comment cette corrélation est liée au poids α_{ij}. Vous pouvez supposer que les variables X sont centrées et standardisées.)

1.3.5 Choix du nombre de composantes

Quatre règles du pouce sont communément employées pour déterminer le nombre de composantes principales à conserver.

La règle du 80%

La preimère règle suggère tout simplement de garder le nombre de composantes principales nécessaires pour expliquer 80% de la variabilité globale. Le choix d'un pourcentage de variabilité globale de 80% est somme toute arbitraire.

La règle de Kaiser

Si l'analyse en composantes principales est effectuée avec la matrice des corrélations $n^{-1}\boldsymbol{X}^*\boldsymbol{X}^{*\top}$, garder $Y_k \Leftrightarrow \ell_k \geq 1$. En général puisque la trace de la matrice de corrélation est égale à $1 + \cdots + 1 = p = \ell_1 + \cdots + \ell_p$, alors la moyenne des valeurs propres est $p/p = 1$, et donc on ne conserve que les composantes associées à une valeur propre supérieure ou égale à cette moyenne.

Ceci suggère une règle un peu plus générale. Peu importe la matrice avec laquelle l'ACP est effectuée, garder $Y_k \Leftrightarrow \ell_k \geq \bar{\ell}$, où $\bar{\ell} = (\ell_1 + \cdots + \ell_p)/p$.

La règle de Joliffe

La règle de Joliffe est un peu plus conservatrice, dans le sens où elle est plus réticente à laisser des composantes. Si l'ACP est effectuée avec la matrice des corrélations, cette règle

dit de garder $Y_k \Leftrightarrow \ell_k \geq 0.7$.

Le règle de Cattell

Cette règle est plus pragmatique, dans le sens où elle nous demande de faire le graphique des paires (k, ℓ_k) (parfois appelé "*scree plot*") et ensuite de ne garder que les ℓ_k précédant le pied de "l'éboulis". La justification ici est que tant que ℓ_k décroit, ajouter une composante explique de la variabilité, mais quand cette décroissance cesse, alors l'ajout de composantes additionnelles n'est pas très utile.

La règle de Horn

Si on croit que les $\boldsymbol{X}_i$ pourraient provenir d'une loi normale multidimensionnelle, alors on va

1. simuler une matrice des corrélations sur la base de n observations de la loi $N_p(\mathbf{0}, \mathbf{I}_p)$;
2. extraire les valeurs propres $m_{11}, \ldots, m_{1p}$;
3. répéter les deux étapes précédentes à K reprises pour obtenir K ensembles de valeurs propres $m_{k1}, \ldots, m_{kp}$ pour $1 \leq k \leq K$;
4. calculer $\bar{m}_i = \frac{1}{K} \sum_{k=1}^{K} m_{ki},\ 1 \leq i \leq p$;
5. garder $Y_k \Leftrightarrow \ell_k \geq \bar{m}_k$.

Puisque $\mathbf{I}_p$ a pour valeurs propres $\mu_1 = \cdots = \mu_p = 1$, on s'attend à ce que $\bar{m}_i > 1$ pour environ $p/2$ valeurs. Cette méthode devrait donc garder les $\ell_i > 1$.

EXEMPLE **1.3** *On considère les données sur la criminalité dans les 50 états américains fournies dans l'aide de la procédure* `PRINCOMP` *de SAS. Dans ce jeu de données, on fournit le taux (nombre d'actes par 100 000 habitants durant l'année 1977) de 7 types de crimes (meurtres, viols, vols, assauts, cambriolages, larcins, vols de voitures).*

En effectuant une analyse en composantes principales sur la matrice des corrélations, on

obtient les deux premières composantes principales suivantes :

$$\begin{aligned} Y_1 &= 0.30\,Meurtres + 0.43\,Viols + 0.40\,Vols + 0.40\,Assauts \\ &\quad + 0.44\,Cambriolages + 0.36\,Larcins + 0.30\,Vol.\,Voitures \\ Y_2 &= -0.63\,Meurtres - 0.17\,Viols + 0.04\,Vols - 0.34\,Assauts \\ &\quad + 0.20\,Cambriolages + 0.40\,Larcins + 0.50\,Vol.\,Voitures. \end{aligned}$$

Clairement, un état avec un score élevé sur le premier axe en est un où le taux de criminalité toutes causes est élevé. En contre-partie, un état avec un score élevé sur le second axe en est un où la proportion des crimes qui sont contre la propriété est élevée en comparaison avec la proportion des crimes qui sont sur la personne. Le graphique des états selon leurs scores dans les deux premiers axes principaux est montré à la figure 1.2. Les états à droite du graphique sont ceux où la criminalité générale est élevée et ceux à droite sont les états où la criminalité générale est faible. Les états au haut du graphique sont ceux où une forte proportion des crimes sont contre la propriété alors que ceux au bas sont ceux où une forte proportion des crimes sont contre la personne (surtout des états du sud-est du pays).

Pour déterminer le nombre de composantes à conserver, on peut se référer au graphique des la figure 1.3. La règle du 80% suggère 3 composantes. La règle de Kaiser y va plutôt pour 2 composantes. Les règles de Joliffe et de Cattell, quant à elles, semblent indiquer de conserver 3 composantes.

1.4 Suites et extensions

Les méthodes développées fonctionnent si les variables $X_1, \ldots, X_p$ sont continues (on veut que $\boldsymbol{X}_i$ soit un vecteur dans $\mathbb{R}^p$). Nous verrons au prochain chapitre quoi faire si les X_j sont en fait des variables catégorielles, par exemple des réponses à des questionnaires à choix multiples. Pour certains types d'applications, les gens ont développé des approches astucieuses pour transformer des données catégorielles ou non structurées (p.ex. images, textes, etc.) en un ensemble de vecteurs dans $\mathbb{R}^p$.

EXEMPLE **1.4** *E. Alpaydin et C. Kaynak on mis à la disposition de tous le jeu de données* `optdigits` *sur le site "Machine Learning Repository" de l'Université de Californie à Irvine*

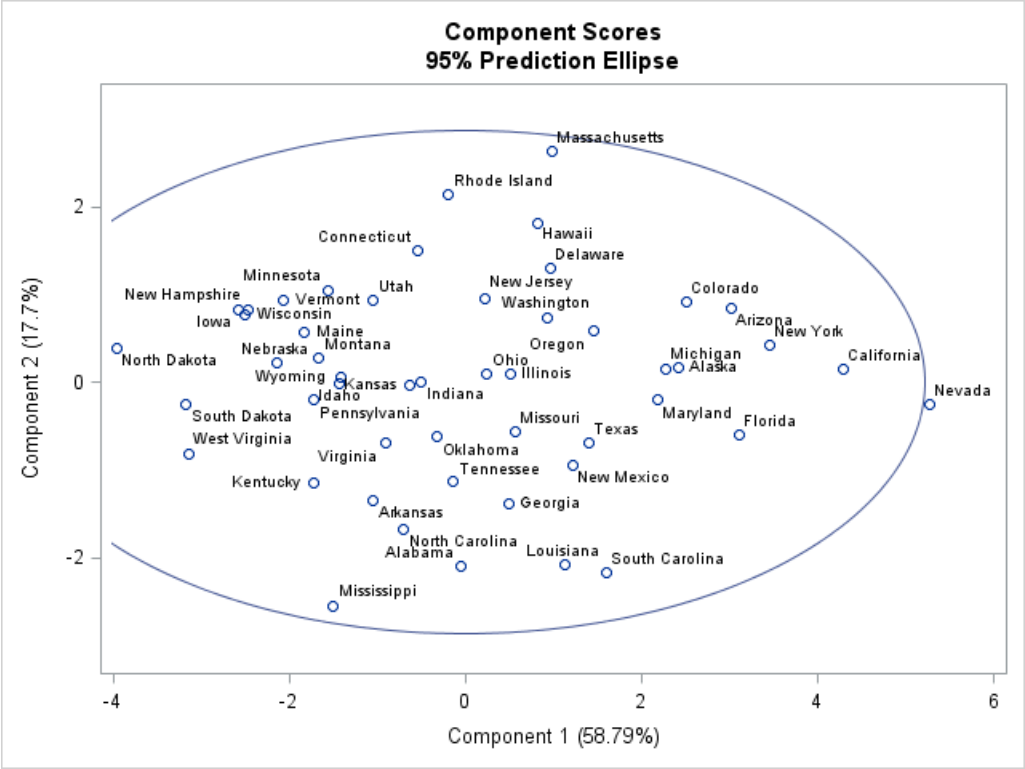

FIGURE 1.2 – Représentation des données sur la criminalité dans les états américains en 1977 selon les deux premiers axes principaux. Graphique tiré de l'aide de SAS sur la procédure `PRINCOMP`.

(Lichman (2013)). Le jeu contient la digitalisation de $n = 1797$ *chiffres de 0 à 9 écrits en noir et blanc à la main par 13 individus. Les données originales consistaient de fichiers bitmap de dimension* 32×32*, mais les données utilisées dans cet exemple sont plutôt le nombre de pixels noirs dans chacun de* $p = 64$ *blocs de dimension* 8×8 *(donc* X_{ij} *est le nombre de pixels noirs dans le* j*-ème bloc de pixels pour le* i*-ème chiffre écrit à la main). Il pourrait être intéressant de voir si les premières composantes principales expliquent suffisamment de variabilité pour permettre de reconnaître les chiffres sans utiliser l'information des 64 variables.*

Le graphique des éboulis de la figure 1.4 montre une variabilité qui diminue très tranquillement en fonction du nombre de composantes principales. D'ailleurs la sortie R ci-dessous

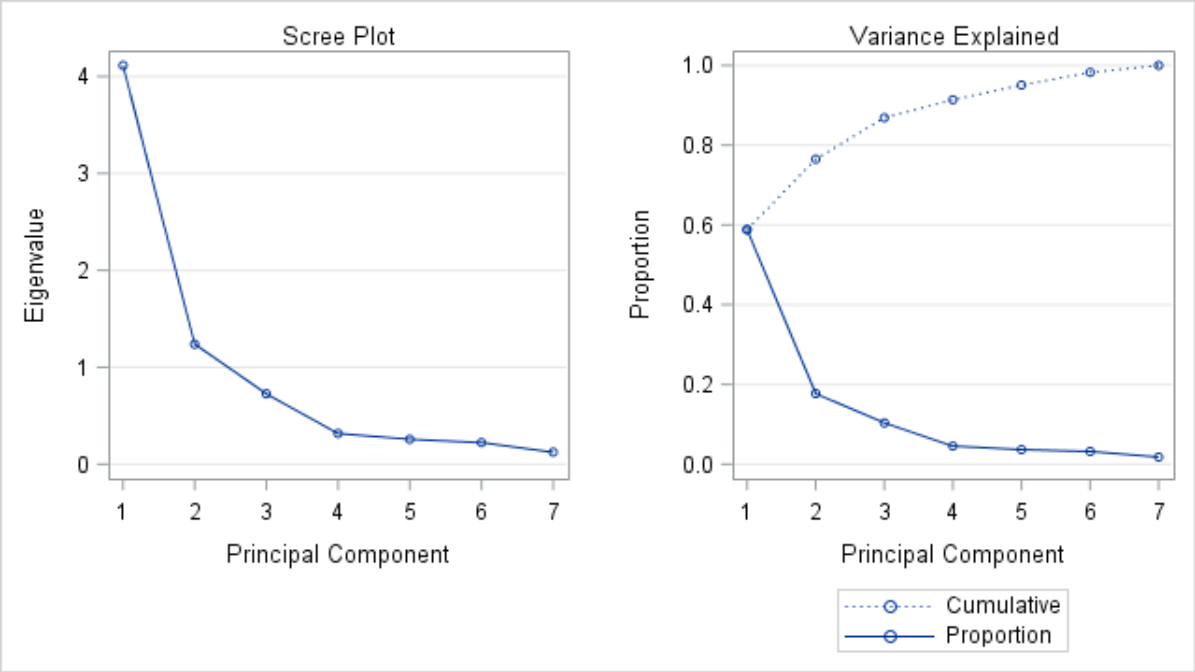

FIGURE 1.3 – Graphiques des éboulis pour les données sur la criminalité dans les 50 états américains en 1977. Graphique tiré de l'aide de SAS sur la procédure `PRINCOMP`.

indique que les 4 premières composantes n'expliquent que 49% de la variabilité.

```
Importance of first k=4 (out of 64) components:
                            PC1      PC2      PC3      PC4
Standard deviation      13.3793 12.7952 11.9075 10.0549
Proportion of Variance   0.1489  0.1362  0.1179  0.0841
Cumulative Proportion    0.1489  0.2851  0.4030  0.4871
```

Ceci étant dit, si on choisit au hasard 500 des 1797 points (pour ne pas surcharger le graphique), qu'on en fait le graphique selon les 2 premières composantes principales en utilisant une couleur correspondant à chacun des chiffres 0 à 9 et que l'on met la valeur du chiffre représenté à côté de chacun des points, on voit que l'on arrive déjà à séparer quelques-uns des chiffres comme le 4, le 6, le 3 et le 1 (voir figure 1.5).

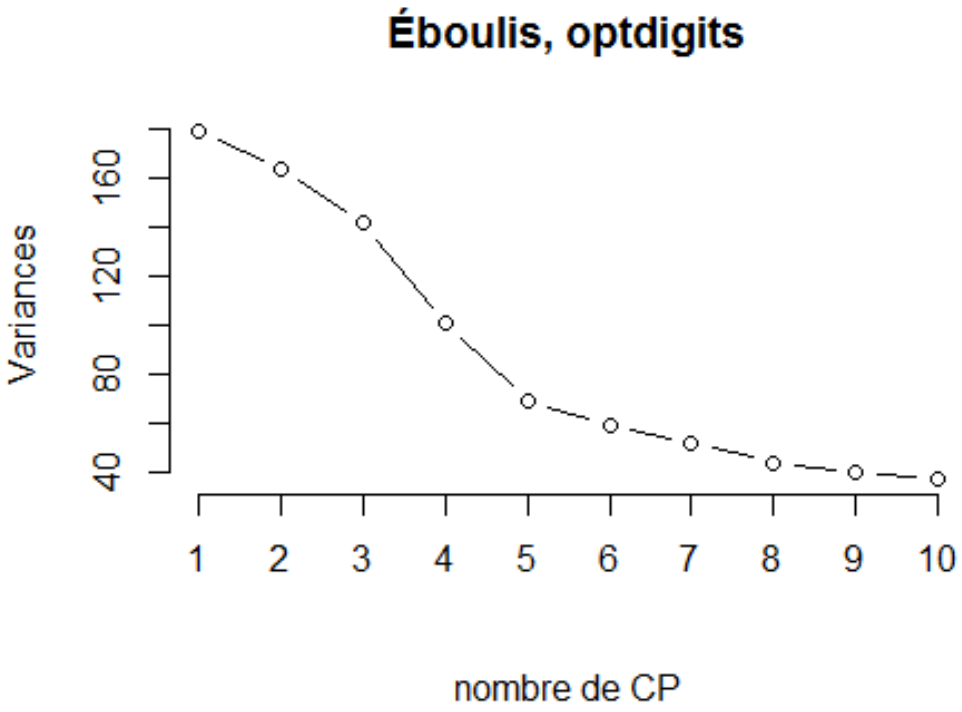

FIGURE 1.4 – Graphique des éboulis pour le jeu de données `optdigits`

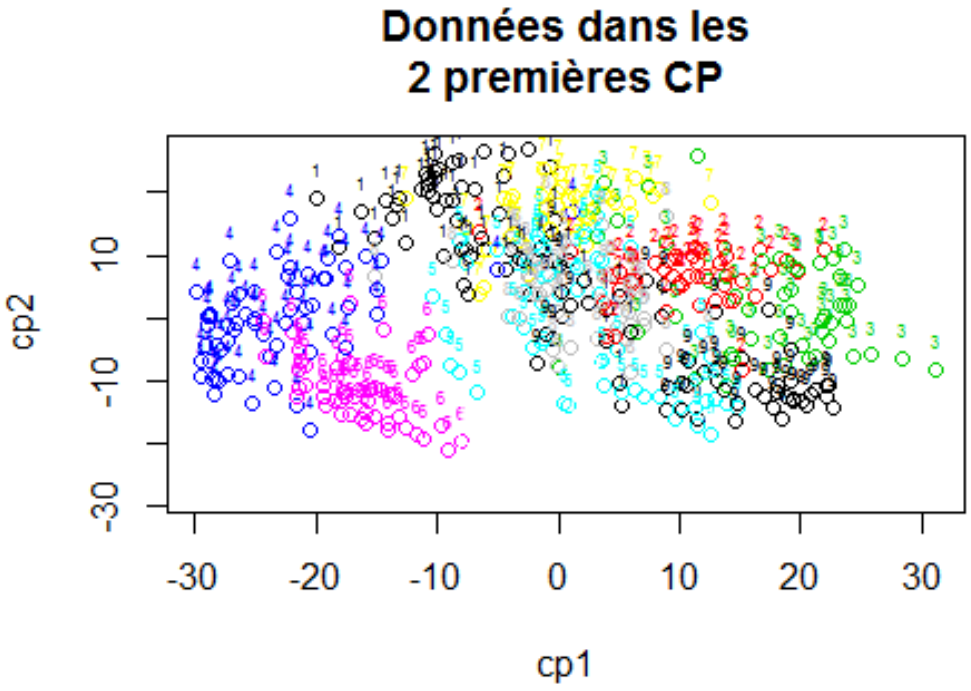

FIGURE 1.5 – 500 observations de `optdigits` selon les 2 premières composantes principales. La couleur et le chiffre à côté de chaque point montrent le chiffre représenté.

Chapitre 2

Données catégorielles : Analyse des correspondances

2.1 Introduction

Au chapitre précédent nous avons que l'ACP pouvait permettre de visualiser, trouver la structure ou réduire la dimension d'ensembles de plusieurs variables continues. Dans ce chapitre, nous verrons comment il est aussi possible de le faire lorsque les variables sont catégorielles, comme par exemple des variables rapportées dans des tableaux de contingences (section 2.2) ou des réponses à des questionnaires à choix multiples (section 2.3).

2.2 Analyse des correspondances binaires

Il est important de noter que la matière à examen est celle des notes projetées en classe. Cette version "imprimable" des notes vient en fait donner plus de théorie et ne contient pas d'exemples.

2.2.1 Introduction

L'analyse des correspondances est une méthode d'origine française. Elle permet de représenter graphiquement des tableaux de fréquences. Certains pourront y voir certaines similitudes avec l'analyse en composantes principales. On verra deux familles de méthodes :

— analyse des correspondances **binaires** : pour des tableaux de fréquences croisant deux variables ;

— analyse des correspondances **multiples** : pour des tableaux faisant intervenir trois variables ou plus.

EXEMPLE **2.1** *On s'intéresse à la relation entre la couleur des yeux et la couleur des cheveux de 592 sujets féminins. Les données sont résumées dans le tableau suivant :*

		CHEVEUX				
		Bruns	*Châtains*	*Roux*	*Blonds*	*Total*
Y	*Marron*	*68*	*119*	*26*	*7*	*220*
E	*Noisette*	*15*	*54*	*14*	*10*	*93*
U	*Verts*	*5*	*29*	*14*	*16*	*64*
X	*Bleus*	*20*	*84*	*17*	*94*	*215*
	Total	*108*	*286*	*71*	*127*	*592*

(Source : Lebart, Morineau & Piron (1995).)

Il serait intéressant de voir si des relations existent entre les modalités des deux variables, aussi de faire une représentation visuelle des données qui fait ressortir les associations les plus fortes.

2.2.2 Description mathématique de la méthode

Concepts et notation

Soit

$$\mathbf{K} = (k_{ij}),$$

le tableau des fréquences où

$$\begin{aligned} k_{ij} &= \text{nombre d'individus appartenant} \\ &\quad \text{à la classe } i \in \{1, \ldots, n\} \text{ et} \\ &\quad \text{à la catégorie } j \in \{1, \ldots, p\}. \end{aligned}$$

Comme ces fréquences sont proportionnelles à la taille d'échantillon, il est plus pertinent de travailler avec le **tableau de fréquences relatives**, Il s'agit du tableau

$$\mathbf{F} = (f_{ij}),$$

dans lequel

$$f_{ij} = \frac{k_{ij}}{k_{\bullet\bullet}} = \frac{k_{ij}}{\sum_{\ell=1}^{n} \sum_{m=1}^{p} k_{\ell m}}\,.$$

EXEMPLE **2.2** *On peut calculer les fréquences relatives pour notre exemple sur la couleur des yeux et des cheveux.*

	Cheveux				
	Bruns	*Châtains*	*Roux*	*Blonds*	*Total*
Marron	*.115*	*.201*	*.044*	*.012*	*.372*
Noisette	*.025*	*.091*	*.024*	*.017*	*.157*
Verts	*.008*	*.049*	*.024*	*.027*	*.108*
Bleus	*.034*	*.142*	*.028*	*.159*	*.363*
Total	*.182*	*.483*	*.120*	*.215*	*1.000*

Nous aurons égelement besoin de la somme des colonnes pour chaque ligne et de la somme des lignes pour chaque colonne, ce qu'on appelle **les marges du tableau** :

$$f_{i\bullet} = \sum_{j=1}^{p} f_{ij} = k_{i\bullet}/k_{\bullet\bullet}\,, \quad 1 \le i \le n\,;$$

$$f_{\bullet j} = \sum_{i=1}^{n} f_{ij} = k_{\bullet j}/k_{\bullet\bullet}, \quad 1 \le j \le p.$$

On peut voir qu'il s'agit en fait des valeurs de la ligne et de la colonne "Total" du tableau de l'exemple 2.2.

Les fréquences relatives estiment des probabilités. On dit que deux variables aléatoires X et Y sont indépendantes quand $P[X = x, Y = y] = P[X = x]P[Y = y]$ pour toutes les paires (x, y). Dans le cas d'un tableau de fréquences croisant deux variables, sous **l'hypothèse d'indépendance** les fréquences relatives devraient être telles qu'on ne s'éloigne pas trop de la relation

$$\forall_{i,j} \quad f_{ij} = f_{i\bullet} f_{\bullet j}.$$

Cette hypothèse est souvent testée à l'aide d'un test du khi-deux. Nous nous servirons de la statistique de ce test pour évaluer notre "distance" de l'hypothèse d'indépendance.

$$X^2 = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{[k_{ij} - E(k_{ij})]^2}{E(k_{ij})}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\left(k_{ij} - \frac{k_{i\bullet} k_{\bullet j}}{k_{\bullet\bullet}}\right)^2}{\left(\frac{k_{i\bullet} k_{\bullet j}}{k_{\bullet\bullet}}\right)}.$$

EXEMPLE **2.3** *Si on effectue le test du khi-deux pour l'exemple sur la couleur des yeux et des cheveux on a*

$$X^2 = k_{\bullet\bullet} \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}$$

$$\approx 138.29.$$

Pour savoir si 138.29 *est une grande valeur ou pas, on peut effectuer un* test d'hypothèse, *où l'hypothèse nulle est que les deux variables sont indépendantes. Dans ce cas précis, on calcule la valeur de p (p-value) à partir d'une loi du khi-deux à* $(n-1)(p-1) = 9$ *degrés de liberté. On obtient une valeur de p de* $P[\chi^2_9 > 138.29] < 0.0001$, *ce qui signifie que si l'hypothèse d'indépendance est vraie, une valeur aussi élevée pour la statistique du khi-deux est très, très peu plausible. On rejette donc l'hypothèse nulle d'indépendance entre les variables et on conclut qu'il semble donc y avoir un lien entre la couleur des yeux et la couleur des cheveux.*

Profils des lignes et colonnes

Les logiciels statistiques permettent d'obtenir la contribution de chaque cellule à la valeur globale de la statistique X^2. Ils permettent aussi d'obtenir le profil de chaque ligne i,

$$L_i = \left(\frac{k_{i1}}{k_{i\bullet}}, \ldots, \frac{k_{ip}}{k_{i\bullet}} \right) = \left(\frac{f_{i1}}{f_{i\bullet}}, \ldots, \frac{f_{ip}}{f_{i\bullet}} \right), \tag{2.1}$$

et le profil de chaque colonne j,

$$C_j = \left(\frac{k_{1j}}{k_{\bullet j}}, \ldots, \frac{k_{nj}}{k_{\bullet j}} \right) = \left(\frac{f_{1j}}{f_{\bullet j}}, \ldots, \frac{f_{nj}}{f_{\bullet j}} \right). \tag{2.2}$$

On peut ensuite définir le profil-ligne moyen est donné par

$$\left(\sum_{i=1}^{n} f_{i\bullet} \, \frac{f_{i1}}{f_{i\bullet}}, \ldots, \sum_{i=1}^{n} f_{i\bullet} \, \frac{f_{ip}}{f_{i\bullet}} \right) = (f_{\bullet 1}, \ldots, f_{\bullet p}). \tag{2.3}$$

De même, le profil-colonne moyen est donné par

$$\left(\sum_{j=1}^{p} f_{\bullet j} \, \frac{f_{1j}}{f_{\bullet j}}, \ldots, \sum_{j=1}^{p} f_{\bullet j} \, \frac{f_{nj}}{f_{\bullet j}} \right) = (f_{1\bullet}, \ldots, f_{n\bullet}). \tag{2.4}$$

En cas d'indépendance, on a

$$\left(\frac{f_{i1}}{f_{i\bullet}}, \ldots, \frac{f_{ip}}{f_{i\bullet}} \right) = (f_{\bullet 1}, \ldots, f_{\bullet p}) \tag{2.5}$$

et

$$\left(\frac{f_{1j}}{f_{\bullet j}}, \ldots, \frac{f_{nj}}{f_{\bullet j}} \right) = (f_{1\bullet}, \ldots, f_{n\bullet}) \tag{2.6}$$

pour tout $i \in \{1, \ldots, n\}$ et $j \in \{1, \ldots, p\}$. La dépendance entre les variables est donc **fonction de la ressemblance entre les profils des lignes et les profils des colonnes**. On peut mesurer et visualiser cette ressemblance de diverses manières, ce qui nous amènera à faire une certaine analogie avec l'ACP.

Distance entre les profils-lignes et entre les profils-colonnes

On peut mesurer la distance entre deux profils-lignes par

$$d^2(i,i') = \sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 .$$

Cependant, cette distance ne tient pas compte de l'importance de chaque colonne. Un choix plus judicieux consiste à prendre la **distance du khi-deux**, qui tient compte de l'importance de chaque colonne.

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 . \tag{2.7}$$

On peut adopter une notation matricielle. Posons

$$D_n = \mathrm{diag}(f_{i\bullet}) \quad \text{et} \quad D_p = \mathrm{diag}(f_{\bullet j}),$$

de sorte que

$$D_n^{-1} F \text{ a pour lignes les profils-lignes}$$
$$D_p^{-1} F^\top \text{ a pour lignes les profils-colonnes.}$$

La distance carrée du khi-deux est de la forme

$$x^\top D_p^{-1} x = \sum_{j=1}^{p} \frac{x_j^2}{f_{\bullet j}}$$

pour un point-ligne $x \in \mathbb{R}^p$ et de la forme

$$x^\top D_n^{-1} x = \sum_{i=1}^{n} \frac{x_i^2}{f_{i\bullet}}$$

pour un point-colonne $x \in \mathbb{R}^n$.

Avantages de la ditance du khi-deux : elle garantit

1. une **équivalence distributionnelle** ;

2. une **relation quasi-barycentrique**[1] entre les deux nuages de points

$$L_1, \ldots, L_n \in \mathbb{R}^p \quad C_1, \ldots, C_p \in \mathbb{R}^n.$$

REMARQUE **2.1** *Quelques constatations ...*

1. *La métrique du khi-deux (entre deux lignes),*

$$\sum_{j=1}^{p} \left(\frac{f_{ij}}{f_{i\bullet}\sqrt{f_{\bullet j}}} - \frac{f_{i'j}}{f_{i'\bullet}\sqrt{f_{\bullet j}}} \right)^2,$$

   s'exprime donc en fonction de la distance euclidienne. De même pour la distance entre deux colonnes.

2. *La distance carrée entre deux lignes reste la même, que ce soit avant ou après l'agrégation de deux colonnes ayant le même profil.*

3. *La distance carrée entre deux colonnes reste la même, que ce soit avant ou après l'agrégation de deux lignes ayant le même profil.*

EXEMPLE **2.4** *Petit exemple numérique pour illustrer les remarques précédentes. Considérons un tableau fictif* 2×3

11	**22**	*16*	*49*
16	**32**	*3*	*51*
27	**54**	*19*	*100*

dont les deux premières colonnes ont le même profil. La distance entre les deux premières lignes est

$$\begin{aligned} d^2(1,2) &= \frac{100}{27}\left(\frac{11}{49} - \frac{16}{51}\right)^2 + \frac{100}{54}\left(\frac{22}{49} - \frac{32}{51}\right)^2 + \\ &\quad \frac{100}{19}\left(\frac{16}{49} - \frac{3}{51}\right)^2 \\ &= 0.088\ 477\ 88. \end{aligned}$$

Maintenant fusionnons les deux premières colonnes :

1. Il existe une relation quasi-barycentrique lorsque les coordonnées des points d'un espace sont proportionnelles aux composantes du facteur de l'autre espace correspondant à la même valeur propre.

11	**22**	*16*	*49*	$\Rightarrow$	**33**	*16*	*49*
16	**32**	*3*	*51*	$\Rightarrow$	**48**	*3*	*51*
27	**54**	*19*	*100*	$\Rightarrow$	**81**	*19*	*100*

La distance entre les deux premières ligne devient

$$D^2(1,2) = \frac{100}{81}\left(\frac{33}{49} - \frac{48}{51}\right)^2 + \frac{100}{19}\left(\frac{16}{49} - \frac{3}{51}\right)^2$$

$$= 0.088\ 477\ 88.$$

La démonstration du phénomène illustré dans cet exemple est assez simple. On suppose que

$$\frac{f_{ij_1}}{f_{\bullet j_1}} = \frac{f_{ij_2}}{f_{\bullet j_2}} = F_i, \quad 1 \leq i \leq n.$$

Si $f_{\bullet j_0} = f_{\bullet j_1} + f_{\bullet j_2}$, alors on a, pour tout i,

$$\frac{f_{ij_0}}{f_{\bullet j_0}} = \frac{f_{ij_1} + f_{ij_2}}{f_{\bullet j_0}} = \frac{F_i(f_{\bullet j_1} + f_{\bullet j_2})}{f_{\bullet j_0}} = F_i.$$

On trouve que

$$d^2(i,i') - D^2(i,i') \equiv \Delta =$$

$$\frac{1}{f_{\bullet j_1}}\left(\frac{f_{ij_1}}{f_{i\bullet}} - \frac{f_{i'j_1}}{f_{i'\bullet}}\right)^2 + \frac{1}{f_{\bullet j_2}}\left(\frac{f_{ij_2}}{f_{i\bullet}} - \frac{f_{i'j_2}}{f_{i'\bullet}}\right)^2 - \frac{1}{f_{\bullet j_0}}\left(\frac{f_{ij_0}}{f_{i\bullet}} - \frac{f_{i'j_0}}{f_{i'\bullet}}\right)^2.$$

Par conséquent,

$$\Delta = f_{\bullet j_1}\left(\frac{F_i}{f_{i\bullet}} - \frac{F_{i'}}{f_{i'\bullet}}\right)^2 + f_{\bullet j_2}\left(\frac{F_i}{f_{i\bullet}} - \frac{F_{i'}}{f_{i'\bullet}}\right)^2 - f_{\bullet j_0}\left(\frac{F_i}{f_{i\bullet}} - \frac{F_{i'}}{f_{i'\bullet}}\right)^2$$
$$= 0.$$

2.2.3 Analyse factorielle des correspondances

L'analyse factorielle des correspondances C'est une approche graphique permettant de représenter simultanément les profils-lignes appartenant à $\mathbb{R}^p$ et les profils-colonnes appartenant à $\mathbb{R}^n$ d'un tableau de fréquences. Comme cette représentation est faite dans le plan cartésien, elle est obtenue à l'aide d'une double analyse en composantes principales.

PROPOSITION **2.1** *Soit deux matrices symétriques,*

$$A \quad et \quad M,$$

comment déterminer le vecteur u tel que

$$u^\top Au \text{ est maximal},$$

sachant que $u^\top Mu = 1$?

$\Rightarrow$ *Il faut prendre*

$$u = \text{vecteur propre de } M^{-1}A$$

associé à

$$\lambda = \text{valeur propre maximale de } M^{-1}A.$$

On obtient ainsi

$$u^\top Au = u^\top \lambda Mu = \lambda(u^\top Mu) = \lambda.$$

On peut démonter la proposition 2.1 ainsi. Il faut maximiser $u^\top Au - \lambda(u^\top Mu - 1)$, où λ est un multiplicateur de Lagrange. En dérivant par rapport à u et en égalisant cette dérivée à zéro, on obtient $Au = \lambda Mu$. En multipliant chaque élément de l'égalité par $u^\top$, on obtient bel et bien $\lambda = u^\top Au$, qui est le terme à maximiser.

Par contre, si nous multiplions chaque élément de l'égalité par M^{-1}, nous avons

$$M^{-1}Au = \lambda u,$$

ce qui signifie que λ est bel et bien une valeur propre de $M^{-1}A$ et que u est le vecteur propre qui lui est associé.

Analyses directe et duale

Analyse directe : Les lignes de $D_n^{-1}F$ sont des éléments de $\mathbb{R}^p$. On cherche à les représenter dans cet espace muni de la fonction-distance $x^\top D_p^{-1}x$.
Analyse duale : Les lignes de $D_p^{-1}F^\top$ sont des éléments de $\mathbb{R}^n$. On cherche à les représenter dans cet espace muni de la fonction-distance $x^\top D_n^{-1}x$.

Liens avec l'analyse en composantes principales

	ACP	AFC
Données	X	$D_n^{-1}F$
Poids	I	D_n^{-1}
Distances	I	D_p^{-1}
Projections	Xu	$(D_n^{-1}F)\ D_p^{-1}u$
À maximiser	$u^\top X^\top Xu$	$\{(D_n^{-1}F)D_p^{-1}u\}^\top\ (D_n^{-1})^{-1}\ \{(D_n^{-1}F)D_p^{-1}u\}$
Contrainte	$u^\top u = 1$	$u^\top D_p^{-1}u = 1$

Premier axe factoriel

Pour **l'analyse directe**, on cherche le vecteur $u \in \mathbb{R}^p$ tel que

$$(u^\top D_p^{-1}F^\top D_n^{-1})D_n(D_n^{-1}FD_p^{-1}u)$$

soit maximal, sachant que $u^\top D_p^{-1}u = 1$. La solution est donnée par le vecteur propre principal de

$$D_p(D_p^{-1}F^\top D_n^{-1}FD_p^{-1}) = F^\top D_n^{-1}FD_p^{-1} \equiv S.$$

Pour **l'analyse duale**, on cherche le vecteur $v \in \mathbb{R}^n$ tel que

$$(v^\top D_n^{-1}FD_p^{-1})D_p(D_p^{-1}F^\top D_n^{-1}v)$$

soit maximal, sachant que $v^\top D_n^{-1}v = 1$. La solution est donnée par le vecteur propre principal de

$$D_n(D_n^{-1}FD_p^{-1}F^\top D_n^{-1}) = FD_p^{-1}F^\top D_n^{-1} \equiv T.$$

REMARQUE **2.2** *1. Les matrices S et T ne sont pas des matrices symétriques. En fait, on a*

$$s_{jj'} = \sum_{i=1}^{n} \frac{f_{ij} f_{ij'}}{f_{i\bullet} f_{\bullet j'}}, \quad j, j' \in \{1, \ldots, p\}.$$

2. Si $p \leq n$, les p premières valeurs propres de S et de T coïncident.

Nous allons démonter la dernière affirmation de cette remarque. Si $Su = F^\top D_n^{-1} F D_p^{-1} u = \lambda u$, alors

$$\begin{aligned} F D_p^{-1} S u &= F D_p^{-1} F^\top D_n^{-1} (F D_p^{-1} u) \\ &= \lambda (F D_p^{-1} u), \end{aligned}$$

de sorte que $v = F D_p^{-1} u$ est un vecteur propre de $T = F D_p^{-1} F^\top D_n^{-1}$. De plus si u est associé à λ et si

$$u^\top D_p^{-1} u = 1,$$

alors $v = F D_p^{-1} u$ est un vecteur propre de T et

$$\begin{aligned} v^\top D_n^{-1} v &= u^\top D_p^{-1} F^\top D_n^{-1} F D_p^{-1} u \\ &= u^\top D_p^{-1} (F^\top D_n^{-1} F D_p^{-1} u) \\ &= \lambda u^\top D_p^{-1} u \\ &= \lambda. \end{aligned}$$

Il faut donc prendre

$$v_j = \frac{1}{\sqrt{\lambda_j}} F D_p^{-1} u_j, \quad j \in \{1, \ldots, p\}$$

pour avoir un vecteur de norme unitaire.

Réciproquement, on a

$$u_j = \frac{1}{\sqrt{\lambda_j}} F^\top D_n^{-1} v_j, \quad j \in \{1, \ldots, p\}.$$

Le j^{e} facteur de l'analyse directe est défini par

$$\varphi_j = D_p^{-1} u_j, \quad 1 \leq j \leq p.$$

Les projections des profils des lignes sur le j^{e} vecteur propre u_j sont les composantes du vecteur $D_n^{-1}F\varphi_j$.

Le j^{e} facteur de l'analyse duale est défini par

$$\Psi_j = D_n^{-1}v_j, \quad 1 \leq j \leq p.$$

Les projections des profils des colonnes sur le j^{e} vecteur propre v_j sont les composantes du vecteur

$$D_p^{-1}F^\top D_n^{-1}v_j = D_p^{-1}F^\top \Psi_j.$$

Relations de transition

Les relations suivantes

$$D_n^{-1}F\varphi_j = \sqrt{\lambda_j}\Psi_j \equiv \hat{\Psi}_j \tag{2.8}$$

$$D_p^{-1}F^\top \Psi_j = \sqrt{\lambda}_j\varphi_j \equiv \hat{\varphi}_j \tag{2.9}$$

établissent un lien **quasi-barycentrique** entre les deux types d'analyse, c.-à-d. que **les coordonnées des points d'un espace sont proportionnelles aux composantes du facteur de l'autre espace correspondant à la même valeur propre**.

Vérification : La première relation se vérifie comme suit.

$$\begin{aligned} D_n^{-1}F\varphi_j &= D_n^{-1}(FD_p^{-1}u_j) \\ &= D_n^{-1}(\sqrt{\lambda_j}v_j) = \sqrt{\lambda_j}\Psi_j. \end{aligned}$$

La seconde identité se démontre de façon semblable.

REMARQUE **2.3** *Des conséquences des relations de transition précédentes sont que ...*

1. $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
2. $(f_{\bullet 1}, \ldots, f_{\bullet p})^\top$ *est un vecteur propre associé à la valeur propre* $\lambda_1 = 1$ *de*

$$S = F^\top D_n^{-1}FD_p^{-1}.$$

(Notez que puisque la première valeur propre est toujours égale à un, bien des logiciels ne mentionnent que les $p-1$ *autres valeurs propres.*

2.2.4 Autres notions utiles

Centre de gravité

Le but usuel de l'analyse des correspondances est de produire un graphique en 2 dimensions qui résume l'information contenues dans le tableaux de fréquences et qui fait bien ressortir les différentes associations intéressantes. Pour ce faire, on doit déterminer l'origine du graphique et les coordonnées des profils ligne et colonne le long de chaque axe.

Le centre de gravité des lignes est défini par $G_L = (g_1, \ldots, g_p)^\top$, où

$$g_j = \sum_{i=1}^{n} f_{i\bullet} \frac{f_{ij}}{f_{i\bullet}} = \sum_{i=1}^{n} f_{ij} = f_{\bullet j}\,, \quad 1 \leq j \leq p. \tag{2.10}$$

De même, le centre de gravité des colonnes est défini par

$$G_C = (f_{1\bullet}, \ldots, f_{n\bullet})^\top. \tag{2.11}$$

Le centrage des lignes est obtenu en calculant

$$\frac{f_{ij}}{f_{i\bullet}} - g_j = \frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} = \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}}$$

de sorte que

$$\sum_{j=1}^{p} \frac{f_{ij} - f_{i\bullet} f_{\bullet j}}{f_{i\bullet}} = 0$$

pour tout $i \in \{1, \ldots, n\}$. La conséquence du centrage des lignes est que l'analyse ne se fait plus sur

$$S = F^\top D_n^{-1} F D_p^{-1},$$

mais plutôt sur $S^* = (s^*_{jj'})$, où

$$s^*_{jj'} = \sum_{i=1}^{n} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})(f_{ij'} - f_{i\bullet} f_{\bullet j'})}{f_{i\bullet} f_{\bullet j'}}\,.$$

Pour détecter les associations entre lignes et colonnes, il faut faire le lien avec la statistique du test du khi-deux (X^2) Par définition,

$$\text{trace}(S^*) = \sum_{j=1}^{p} \sum_{i=1}^{n} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}\,.$$

On retrouve l'expression de la statistique X^2 servant à tester l'indépendance entre deux variables !

REMARQUE **2.4** *1. On peut montrer que S et S^* les mêmes p premiers vecteurs propres normalisés.*

2. G_L est le vecteur propre de S associé à sa valeur propre unitaire.

3. G_L est un vecteur propre de S^* associé à la valeur propre 0.

Calcul des coordonnées

Chacun des points-lignes est un élément de $\mathbb{R}^p$. Sur le graphique ils sont représentés par leur projection dans les deux premiers axes principaux.

Les p axes principaux sont déterminés par des vecteurs u_j tels que

$$u_j^\top D_p^{-1} u_j = 1, \quad j \in \{1, \ldots, p\}.$$

Puisque le j^{e} facteur de l'analyse directe est

$$\varphi_j = D_p^{-1} u_j,$$

on voit que

$$u_j^\top D_p^{-1} u_j = 1 \quad \Rightarrow \quad \varphi_j^\top D_p \varphi_j = 1,$$

c'est-à-dire que

$$\sum_{k=1}^{p} f_{\bullet k} \varphi_{jk}^2 = 1.$$

Puisque le j^{e} facteur de l'analyse duale est

$$\Psi_j = D_n^{-1} v_j,$$

on voit que

$$v_j^\top D_n^{-1} v_j = 1 \quad \Rightarrow \quad \Psi_j^\top D_n \Psi_j = 1,$$

c'est-à-dire que

$$\sum_{i=1}^{n} f_{i\bullet} \Psi_{ji}^2 = 1.$$

La projection du i^{e} point-ligne sur l'axe j est donnée par

$$\left(D_n^{-1}F\varphi_j\right)_i = \frac{1}{f_{i\bullet}}\sum_{j'=1}^{p} f_{ij'}\varphi_{jj'} = \sqrt{\lambda_j}\,\Psi_{ji} \equiv \hat{\Psi}_{ji}.$$

De plus,

$$\sum_{i=1}^{n} f_{i\bullet}\hat{\Psi}_{ji}^2 = \sum_{i=1}^{n} f_{i\bullet}\left(\sqrt{\lambda_j}\Psi_{ji}\right)^2 = \lambda_j.$$

La projection du k^{e} point-colonne sur l'axe j est donnée par

$$\left(D_p^{-1}F^\top\psi_j\right)_k = \frac{1}{f_{\bullet k}}\sum_{i=1}^{n} f_{ik}\Psi_{ji} = \sqrt{\lambda_j}\,\varphi_{jk} \equiv \hat{\varphi}_{jk}.$$

De plus,

$$\sum_{k=1}^{p} f_{\bullet k}\hat{\varphi}_{jk}^2 = \sum_{k=1}^{p} f_{\bullet k}\left(\sqrt{\lambda_j}\varphi_{jk}\right)^2 = \lambda_j.$$

La contribution à la statistique du khi-deux est proportionnelle à la distance d'un point à l'origine. Chaque axe principal a une **inertie** et chaque point correspond à l'inertie de cet axe. Plus la coordonnée du point est élevée (en valeur absolue) sur cet axe, plus sa contribution à l'inertie de cet axe est grande. Mathématiquement, on définit
l'inertie absolue du i^{e} point-ligne sur l'axe j.

$$f_{i\bullet}\hat{\Psi}_{ji}^2;$$

l'inertie relative du i^{e} point-ligne sur l'axe j.

$$\frac{f_{i\bullet}\hat{\Psi}_{ji}^2}{\lambda_j}.$$

Similairement, l'inertie absolue du k^{e} point-colonne sur l'axe j est

$$f_{\bullet k}\hat{\varphi}_{jk}^2.$$

et l'inertie relative du k^{e} point-colonne sur l'axe j est

$$\frac{f_{\bullet k}\hat{\varphi}_{jk}^2}{\lambda_j}.$$

Qualité de la représentation des observations

La qualité de la représentation d'un point est liée à la distance entre ce point et les axes du graphique. On la mesure par le cosinus de l'angle entre le point et l'axe en question. En fait dans $\mathbb{R}^n$ muni de la métrique D_n^{-1},

$$d^2(k, G_C) = \sum_{i=1}^{n} \frac{1}{f_{i\bullet}} \left(\frac{f_{ik}}{f_{\bullet j}} - f_{i\bullet} \right)^2$$

est (le carré de) la distance entre G_C et le k^{e} point- colonne. Le carré de la projection du k^{e} point-colonne sur l'axe j vaut

$$d_j^2(k, G_C) = \left(\sqrt{\lambda_j} \varphi_{jk} \right)^2$$

et

$$\sum_{j=1}^{p} d_j^2(k, G_C) = \sum_{j=1}^{p} \left(\sqrt{\lambda_j} \varphi_{jk} \right)^2 = d^2(k, G_C).$$

La **qualité de la représentation** du k^{e} point-colonne dans l'axe j est donnée par

$$\frac{d_j^2(k, G_C)}{d^2(k, G_C)} = \cos^2(\theta_{kj}),$$

où

$$\theta_{kj} = \text{angle entre le point } k \text{ et sa projection sur l'axe } j.$$

Plus les $\cos^2(\theta_{kj})$ sont élevés, plus l'angle entre les points et l'axe est près de zéro et mieux les points sont représentés sur l'axe j.

Bien sûr une définition analogue existe pour les points-ligne. On peut sommer les qualités d'un même point sur plusieurs axes pour obtenir la qualité total de la représentation de ce point par ces axes. Par exemple on définit

$$\text{Qualité} = \sum_{i=1}^{N} \cos^2(\theta_{ij})$$

pour le i^{e} point-ligne.

2.3 Analyse des correspondances multiples

2.3.1 Introduction

L'analyse des correspondances multiple sert à résumer, dans un graphique en 2 dimensions, les associations présentes dans des tableaux de fréquences croisant plus de deux variables. Cette analyse est particulièrement populaire pour faire une première analyse exploratoire d'enquêtes ou d'études faites à partir de questionnaires formés de plusieurs questions à choix multiples.

2.3.2 Notation et codification des données

Nous nous concentrerons dans ce cours sur l'analyse de questionnaires à choix multiples. Il faut coder les données d'une façon similaire à celle qui suit :

ID	Type d'employé					Type de fumeur			Total
	1	2	3	4	5	1	2	3	Q
1	0	0	1	0	0	0	1	0	2
2	0	1	0	0	0	1	0	0	2
3	1	0	0	0	0	0	1	0	2
4	0	0	0	0	1	0	0	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
193	0	1	0	0	0	1	0	0	2

Dans l'exemple précédent, le nombre de questions Q est égal à deux. On a ainsi un tableau de la forme suivante.

$$Z = [Z_1 \mid Z_2].$$

En général, pour un questionnaire contenant Q questions, on a un tableau de la forme

$$Z = [Z_1 \mid \cdots \mid Z_Q].$$

Un peu de notation additionnelle :

$$\begin{aligned} Q &= \text{nombre de questions} \\ n &= \text{nombre d'individus répondant au questionnaire} \\ p_q &= \text{nombre de modalités de la question } q \\ p &= p_1 + \cdots + p_Q \end{aligned}$$

Il est à noter que plus Q est grand, plus il y aura de cellules vides ; en fait, la proportion de cellules **non vides** est

$$\frac{nQ}{np} = \frac{Q}{p}\,.$$

Dans le cas particulier où toutes les questions ont le même nombre de choix de réponses, on a

$$p_1 = \cdots = p_Q = \frac{p}{Q}\,,$$

de sorte que

$$\frac{Q}{p} = \frac{1}{p_1} \to 0 \text{ quand } p_1 \to \infty.$$

Tableau de Burt

C'est une autre façon de présenter un tableau de fréquences contenant plus de deux variables. Étant donné un tableau logique

$$Z = [Z_1 \mid \cdots \mid Z_Q],$$

le tableau de Burt associé à ce tableau logique est la matrice carrée $p \times p$ définie comme étant

$$B = ZZ^\top$$

$$B = \begin{bmatrix} Z_1^\top Z_1 & Z_1^\top Z_2 & \cdots & Z_1^\top Z_Q \\ Z_2^\top Z_1 & Z_2^\top Z_2 & \cdots & Z_2^\top Z_Q \\ \vdots & \vdots & \cdots & \vdots \\ Z_Q^\top Z_1 & Z_Q^\top Z_2 & \cdots & Z_Q^\top Z_Q \end{bmatrix}.$$

REMARQUE **2.5** *Quelques remarques sur la forme de $Z_q^\top Z_{q'}$ ...*

1. *$Z_q^\top Z_q$ est une matrice diagonale $p_q \times p_q$ présentant les réponses à la q^{e} question.*
2. *L'élément (j, j) de la matrice $Z_q^\top Z_q$ est égal au nombre d'individus d_{jj} qui appartiennent à la j^{e} catégorie de la q^{e} question.*
3. *$Z_q^\top Z_{q'}$ est le tableau de fréquences présentant les réponses aux q^{e} et q'^{e} questions.*

4. *L'élément (j, j') de la matrice $Z_q^\top Z_{q'}$ est égal au nombre d'individus $d_{jj'}$ qui appartiennent à la j^{e} catégorie de la q^{e} question et à la j'^{e} catégorie de la q'^{e} question.*

Si on retourne à l'exemple précédent, on obtient le tableau du Burt qui suit.

	AS	AJ	ES	EJ	SE	NON	MOY	GRO
AS	11	0	0	0	0	4	5	2
AJ	0	18	0	0	0	4	10	4
ES	0	0	51	0	0	25	22	4
EJ	0	0	0	88	0	18	57	13
SE	0	0	0	0	25	10	13	2
NON	4	4	25	18	10	61	0	0
MOY	5	10	22	57	13	0	107	0
GRO	2	4	4	13	2	0	0	25

On pose $D_i = Z_i^\top Z_i$ et

$$D = \begin{bmatrix} Z_1^\top Z_1 & 0 & 0 & \cdots & 0 \\ 0 & Z_2^\top Z_2 & 0 & \cdots & 0 \\ 0 & 0 & Z_3^\top Z_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & Z_Q^\top Z_Q \end{bmatrix}$$

2.3.3 Analyse des correspondances multiples

D'un point de vue mathématique, l'analyse des correspondances multiples est une analyse des correspondances binaires effectuée sur la matrice logique Z ou sur le tableau de Burt B. Nous allons montrer que l'on obtient les mÃªmes facteurs, et ce, peu importe la matrice utilisée pour l'analyse.

Pour l'analyse des correspondances binaires, on avait

$$F = (F_{ij})\,.$$

Pour l'analyse des correspondances multiples avec la matrice Z, on a

$$F = \frac{Z}{nQ}$$

avec

$$\sum_{i=1}^{n}\sum_{j=1}^{p} f_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{Z_{ij}}{nQ} = 1.$$

Pour l'analyse des correspondances multiples avec le tableau de Burt, on a

$$F = \frac{B}{nQ^2},$$

puisque chacun des Q blocs de B est composé d'entiers dont la somme est égale à n. De plus, B est une matrice symétrique. L'analyse des correspondances multiples avec le tableau de Burt est effectuée avec $n = p$. Dans ce cas particulier, on a

$$D_n = D_p = \frac{D}{nQ}.$$

Les facteurs de l'analyse des correspondances multiples avec le tableau de Burt sont donnés par l'équation

$$\varphi_j^* = D_n^{-1} v_j = nQD^{-1} v_j,$$

où

$$FD_p^{-1}F^\top D_n^{-1} v_j = \lambda_j^* v_j.$$

De façon équivalente, on a

$$\frac{1}{Q^2} BD^{-1}B^\top D^{-1} v_j = \lambda_j^* v_j.$$

Le facteur φ_j^* est la solution de l'égalité

$$\frac{1}{Q^2} BD^{-1}B^\top \varphi_j^* = \lambda_j^* D\varphi_j^*,$$

ou encore, en multipliant les deux côtés de l'équation par D^{-1}, le facteur φ_j^* est la solution de l'éalité

$$\frac{1}{Q^2} D^{-1}BD^{-1}B^\top \varphi_j^* = \lambda_j^* \varphi_j^*.$$

REMARQUE **2.6** *rem :bvsz En ce qui concerne l'analyse basée sur la matrice Z, on sait que*

$$\frac{1}{Q} D^{-1}B^\top \varphi_j = \lambda_j \varphi_j,$$

de sorte qu'en multipliant les deux côtés de l'égalité par $D^{-1}B/Q$, on obtient que

$$\frac{1}{Q^2} D^{-1}BD^{-1}B^\top \varphi_j = \lambda_j \frac{D^{-1}B\varphi_j}{Q} = \lambda_j^2 \varphi_j.$$

On en conclut donc que

$$\lambda_j^* = \lambda_j^2, \quad 1 \le j \le p \quad \textit{et que} \quad \varphi_j^* = \varphi_j, \quad 1 \le j \le p.$$

REMARQUE **2.7** *rem :z1z2 Dans le cas où $Q = 2$, l'analyse des correspondances multiples avec la marice Z est équivalente à l'analyse des correspondances binaires avec la matrice $Z_2^\top Z_1$. En fait, le j^{e} vecteur Φ_j de l'analyse des correspondances multiples avec la matrice $Z = [Z_1 \mid Z_2]$ est de la forme*

$$\Phi_j = (\varphi_j, \psi_j)^\top,$$

où φ_j et ψ_j sont respectivement les j^{e} facteurs direct et dual de l'analyse de $Z_2^\top Z_1$. Si

$$\lambda_j^* = j^{\text{e}} \text{ valeur propre issue de } Z_2^\top Z_1,$$

alors

$$\lambda_j = \frac{1 + \sqrt{\lambda_j^*}}{2}, \quad 1 \leq j \leq p.$$

REMARQUE **2.8** *On crée un graphique comme pour l'analyse des correspondances binaires. Cependant, en analyse des correspondances multiples, la distance entre les points et la géométrie globale du graphique* **ne peuvent pas s'interpréter** *comme en analyse des correspondances binaires. En fait,*

1. *on s'intéresse aux points qui sont* **dans une même région ou dans un même quadrant du graphique** *;*
2. *on s'intéresse aux points qui sont* **dans une même direction par rapport à l'origine**.

2.4 Autres approches pour données discrètes

Il existe d'autres approches pour trouver des associations entre des variables discrètes qui ne seront pas couvertes dans ce cours. Par exemple dans le commerce de détail, on est souvent intéressé à trouver les modalités de grands ensembles de variables discrètes qui surviennent en même temps, ce que l'on appelle une analyse de panier de biens ("market basket analysis"). Dans ce type d'analyse, on cherche à savoir quels items dans un (grand) magasin ont tendance à être achetés par les mêmes clients. Des algorithmes comme l'algorithme "Apriori" permettent d'identifier des règles d'association (par exemple les hommes dans la vingtaine qui achètent de l'huile ont souvent aussi tendance à acheter du savon dégraisseur).

Chapitre 3

Classification non supervisée

3.1 Introduction

La classification a pour but de **regrouper/partionner** n observations en un certain nombre de groupes ou de classes **homogènes**. Il existe deux principaux types de classification :

1. la classification supervisée, souvent appelée simplement classification ("Classification" en anglais) ;
2. la classification non-supervisée, parfois appelée partitionnement, segmentation ou regroupement ("Clustering" en anglais).

En classifications supervisée,

— on connaît déjà le nombre de groupes qui existent dans la population ;

— on connaît **le groupe auquel appartient chaque observation** de la population ;

— on veut classer les observations dans les bons groupes à partir de différentes variables.

On peut ensuite utiliser une règle de classification pour prédire les groupes auxquels appartiennent de nouvelles observations. Des exemples classiques d'applications incluent

- identifier si une transaction bancaire est frauduleuse ou pas ;
- reconnaître des chiffres écrits à la main ;
- identifier le type de cancer dont souffre un patient.

En contre-partie en classification non-supervisée,

— on ne connaît souvent pas le nombre de groupes qui existent dans la population ;
— on ne connaît pas le groupe auquel appartient chaque observation de la population ;
— on veut classer les observations dans des groupes homogènes à partir de différentes variables.

Les applications typiques sont nombreuses. Par exemple,

- en biologie : l'élaboration de la taxonomie animale ;
- en psychologie : la détermination des types de personnalités présents dans un groupe d'individus ;
- en "text mining" : le partitionnement de courriels ou de textes en fonction du sujet traité ;
- en assurance : la segmentation des assurés en fonction du risque qu'ils représentent.

Il existe plusieurs familles de méthodes de classification non-supervisée. Les plus communes incluent

— la classification hiérarchique ;
— la classification non-hiérarchique, par exemple la méthode des k-moyennes ("k-means") ;
— la classification basée sur une densité ;
— la classification basée sur des modèles statistiques/probabilistes, par exemple un mélange de lois normales.

Dans ce cours, nous nous concentrerons sur la classification hiérarchique, mais nous commencerons par un survol rapide de la méthode des k-moyennes. Le lecteur intéressé à en savoir plus peut trouver beaucoup d'information sur le sujet dans le livre de Bandyopadhyay and Saha (2013), disponible en format PDF sur le site de la bibliothèque.

3.2 Distance et similarité entre deux observations

Pour regrouper des observations en groupes homogènes, il faut tout d'abord avoir une définition de ce que sont des observations *similaires* ou des observations *différentes*. Il faut donc être en mesure de **quantifier la similarité ou la distance entre deux observations**. Cette première étape peut parfois être la plus difficile de tout le processus de classification,

mais elle est essentielle et le premier pas de toute analyse de partitionnement.

Si les observations sont constituées de p nombres réels de valeurs du même ordre de grandeur, alors la distance euclidienne entre les deux vecteurs dans $\mathbb{R}^p$ est un mesure tout-à-fait raisonnable. Mais que fait-on quand les observations sont constituées de p variables binaires (oui/non, homme/femme, présent/absent, etc.), ou p variables catégorielles, des images, des textes, ou encore plus difficile, un mélange de tout cela (par exemple pour chacun de n individus on a l'âge, le revenu, le sexe, le niveau d'éducation et une description de l'emploi en un paragraphe) ?

En fait plusieurs mesures ont été développées sur mesure pour leur application particulière, à force d'expérience et d'expérimentation. Dans ce cours, nous survolerons les mesures les plus classiques qui permettent de traiter une bonne proportion des problèmes standards. Nous commencerons par les mesures plus simples ou générales pour terminer avec des mesures plus complexes ou spécialisées.

3.2.1 Mesures de distance

DÉFINITION **3.1** *Une* mesure de distance d *doit satisfaire les propriétés suivantes pour tout* $i, j, k \in \{1, \ldots, n\}$ *:*

1. $d(i, j) \geq 0$ *;*
2. $d(i, i) = 0$ *;*
3. $d(i, j) = d(j, i)$ *;*
4. $d(i, k) \leq d(i, j) + d(j, k)$.

La distance $\mathcal{L}_q$ entre deux vecteurs dans $\mathbb{R}^p$ est définie par

$$||x_i - x_j||_q = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right)^{1/q}.$$

La distance euclidienne correspond au cas ou $q = 2$.

REMARQUE **3.1** *La distance* $\mathcal{L}_q$ *n'est* **PAS** *invariante à un changement d'échelle.*

La remarque 3.1 a des conséquences majeures en pratique. Par exemple, considérons le jeu de données suivant :

Objet	poids (g)	taille (cm)
1	10	7
2	20	2
3	30	10

On trouve les distances suivantes :

$$d_{12} = 11.2, \quad d_{13} = 20.2, \quad d_{23} = 12.8.$$

Si la taille est exprimée en millimètres, on trouve les distances suivantes.

$$d_{12} = 51.0, \quad d_{13} = 36.1, \quad d_{23} = 80.6.$$

On peut donc se demander si le premier objet est plus près du deuxième objet ou du troisième objet ! ?

Cette remarque et cet exemple expliquent pourquoi dans plusieurs situations en pratique on préfère travailler avec **la distance standardisée** entre les variables,

$$d^2(x_i, x_j) = \sum_{k=1}^{p} \{(x_{ik} - \mu_k)/s_k - (x_{jk} - \mu_k)/s_k\}^2 = \sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2 ,$$

où

$$\mu_k = \text{moyenne de la variable } k; \quad s_k = \text{écart-type de la variable } k.$$

3.2.2 Indices de similarité

DÉFINITION **3.2** *Un* indice de similarité s *entre des objets doit satisfaire les propriétés suivantes pour tout* $i, j, k \in \{1, \ldots, n\}$ *:*

1. $s(i, j) \geq 0$;
2. $s(i, j) = s(j, i)$;
3. $s(i, i) = 1 \geq s(i, j)$.

REMARQUE **3.2** *Une distance peut se transformer en similarité en posant*

$$s_{ij} = \frac{1}{1 + d_{ij}}.$$

La relation inverse n'est pas vraie, en raison de l'inégalité du triangle.

On peut aussi définir la **dissemblance** *("dissimilarity" en anglais) entre deux objets, soit*

$$d^*_{ij} = 1 - s_{ij}.$$

Types de variables

L'indice de similarité dépend du(des) type(s) de variables utilisées dans l'analyse. Les principaux types de variables avec lesquels on doit composer sont les suivants.

- **Numériques/réelles :** Il s'agit de variables dont la valeur numérique mesure quelque chose de quantifiable et dont la différence entre les valeurs reflète la différence entre les objets. On peut ainsi parler du revenu en dollars, de la masse, de l'âge, etc.
- **Ordinales :** Il s'agit de variables qui ne donnent pas une quantification précise d'un phénomène, mais dont les modalités peuvent être naturellement ordonnées. On peut penser par exemple à un revenu faible, moyen ou élevé, ou à un niveau d'accord entre "tout-à-fait en désaccord", "en désaccord", "pas d'avis", "d'accord", "tout-à-fait d'accord".
- **Nominales symétriques :** Il s'agit de variables qualitatives (donc qui ne sont ni numériques ni ordinales) dont toutes les modalités sont aussi informatives l'une que l'autre. On peut penser par exemple au sexe (homme ou femme), à laquelle de l'une de 4 sections d'un cours des étudiants appartiennent, etc.
- **Nominales asymétriques :** Il s'agit de variables qualitatives dont les modalités ne contiennent pas toutes le même niveau d'information. Ceci se produit habituellement lorsque l'une des modalités est très fréquente, un peu la modalité "par défaut", mais que les autres ne le sont pas. Par exemple si une variable indique si un indivdu est daltonien ou pas, deux individus daltoniens ont quelque chose en commun, mais deux individus non daltoniens n'ont pas nécessairement quelque chose en commun. Un autre exemple pourrait être si une transaction est frauduleuse ou pas dans une analyse où une très faible proportion des transactions sont frauduleuses.

Variables nominales binaires

Pour des vecteurs de variables binaires symétriques, on utilise la proportion d'accords ("matching coefficient") dans les éléments des vecteurs. On commence par coder l'une modalité à 0 et l'autre à 1. Si on mesure p variables binaires pour chacun de deux individus i et j, on compte le nombre de variables pour lesquelles ces deux individus ont la même valeur pour une même variable, soit $m = \sum_{k=1}^{p} I(x_{ik} = x_{jk})$ et la similarité est définie par $s(i, j) = m/p$. Par exemple supposons que deux individus remplissent un questionnaire de 10 questions et que la valeur 1 représente une réponse "oui" et une valeur 0 une réponse "non".

Individu	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
i	1	0	0	0	0	1	0	0	0	0
j	1	0	0	0	0	0	1	0	0	0

Alors la similarité entre x_i et x_j ici serait $s(i, j) = 8/10 = 0.8$, puisqu'ils ont donné la même réponse pour 8 des 10 questions posées.

Pour des vecteurs de variables binaires asymétriques, on assigne la modalité 1 à la valeur la plus rare (ou la plus importante) et la valeur 0 à l'autre modalité. On peut ensuite utiliser l'indice de Jaccard défini par le nombre de variables pour lesquelles i et j prennent simultanément la valeur 1 sur le nombre de variables pour lesquelles au moins l'un de i ou j n'a pas la valeur 0, soit

$$J(i, j) = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sum_{k=1}^{p} \{1 - (1 - x_{ik})(1 - x_{jk})\}}.$$

Dans l'exemple ci-dessus, on aurait donc $J(i, j) = 1/3 = 0.33$, puisque les 7 questions pour lesquelles les deux individus donnent simultanément la valeur 0 ne sont pas comptées ni au numérateur, ni au dénominateur ; pour les 3 questions qui comptent (Q1, Q6 et Q7), ils sont en accord 1 fois, d'où le $1/3$.

Variables nominales polytomiques

Si une variable est composée de $M > 2$ modalités, on peut la coder en utilisant $M - 1$ variables binaires. Par exemple si les réponses possibles à une question sont "Oui", "Non", "Je

ne sais pas", on pourrait coder les trois réponses possibles ainsi :

	x_{i1}	x_{i2}
Oui	1	0
Non	0	1
Je ne sais pas	0	0

On peut ensuite calculer la similiraté entre i et j en utilisant les méthodes ci-dessus.

Si on a q variables binaires ou polytomiques de même nature, il est d'usage de calculer la similarité séparément pour chacune des q variables et ensuite de faire la moyenne des q similarités.

Par exemple supposons que nos deux individus i et j on répondu à deux questions pour lesquelles ils avaient trois choix : i répond "b" aux deux questions alors que j répond "b" à la première et "c" à la deuxième. Suppons que les choix sont codés selon le tableau ci-dessous.

Individu	Q1a	Q1b	Q2a	Q2b
i	0	1	0	1
j	0	1	0	0

Calculons les similarités pour chaque question en supposant les modalités aussi importantes les unes que les autres. On aura $s_{Q1}(i,j) = 2/2$ et $s_{Q2}(i,j) = 1/2$. En prenant la moyenne des deux similarités on obtient $s(i,j) = \{s_{Q1}(i,j) + s_{Q2}(i,j)\}/2 = 3/4$. Maintenant supposons que les variables correspondant à la question 1 sont symétriques et celles correspondant à la question 2 sont asymétriques. La similarité $s_{Q1}(i,j) = 2/2$ demeure inchangée puisqu'on conserva la même règle. Pour la 2e question, on prend maintenant l'indice de Jaccard et donc $s_{Q2}(i,j) = 0/1$ car Q2a ne contribue pas au calcul. On obtient donc $s(i,j) = \{2/2+0/1\}/2 = 1/2$.

Variables ordinales

On assigne habituellement un score numérique à chaque modalité de la variable ordinale, ensuite on la traite comme une variable numérique. Il n'y a pas de règle sur les scores

numériques à donner, à part qu'ils doivent être positifs et refléter l'ordre des modalités. Par exemple pour une question sur le revenu, on pourrait accorder respectivement des scores de 1, 2, 3 pour des revenus "faible", "moyen", "élevé", ou on pourrait aussi accorder 15000, 50000, 150000 ; c'est un exercice de jugement et il n'existe pas de règle mathématique claire.

3.2.3 Observations constituées de plusieurs types de variables

Que doit-on faire si chaque observation est constituée de variables de plusieurs types, par exemple si pour chacun de n individus nous mesurons l'âge (numérique), le sexe (nominale symétrique), s'il est porteur d'une mutation génétique rare (nominale asymétrique) et son niveau d'accord avec une certaine politique (ordinale) ? Il existe quelques façons de procéder, mais la plus commune semble être de mesurer la similarité de Gower. On doit tout d'abord recoder toute variable nominale sous forme de variables binaires et toute variable ordinale sous forme de variable numérique et supposons qu'une fois cette opération effectuée, on obtient p variables par individu. La similarité de Gower [1] entre x_i et x_j est définie ainsi :

$$G(i,j) = \frac{\sum_{k=1}^{p} w_k \gamma_k(i,j) s_k(i,j)}{\sum_{k=1}^{p} w_k \gamma_k(i,j)},$$

où w_k est un poids accordé à la variable k et $\gamma_k(i,j)$ et $s_k(i,j)$ sont définies différemment selon le type de la variable k,

- **variable k numérique ou ordinale :** $\gamma_k(i,j) = 1$ et $s_k(i,j) = 1 - |x_{ik} - x_{jk}|/r_k$;
- **variable k nominale symétrique :** $\gamma_k(i,j) = 1$ et $s_k(i,j) = I(x_{ik} = x_{jk})$;
- **variable k nominale asymétrique :** $\gamma_k(i,j) = \{1-(1-x_{ik})(1-x_{jk})\}$ et $s_k(i,j) = I(x_{ik} = x_{jk})$.

La valeur r_k dans la similarité pour les variables numériques/ordinales est l'étendue ("range") de la variable k. Il est fortement recommandé de standardiser les variables numériques/ordinales au préalable. Les poids w_k permettent de moduler l'importance de chaque variable dans la mesure de similarité.

Si on poursuit avec notre exemple, supposons que nous avons recodé les réponses de nos individus i et j de sorte que la variable 1 est numérique, la variable 2 est ordinale, les

1. Il s'agit ici de la définition utilisée par la procédure `DISTANCE` de SAS et la fonction `daisy` de la librairie `cluster` de R, avec la nuance que la fonction R calcule la **dis**similarité de Gower.

variables 3 et 4 sont des indicatrices binaires correspondant à une variable symétrique et la variable 5 est une indicatrice correspondant à une variable asymétrique. Les valeurs pour les variables 1 et 2 ont été standardisées et supposons qu'elles prennent des valeurs entre -2.5 et 2.5 (donc une étendue de 5).

Individu	Q1	Q2	Q3	Q4	Q5
i	1	2	0	1	0
j	-1	1	0	0	1

Supposons que nous voulons que la question 1 soit trois fois plus importante que les autres dans la mesure de similarité ; il faut lui accorder un poids w_1 qui est trois fois plus élevé que $w_2 = w_3 = w_4 = w_5$. On peut prendre $w_1 = 3$ et $w_2 = w_3 = w_4 = w_5 = 1$. On a que $\gamma_1(i,j) = \gamma_2(i,j) = \gamma_3(i,j) = \gamma_4(i,j) = 1$ et $\gamma_5(i,j) = 1 - (1 - x_{i5})(1 - x_{j5}) = 1 - (1-1)(1-0) = 1$. On a ensuite $s_1(i,j) = 1 - |1 - (-1)|/5 = 3/5$, $s_2(i,j) = 1 - |2-1|/5 = 4/5$, $s_3(i,j) = 1/1$, $s_4(i,j) = 0/1$ et $s_5(i,j) = 0/1$. (Note : Si on a $s_k(i,j) = 0/0$ pour une variable asymétrique, c'est que $\gamma_k(i,j)$ sera 0 et donc cette variable ne contribuera pas au calcul de $G(i,j)$.)

$$\begin{aligned} G(i,j) &= \frac{\sum_{k=1}^{5} w_k \gamma_k(i,j) s_k(i,j)}{\sum_{k=1}^{5} w_k \gamma_k(i,j)} \\ &= \left(3 \times 1 \times \frac{3}{5} + 1 \times 1 \times \frac{4}{5} + 1 \times 1 \times \frac{1}{1} + 1 \times 1 \times \frac{0}{1} + 1 \times 1 \times \frac{0}{1}\right) \\ &\quad \div (3 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 1 + 1 \times 1) \\ &= \frac{18/5}{7} = \frac{18}{35} \approx 0.514. \end{aligned}$$

3.2.4 Mesures pour des applications spécifiques

Textes

Pour le partitionnement de n documents écrits (p.ex. pages web, courriels, lettres à l'éditeur, plaintes, documents légaux, etc.), on se crée une matrice de documents par termes où chacune des lignes de la matrice correspond à un des documents et chacune des p colonnes

de la matrice correspond à l'un des mots présents dans l'ensemble des documents [2]. L'élément de la ligne i, colonne k est simplement le nombre de fois (fréquence) où le k-ème mot apparait dans le i-ème document. Ces matrices sont généralement de très grandes (nombre p de colonnes élevés) matrices éparses (la vaste majorité des éléments sont des zéros). On va fréquemment remplacer les fréquences non nulles par 1 (donc l'élément en ligne i, colonne k est 1 si le mot j est dans le document i, 0 sinon) ou remplacer les fréquences par des encodages qui modifient les fréquences pour tenir compte de l'importance des mots comme tf-idf, comme l'expliquent par exemple Weiss et al. (2015).

Si on utilise un codage présence/absence et une mesure de similarité comme la proportion d'accords ou la similarité de Jaccard, alors on aura que pratiquement toutes les paires de documents auront une similarité nulle ou très faible. C'est pourquoi on utilise habituellement cette approche pour détecter des textes où des extraits sont recopiés (par exemple détection de plagiat).

Règle générale, une mesure qui fonctionne bien pour mesurer la similarité entre des textes est celle du cosinus. Elle est définie pour des variables binaires ou numériques ainsi :

$$s_{cos}(i,j) = \frac{\sum_{k=1}^{p} w_k \; x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{p} w_k x_{ik}^2 \;\; \sum_{k=1}^{p} w_k x_{jk}^2}}.$$

Les exemples de calcul de similarité entre des documents sont très nombreux sur internet, et une bonne proportion de ces calculs utilisent la similarité du cosinus.

3.2.5 Autres possibilités

Les techniques de réduction de la dimension vues aux chapitres 1 et 2 peuvent aussi être utilisées pour simplifier le calcul des similarités/distances entre les observations. Par exemple

2. Il est d'usage d'éliminer des mots qui ne contiennent pas d'information ("stopwords") comme les articles, les prépositions, les conjonctions, les pronoms, etc. et de ne conserver que la racine des mots, par exemple seulement "météo" pour "météorologie" et "météorologique" et d'effectuer quelques autres opérations pour diminuer le nombre de colonnes et ne se limiter qu'à l'information utile. Il est donc recommandé de consulter des ouvrages sur l'analyse automatisée de textes, comme par exemple Weiss et al. (2015), afin de bien comprendre les étapes préliminaires à la création d'une matrice de documents par termes.

si on a p variables numériques/ordinales et que p est une très grande valeur, on peut utiliser l'analyse en composantes principales pour calculer les scores de chaque observation dans les $k << p$ axes principaux. On utilise ensuite ces k scores et la distance euclidienne pour calculer la distance entre les objets. Si on a plutôt Q réponses à un questionnaire à choix multiples (ou plus généralement la valeur de Q variables catégorielles), on peut calculer les scores de chaque observation dans les k premiers axes principaux d'une analyse des correspondances multiples, ensuite encore une fois utiliser ces k scores et la distance euclidienne pour calculer les distances.

Remarque : Cette stratégie est utilisée dans de nombreuse situations en pratique, en particulier quand p est très grand, et fonctionne généralement bien. Mais dans certaines situations il n'est pas impossible que des groupes nettement séparés en dimension p soient difficilement distinguables en dimension $k < p$. La figure 3.1 montre un exemple où l'on voit bien les 2 groupes (points noirs et triangles rouges) en dimension $p = 2$, mais quand on regarde les scores de ces mêmes observations dans la première composante principale ($k = 1$), les deux groupes ne sont plus aussi clairement distinguables.

3.3 Méthode des k-moyennes

3.3.1 Problème général

Maintenant que nous pouvons mesurer la distance ou la similarité entre deux observations, on voudrait partitionner les n observations du jeu de données en K classes (groupes, catégories) avec comme objectifs :

— que les observations dans une même classe soient le plus similaires possible ;

— que les observations dans des classes différentes soient les moins similaires possibles.

Soit

$$\begin{aligned} C : \{1, \ldots, n\} &\rightarrow \{1, \ldots, K\} \\ i &\mapsto C(i), \end{aligned}$$

où $C(i)$ nous dit auquel des groupes $1, \ldots, K$ l'observation i appartient. On cherche la fonction C qui va nous permettre de remplir les deux objectifs. Autrement dit, on cherche

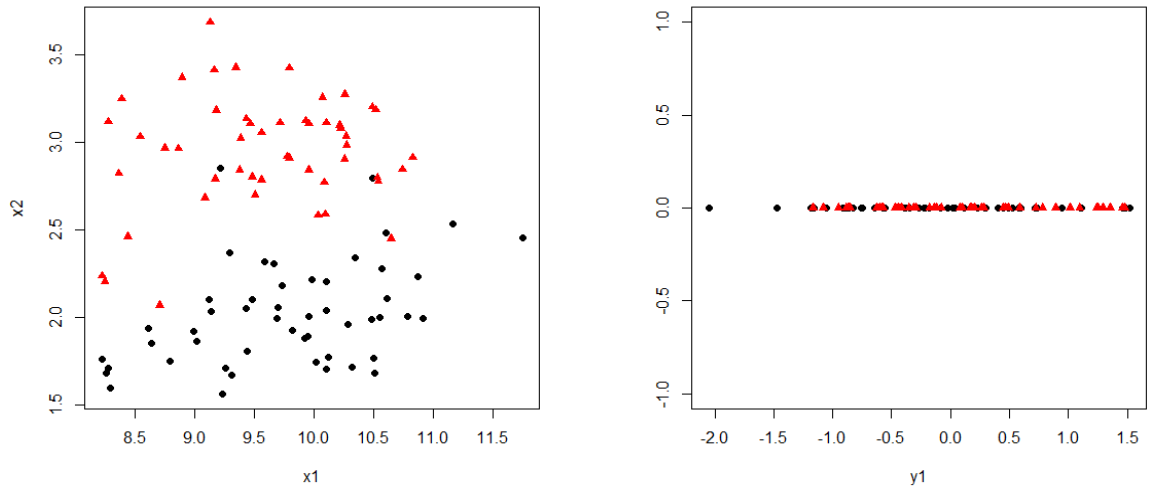

FIGURE 3.1 – Gauche : Données originales (x_1, x_2) dont 50 observations sont du groupe 1 (points noirs) et 50 observations sont du groupe 2 (triangles rouges). Droite : Mêmes observations dans le premier axe principal de l'ACP.

C qui va minimiser la fonction objectif

$$W(C) = \sum_{k=1}^{K} \sum_{i:C(i)=k} \sum_{j:C(j)=k} d^*(x_i, x_j), \tag{3.1}$$

où $d^*(x_i, x_j)$ est la dissimilarité entre les observations i et j. On peut démontrer que la règle qui minimise la dissimilarité $W(C)$ à l'intérieur des classes est aussi la règle qui maximise la dissimilarité entre les observations de classes différentes (voir par exemple Hastie et al. (2009)), donc la règle qui remplit les deux objectifs.

Une première approche brute pourrait être d'essayer toutes les assignations possibles des n observations à K groupes et de voir laquelle de ces assignations donne la valeur la plus faible pour $W(C)$. Malheureusement, il y a

$$\frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n$$

possibilités ici, ce qui est une valeur beaucoup trop élevée pour que cette stratégie soit applicable en pratique. Hastie et al. (2009) donnent comme exemple le cas d'un tout petit échantillon de taille $n = 19$ que l'on voudrait diviser en $K = 4$ groupes, ce qui donne approximativement 10^{10} possibilités ! Les algorithmes que nous verrons seront donc des algorithmes dit "gloutons", c'est-à-dire qu'ils vont nous donner une règle C qui minimise (3.1) sur un espace restreint et qui ne nous garantissent pas que nous avons bien trouvé le C qui minimise globalement $W(C)$.

3.3.2 Méthode des k-moyennes

La méthode des k-moyennes s'applique à des situations où les p variables sont numériques/ordinales (et habituellement standardisées). L'objectif de la méthode est de partitionner les données en K groupes et **la valeur de K est fixée**. L'algorithme est relativement simple et on peut démontrer qu'à chaque étape de son exécution, la valeur de $W(C)$ est diminuée.

Algorithme des k-moyennes

1. On choisit le nombre de classes/groupes K que l'on désire obtenir.
2. On partitionne aléatoirement les n observations en K groupes.
3. On calcule le vecteur-moyenne pour chacun des K groupes, soit

$$\mu_k = \frac{1}{N_k} \sum_{i:C(i)=k} x_i, \quad k = 1, \ldots, K,$$

   où N_k est le nombre d'observations dans le groupe k.
4. On calcule la distance entre chaque observation et chacun des K vecteurs-moyennes.
5. On assigne chacune des n observations au groupe dont le vecteur-moyenne est le plus près.
6. On répète les étapes 3 à 5 jusqu'à ce qu'aucune observation ne soit réassignée à un nouveau groupe.

EXEMPLE **3.1** *Supposons les 5 observations des variables x_1 et x_2 suivantes que nous désirons partitionner en $K = 2$ groupes selon la méthode des k-moyennes :*

i	*1*	*2*	*3*	*4*	*5*
x_{i1}	*-1*	*-0.5*	*0*	*0.5*	*1*
x_{i2}	*-1*	*0*	*0.5*	*-0.5*	*1*

Supposons que l'on assigne aléatoirement les observations 1, 2 et 5 au groupe 1 et les observations 3 et 4 au groupe 2. On calcule μ_1, le vecteur des valeurs moyennes de x_{i1} et de x_{i2} pour les observations (i) du groupe 1 et μ_2, le même vecteur pour le groupe 2 :

$$\begin{aligned} \mu_1 &= \frac{1}{3}\left\{\begin{pmatrix}-1\\-1\end{pmatrix} + \begin{pmatrix}-0.5\\0\end{pmatrix} + \begin{pmatrix}1\\1\end{pmatrix}\right\} = \begin{pmatrix}-1/6\\0\end{pmatrix} \\ \mu_2 &= \frac{1}{2}\left\{\begin{pmatrix}0\\0.5\end{pmatrix} + \begin{pmatrix}0.5\\-0.5\end{pmatrix}\right\} = \begin{pmatrix}1/4\\0\end{pmatrix}. \end{aligned}$$

On calcule ensuite la distance entre chacun des points et chacun des vecteurs-moyennes et on assigne chaque point au groupe dont le vecteur-moyenne est le plus près.

i	*1*	*2*	*3*	*4*	*5*
$d^2(i,\mu_1)$	**1.69**	**0.11**	**0.28**	*0.69*	*2.36*
$d^2(i,\mu_2)$	*2.56*	*0.56*	*0.31*	**0.31**	**1.57**

On a donc un groupe avec les observations 1, 2 et 3 et un autre avec les observations 4 et 5. On calcule les nouvelles moyennes de chaque groupe :

$$\begin{aligned}\mu_1 &= \frac{1}{3}\left\{\begin{pmatrix}-1\\-1\end{pmatrix}+\begin{pmatrix}-0.5\\0\end{pmatrix}+\begin{pmatrix}0\\0.5\end{pmatrix}\right\}=\begin{pmatrix}-1/2\\-1/6\end{pmatrix}\\ \mu_2 &= \frac{1}{2}\left\{\begin{pmatrix}0.5\\-0.5\end{pmatrix}+\begin{pmatrix}1\\1\end{pmatrix}\right\}=\begin{pmatrix}3/4\\1/4\end{pmatrix}.\end{aligned}$$

On calcule ensuite la distance entre chacun des points et chacun des vecteurs-moyennes et on assigne chaque point au groupe dont le vecteur-moyenne est le plus près.

i	*1*	*2*	*3*	*4*	*5*
$d^2(i,\mu_1)$	**0.94**	**0.03**	*0.69*	*1.11*	*3.61*
$d^2(i,\mu_2)$	*4.63*	*1.63*	**0.63**	**0.63**	**0.63**

On a donc un groupe avec les observations 1 et 2 et un autre avec les observations 3, 4 et 5. On calcule les nouvelles moyennes de chaque groupe :

$$\begin{aligned}\mu_1 &= \frac{1}{2}\left\{\begin{pmatrix}-1\\-1\end{pmatrix}+\begin{pmatrix}-0.5\\0\end{pmatrix}\right\}=\begin{pmatrix}-3/4\\-1/2\end{pmatrix}\\ \mu_2 &= \frac{1}{3}\left\{\begin{pmatrix}0\\0.5\end{pmatrix}+\begin{pmatrix}0.5\\-0.5\end{pmatrix}+\begin{pmatrix}1\\1\end{pmatrix}\right\}=\begin{pmatrix}1/2\\1/3\end{pmatrix}.\end{aligned}$$

On calcule ensuite la distance entre chacun des points et chacun des vecteurs-moyennes et on assigne chaque point au groupe dont le vecteur-moyenne est le plus près.

i	*1*	*2*	*3*	*4*	*5*
$d^2(i,\mu_1)$	**0.31**	**0.31**	*1.56*	*1.56*	*5.31*
$d^2(i,\mu_2)$	*4.03*	*1.11*	**0.28**	**0.69**	**0.69**

Aucune observation ne change de groupe, l'algorithme cesse donc ici, avec ces groupes et ces vecteurs-moyennes.

Exercice : *Calculez la statistique $W(C)$ donnée par l'équation (3.1) pour chacune des 3 assignations des 5 observations aux 2 groupes en utilisant la distance euclédienne. Qu'arrive-t-il à la valeur de cette statistique après chaque nouvelle assignation ?*

3.4 Classification hiérarchique

3.5 Introduction

Si on ne sait pas quelle valeur de K utiliser ou si les observations ne sont pas constituées de variables numériques ou ordinales, la méthode du k-means ne s'applique pas (il faut une distance en bonne et due forme pour que les assignations aux groupes via les distance à un vecteur-moyenne ait du sens). On peut alors utiliser un des nombreux algorithmes de classification hiérarchique. La classification hiérarchique permet d'obtenir des partitions toutes imbriquées les unes dans les autres. Il existe deux types d'algorithmes pour effectuer de la classification hiérarchique :

— les algorithmes ascendants ;
— les algorithme descendants.

Dans les deux cas, on obtient des ensembles de partitions hiérarchiques contenant d'un à n groupes.

3.5.1 Algorithmes descendants

Un algorithme descendant fonctionne ainsi :

— au départ, toutes les observations sont dans un seul et même groupe de n observations ;
— à chaque étape, on divise le groupe le moins homogène en deux groupes ;
— à la fin, après n étapes, chaque observation est son propre groupe, c'est-à-dire qu'on obtient n groupes contenant une seule observation.

L'exécution d'un tel algorithme ne donne pas une seule partition, mais n partitions : une partition avec un groupe, une partition avec deux groupes, ..., une partition avec n groupes. Nous verrons plus tard comment résumer de façon visuelle le résultat d'une classification hiérarchique à l'aide d'un graphique en forme d'arbre appelé "dendogramme". Nous verrons aussi des critères qui peuvent aider à choisir l'une parmi les n partitions proposées par l'algorithme.

Caractéristiques

— Les algorithmes descendants demandent beaucoup de temps de calcul (ce n'est pas tout de déterminer quel groupe scinder en 2, mais on doit déterminer comment se découpage doit se faire).
— Ils sont moins utilisés en pratique que les algorithmes ascendants.

3.5.2 Algorithmes ascendants

Un algorithme ascendant fonctionne de manière opposée à un algorithme descendant :
— au départ chaque observation est son propre groupe, c'est-à-dire qu'on démarre avec n groupes contenant chacun une seule observation ;
— à chaque étape on fusionne les deux groupes les plus similaires ;
— à la fin des n étapes, on obtient un seul groupe contenant toutes les n observations.

Les implémentations de la classification hiérarchique se distinguent de trois façons :

1. leur caractère ascendant ou descendant ;
2. leur façon de mesurer les distances ou les similarités entre deux observations ;
3. leur façon de mesurer les distances ou les similarités entre deux groupes.

3.5.3 Distance et similarité entre deux groupes

Pour mettre en oeuvre les algorithmes mentionnés précédemment, on doit définir $d(A, B)$, la distance entre deux groupes groupes d'observations A et B tels que
— les groupes A et B sont des sous-ensembles des observations du jeu de données ;
— l'intersection entre les groupes A et B est l'ensemble vide.

Nous avons vu toutes sortes de méthodes pour calculer la distance entre une paire d'observations, mais nous devons maintenant considérer des méthodes pour calculer la distance entre une paire de groupes d'observations. Il existe plusieurs façons de calculer une telle distance entre deux groupes lors de l'exécution d'une classification hiérarchique ascendante. Nous commençons par les lister, nous présenterons ensuite les points forts et les points faibles de chacune des méthodes.

Méthode du plus proche voisin (Single linkage)

La distance entre deux groupes se définit comme suit pour cette méthode :

$$d(A, B) = \min \{d_{ij} : i \in A, j \in B\}.$$

Si on travaille plutôt avec des indices de similarité, on pose

$$s(A, B) = \max \{s_{ij} : i \in A, j \in B\}.$$

On voit donc d'où la méthode tire son nom : la distance/similarité entre deux groupes d'observations est tout simplement la distance/similarité entre les points de chaque groupe qui sont les plus rapprochés/similaires.

Méthode du voisin le plus distant (Complete linkage)

La distance entre deux groupes se définit comme suit pour cette méthode :

$$d(A, B) = \max \{d_{ij} : i \in A, j \in B\}.$$

Si on travaille plutôt avec des indices de similarité, on pose

$$s(A, B) = \min \{s_{ij} : i \in A, j \in B\}.$$

Encore une fois comme le nom l'indique, la distance/similarité entre deux groupes d'observations est tout simplement la distance/similarité entre les points de chaque groupe qui sont les plus éloignés/dissimilaires.

Méthode de la moyenne (Average linkage)

La distance entre deux groupes se définit comme suit pour cette méthode :

$$d(A, B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d(x_i, x_j),$$

où n_A est le nombre d'observations dans le groupe A et n_B est le nombre d'observations dans le groupe B.

On doit donc calculer les $n_A \times n_B$ distances possibles entre les points des deux groupes, ensuite on prend la moyenne de ces distances comme étant celle qui sépare les deux groupes.

Méthode du centroïde (Centroid method)

La distance entre deux groupes se définit comme suit pour cette méthode :

$$d\left(A,B\right)=d\left(\bar{x}_A,\bar{x}_B\right),$$

où

$$\bar{x}_A=\frac{1}{n_A}\sum_{i\in A}x_i,\quad \bar{x}_B=\frac{1}{n_B}\sum_{j\in B}x_j.$$

La moyenne $\bar{x}_{AB}$ du nouveau groupe résultant de la fusion des groupes A et B se calcule comme suit :

$$\bar{x}_{AB}=\frac{n_A\bar{x}_A+n_B\bar{x}_B}{n_A+n_B}.$$

Méthode de la médiane (Median method)

À une étape donnée nous avons toujours à notre disposition la distance entre les groupes déjà formés. On fusionne les deux groupes les plus similaires, disons A et B pour obtenir un groupe AB. Avec la méthode de la médiane la distance entre le nouveau groupe AB et tout autre groupe C est donnée par

$$d(AB,C)=\frac{d(A,C)+d(B,C)}{2}-d(A,B)/4.$$

Méthode de Ward (Ward method)

La méthode de Ward est une variante de la méthode du centroïde pour tenir compte de la taille des groupes. Elle a été conçue de sorte à être optimale si les n vecteurs $x_1,\ldots,x_n$ suivent des lois normales multivariées de K moyennes différentes mais toutes de même matrice de variance-covariance. Elle est basée sur les sommes de carrés suivantes :

$$\begin{aligned}SC_A&=\sum_{i\in A}(x_i-\bar{x}_A)^\top(x_i-\bar{x}_A),\\ SC_B&=\sum_{j\in B}(x_j-\bar{x}_B)^\top(x_j-\bar{x}_B),\\ SC_{AB}&=\sum_{k\in A\cup B}(x_k-\bar{x}_{AB})^\top(x_k-\bar{x}_{AB}),\end{aligned}$$

où $\bar{x}_A$, $\bar{x}_B$ et $\bar{x}_{AB}$ sont calculées comme dans la méthode du centroïde.

On regroupe ensuite les classes A et B pour lesquelles

$$\begin{aligned} I_{AB} &= SC_{AB} - SC_A - SC_B \\ &= \frac{n_A n_B}{n_A + n_B} \left(\bar{x}_A - \bar{x}_B\right)^\top \left(\bar{x}_A - \bar{x}_B\right) \\ &= \frac{d^2(\bar{x}_A, \bar{x}_B)}{\frac{1}{n_A} + \frac{1}{n_B}} \end{aligned}$$

est minimale.

Méthode flexible (Flexible method)

Les auteurs de cette méthode ont remarqué que pour plusieurs méthodes connues, on a les relations suivantes :

$$\begin{aligned} d(C, A \cup B) = \alpha_A \, d(C, A) + \\ \alpha_B \, d(C, B) + \\ \beta \, d(A, B) + \\ \gamma \, |d(C, A) - d(C, B)| \end{aligned}$$

Méthode	α_A	α_B	β	γ
Plus proche	$1/2$	$1/2$	0	$-1/2$
Plus distant	$1/2$	$1/2$	0	$1/2$
Médiane	$1/2$	$1/2$	$-1/4$	0
Moyenne	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	0	0
Centroïde	$\frac{n_A}{n_A+n_B}$	$\frac{n_B}{n_A+n_B}$	$-\frac{n_An_B}{n_A+n_B}$	0
Ward	$\frac{n_A+n_C}{n_A+n_B+n_C}$	$\frac{n_B+n_C}{n_A+n_B+n_C}$	$-\frac{n_C}{n_A+n_B+n_C}$	0

Avec la méthode flexible, on impose arbitrairement les contraintes suivantes :

$$\alpha_A + \alpha_B + \beta = 1, \quad \alpha_A = \alpha_B, \quad \gamma = 0.$$

Ainsi,

$$\alpha_A = \alpha_B = (1-\beta)/2$$

et il ne reste qu'à choisir β.

Les auteurs suggèrent de poser $\beta = -0.25$, mais ils recommandent plutôt de poser $\beta = -0.5$ lorsque l'on soupçonne la présence de données aberrantes.

3.5.4 Choix de l'algorithme

Comment choisir l'une des méthodes vues ci-dessus en pratique ? La connaissance des points forts et des points faibles de chaque algorithme peut nous aider.

Méthode du plus proche voisin

Points forts :

— Donne de bons résultats lorsque les variables sont de types différents
— Possède d'excellentes propriétés théoriques
— Permet de créer des groupes dont la forme est très irrégulière
— Est robuste aux données aberrantes

Points faibles :

— Tend à former un grand groupe avec plusieurs petits groupes satellites
— Perd de l'efficacité si les vrais groupes sont de forme régulière
— Possède des propriétés théoriques ne semblant pas se réaliser en pratique dans certaines études

Méthode du voisin le plus distant

Points forts :

— Donne de bons résultats lorsque les variables sont de types différents
— Tend à former des groupes de taille égale

Points faibles :

— Est extrêmement sensible aux données aberrantes
— Tend à former des groupes de taille égale
— Est très peu utilisée en pratique

Méthode de la moyenne

Point fort :

— Tend à former des groupes de faible variance

Point faible :

— Tend à former des groupes de même variance

Méthode du centroïde

Point fort :

— Est très robuste aux données aberrantes

Point faible :

— Est peu efficace en l'absence de données aberrantes

Méthode de la médiane

Point fort :

— Est encore plus robuste en présence de données aberrantes

Point faible :

— Est très peu efficace en l'absence de données aberrantes

Méthode de Ward

Point fort :

— Est optimale si les observations sont approximativement distribuées selon une loi normale multidimensionnelle de même matrice de variances-covariances

Points faibles :

— Est sensible aux données aberrantes
— Tend à former des groupes de petite taille
— Tend à former des groupes de même taille

Choix du nombre de groupes

L'exécution d'un algorithme nous donne une séquence de n partitions ayant de n à 1 groupe. Quelle partition de cette séquence devrions-nous choisir ? C'est une question très difficile à laquelle il existe plusieurs réponses. En premier lieu, si on est familier avec les données à l'étude on peut voir si l'une des n partitions est particulièrement interprétable, si elle a un sens pratique. Sinon, alors voici trois critères disponibles dans les procédures de base des principaux logiciels.

Critère de classification cubique (CCC)

Règle du pouce pour son interprétation :

On fait habituellement un graphique avec le CCC en ordonnée et le nombre de groupes en abscisse. Pour le nombre de groupes, on ne considère que les partitions de $K = 1$ à $K = n/10$.

— Si $CCC > 2$, on est en présence d'une classification de bonne qualitéé
— Si $0 < CCC < 2$, on est en présence d'une classification de qualité moyenne.
— Si $CCC < 0$, on est en présence d'une classification de mauvaise qualité.
— Pour choisir le nombre de classes à retenir, on peut considérer les nombres de classes associés aux fortes hausses du critère CCC entre deux nombres de classes subséquents.
— On considère les « pics » atteignant des valeurs du critère supérieures à 2 ou à 3 comme étant de fortes hausses de ce critère.

Attention :

— Il ne faut pas utiliser le critère CCC avec la méthode du plus proche voisin, ou lorsque l'on suspecte que les groupes sont de forme très allongée ou irrégulière.
— Le critères ne fonctionne pas bien quand le nombre d'observations dans certains groupes est inférieur à 10.

8

Statistique pseudo-F

Le nom de cette statistique vient du fait qu'elle est presque distribuée selon une loi F lorsque la loi des données n'est pas trop loin de la normale multivariée avec des variances égales dans toutes les classes. Mais même si on est loin de la normalité, en pratique cette statistique peut quand-même être informative.

On cherche des nombres de classes pour lesquels la statistique du pseudo-F se démarque par une grande valeur. Sur un graphique de la statistique du pseudo-F en fonction du nombre de classes, ceci se traduit par la recherche de pics. Encore une fois, il ne faut pas utiliser la statistique du pseudo-F avec la méthode du plus proche voisin.

Statistique du pseudo-t^2

Le nom de cette statistique vient du fait qu'elle est presque distribuée selon une loi t lorsque la loi des données n'est pas trop loin de la normale multivariée avec des variances égales dans toutes les classes. En pratique, on regarde le graphique de la statistique du pseudo-t^2 en fonction du nombre de classes. En regardant ce graphique de droite à gauche, on essaie de trouver des valeurs de la statistique qui sont beaucoup plus élevées que la valeur précédente. Supposons que la forte hausse se produit entre k et $k-1$ classes. On choisit k classes dans le partitionnement de nos observations. Bien sûr, il ne faut pas utiliser la statistique du pseudo-t^2 avec la méthode du plus proche voisin.

Troisième partie

Méthodes supervisées

Dans la seconde partie du cours, nous couvrons les méthodes dites “d’apprentissage supervisé”. Contrairement aux situations de la première partie où nous n’avions, pour chacun de n objets, que les valeurs de p variables $X_1, \ldots, X_p$, nous avons en plus une “étiquette” Y_i, $i = 1, \ldots, n$ pour chacun des n objets. L’objectif des méthodes des prochains chapitres est de présenter des méthodes qui prennent les valeurs des p variables $X_1, \ldots, X_p$ en entrée et qui donnent une prévision de l’étiquette Y en sortie. C’est exactement ce que font les algorithmes d’apprentissage statistique (aussi appelé “apprentissage automatisé” ou en anglais “machine learning”). Les méthodes vues aux chapitres 4 et 5 sont les méthodes de base. Plusieurs cours offerts à l’Université Laval couvrent des méthodes plus sophistiquées (STT-2100, STT-4500, IFT-4027, GIF-4101, ACT-2003, ACT-2008).

Les exemples d’application de ces méthodes supervisées pleuvent. En voici quelques-uns :

— détection de fraude ;
— identification de nouveaux clients potentiels ;
— prévision du nombre ou du montant des réclamations ;
— filtrage de courriels indésirables ;
— prévision des gagnants lors d’événements sportifs ;
— simulation du déplacement d’animaux ;
— traduction automatisée de textes ;
— reconnaissance d’images.

Chapitre 4

Analyse discriminante

4.1 Introduction

4.1.1 Méthodologie de la classification supervisée

L'approche générale au problème de classification supervisée peut se résumer ainsi :

1. sélectionner un certain nombre d'individus dont on connaît le groupe d'appartenance ;
2. mesurer p caractéristiques $X_1, \ldots, X_p$ sur ces individus ;
3. diviser ce jeu de données en deux :
   — un jeu de données pour la modélisation (entrainement, "train") ;
   — un jeu de données pour la vérification (validation, "test").
4. développer modèle/algo pour classer le mieux possible les individus du jeu de données d'entrainement ;
5. évaluer notre modèle/algo sur le jeu de données de validation ;
6. répéter étapes 3-4-5 avec d'autres modèles/algo et choisir le meilleur.

Un grand nombre de méthodes de classification supervisée existent. Pour n'en nommer que quelques-unes ...

— l'analyse discriminante ;
— les arbres de classification ;

— la régression logistique et ses évolutions (GLM, GAM, GLMM, GAMM, etc.) ;
— les machines de vecteur de soutien (support vector machine, SVM) ;
— les réseaux de neurones.

Dans ces notes, nous ne couvrirons que les deux premières méthodes.

4.2 Définition de l'analyse discriminante

4.2.1 Notation et formulation du problème

C'est en 1936 que Sir R. A. Fisher a introduit la méthode. Il s'intéressait à la taxonomie végétale, p.ex. déterminer l'espèce de fleurs à partir de diverses mesures (Ronald A. Fisher (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179–188.)

Soit $\mathbf{X} = (X_{ij})$, qui est une matrice de dimension $n \times p$, où n est le nombre d'individus dans l'échantillon, p est le nombre de variables et X_{ij} est la valeur de la j^e variable pour le i^e individu. On définit

— I_k = ensemble des individus du groupe k ;
— $n_k = |I_k|$ = cardinalité de I_k ;
— $\Rightarrow n_1 + \cdots + n_q = n$, où q est le nombre de groupes.

On a donc des observations dans $\mathbb{R}^p$. Pour faire la classification à partir de $X_1, \ldots, X_p$, on doit **partitionner** $\mathbb{R}^p$ en q sous-ensembles de sorte que chacun des q sous-ensembles est associée à un des q groupes.

La stratégie de Fisher :

— passer de la dimension p à la dimension 1 en calculant un score

$$f(x_1, \ldots, x_p) \in \mathbb{R},$$

pour chaque observation ;
— utiliser ce score pour déterminer le groupe d'appartenance (donc partitionner $\mathbb{R}$).

Le score proposé par Fisher est une **combinaison linéaire des variables**, c'est-à-dire

$$\begin{aligned} f(X_1, \ldots, X_p) &= \mathbf{a}^\top \mathbf{X} + b \\ &= a_1 X_1 + \cdots + a_p X_p + b. \end{aligned}$$

On en déduira q intervalles de décision $\mathcal{I}_1, \ldots, \mathcal{I}_q$ associés aux groupes.

Sans perte de généralité, on peut choisir

$$-b = a_1 \bar{X}_1 + \cdots + a_p \bar{X}_p = \mathbf{a}^\top \bar{\mathbf{X}},$$

ce qui permet de centrer les variables en enlevant le vecteur de moyenne

$$\bar{\mathbf{X}} = \left(\tilde{X}_1, \ldots, \tilde{X}_p\right)^\top.$$

Il ne reste plus qu'à choisir le vecteur $\mathbf{a} = (a_1, \ldots, a_p)$.

4.2.2 Calcul du meilleur score

On voudrait choisir le vecteur $\mathbf{a}$ de sorte que les scores soient

— très différents entre les groupes ;

— très similaires à l'intérieur d'un groupe.

On s'intéresse donc à la variabilité des scores à l'intérieur des groupes et entre les groupes.

Étant donné $\mathbf{a} \in \mathbb{R}^p$, on a

$$\text{var}(\mathbf{a}^\top (X_1 \ \cdots X_p)^\top) = \mathbf{a}^\top \text{var}((X_1 \ \cdots X_p)^\top)\mathbf{a},$$

que nous estimons à partir des n observations par

$$\frac{1}{n}\, \mathbf{a}^\top \mathbf{S}\mathbf{a}.$$

La base de l'analyse discriminante repose sur le fait que

$$\mathbf{S} = \mathbf{W} + \mathbf{B},$$

où

$$\mathbf{W} = \text{matrice de variance intragroupe,}$$
$$\mathbf{B} = \text{matrice de variance intergroupe.}$$

(**W** pour "within" et **B** pour "between"). On peut prouver ce résultat en considérant la définition des matrices **S**, **W** et **B**.

La moyenne de la variable j pour tous les individus de l'échantillon est

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$$

La moyenne de la variable j pour les individus du groupe k est

$$\bar{X}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} X_{ij}$$

La somme des carrés totale est

$$s_{jj'} = \sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right) \left(X_{ij'} - \bar{X}_{j'}\right),$$

(On tirerait de la matrice **S** une estimation de $\text{cov}(X_j, X_{j'})$ si toutes les observations provenaient d'un même groupe.) On définit $s_{jj'}$ comme étant

$$s_{jj'} = w_{jj'} + b_{jj'}$$

où

$$w_{jj'} = \sum_{k=1}^{q} \sum_{i \in I_k} (X_{ij} - \bar{X}_{kj})(X_{ij'} - \bar{X}_{kj'}),$$
$$b_{jj'} = \sum_{k=1}^{q} n_k (\bar{X}_{kj} - \bar{X}_j)(\bar{X}_{kj'} - \bar{X}_{j'}).$$

On obtient

$$\widehat{\text{Var}}(\mathbf{a}^\top (X_1 \cdots X_p)^\top) = \frac{1}{n} \mathbf{a}^\top \mathbf{S} \mathbf{a} = \frac{1}{n} \left(\mathbf{a}^\top \mathbf{W} \mathbf{a} + \mathbf{a}^\top \mathbf{B} \mathbf{a}\right).$$

On se rappelle que l'on veut choisir le vecteur $\mathbf{a}$ pour que les scores puissent facilement séparer les groupes. En d'autres mots, on veut des scores les plus **similaires** possible à **l'intérieur** d'un groupe et des scores les plus **différents** possible **entre** les groupes.

On propose de choisir le vecteur $\mathbf{a} \in \mathbb{R}^p$ pour maximiser

$$\frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}} \quad \text{ou} \quad \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{S}\mathbf{a}} = \frac{1}{1 + \frac{\mathbf{a}^\top \mathbf{W}\mathbf{a}}{\mathbf{a}^\top \mathbf{B}\mathbf{a}}}\,.$$

Ce vecteur est unique à une constante près. On peut formuler le problème de trois manières équivalentes.

— Maximiser $\mathbf{a}^\top \mathbf{B}\mathbf{a}/\mathbf{a}^\top \mathbf{S}\mathbf{a}$ sous la contrainte que $\mathbf{a}^\top \mathbf{a} = 1$.

— Maximiser $\mathbf{a}^\top \mathbf{B}\mathbf{a}$ sous la contrainte que $\mathbf{a}^\top \mathbf{S}\mathbf{a} = 1$.

— Maximiser $\mathbf{c}^\top \mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}\mathbf{c}$ sous la contrainte que $\mathbf{c}^\top \mathbf{c} = 1$, où $\mathbf{c} = \mathbf{S}^{1/2}\mathbf{a}$.

En récrivant la 3e formulation

$$\mathbf{c}^\top \left(\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}\right) \mathbf{c},$$

du chapitre sur l'ACP, on se souvient qu'il faut prendre $\mathbf{a} = \mathbf{S}^{-1/2}\mathbf{c}$, où

$$\mathbf{c} = \text{vecteur propre normé}$$

associé à

$$\lambda_1 = \text{première valeur propre de } \mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}.$$

De façon équivalente, de la 2e formulation et du chapitre sur l'ACB on se souvient qu'on peut prendre

$$\mathbf{a} = \text{vecteur propre normé}$$

associé à

$$\lambda_1 = \text{première valeur propre de } \mathbf{S}^{-1}\mathbf{B}.$$

À noter que si

$$\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}\mathbf{c} = \lambda\mathbf{c} \quad \text{et} \quad \mathbf{a} = \mathbf{S}^{-1/2}\mathbf{c},$$

alors

$$\mathbf{S}^{-1/2}\mathbf{B}\mathbf{a} = \lambda\mathbf{S}^{1/2}\mathbf{a} \quad \Rightarrow \quad \mathbf{S}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}.$$

Les valeurs propres de $\mathbf{S}^{-1}\mathbf{B}$ et de $\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$ sont donc les mêmes !

La fonction discriminante de Fisher est donc

$$f(\mathbf{x}) = \mathbf{a}^\top \left(\mathbf{x} - \bar{\mathbf{X}}\right),$$

où $\mathbf{a}$ est le vecteur propre normé associé à la plus grande valeur propre de $\mathbf{S}^{-1}\mathbf{B}$.

Les scores $Y_i = \mathbf{a}^\top \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)$ sont les scores linéaires en $\mathbf{X}_i$ qui ont le rapport (variance inter)/(variance intra) le plus élevé.

Puisque la matrice $\mathbf{S}^{-1/2}\mathbf{B}\mathbf{S}^{-1/2}$ est symétrique et définie positive, ses valeurs propres sont toutes réelles et positives. De plus, on a que $\mathbf{S}^{-1}\mathbf{B}\mathbf{a} = \lambda_1\mathbf{a}$. Ainsi,

$$\begin{aligned}\mathbf{B}\mathbf{a} = \lambda_1\mathbf{S}\mathbf{a} &\Rightarrow \mathbf{a}^\top\mathbf{B}\mathbf{a} = \lambda_1\mathbf{a}^\top\mathbf{S}\mathbf{a} \\ &\Rightarrow \lambda_1 = \frac{\mathbf{a}^\top\mathbf{B}\mathbf{a}}{\mathbf{a}^\top\mathbf{S}\mathbf{a}}.\end{aligned}$$

On a donc

$$0 \leq \lambda_1 \leq 1.$$

La valeur propre λ_1 peut donc être vue comme le **pouvoir discriminant** de f :

- $\lambda_1 = 1 \Rightarrow \mathbf{a}^\top\mathbf{B}\mathbf{a} = \mathbf{a}^\top\mathbf{S}\mathbf{a}$, donc 100% de la variabilité entre les groupes et 0 variabilité à l'intérieur des groupes ;
- $\lambda_1 = 0 \Rightarrow \mathbf{a}^\top\mathbf{B}\mathbf{a} = 0$, donc 0 variabilité entre les groupes et 100% de la variabilité à l'intérieur des groupes.

4.3 Fonction discriminante et classification

4.3.1 Règle de classification

Après avoir défini la fonction discriminante $f(\mathbf{x})$, on peut calculer le score moyen de chaque groupe défini comme étant

$$m_k = \mathbf{a}^\top \left(\bar{X}_{k1}, \ldots, \bar{X}_{kp}\right)^\top,$$

où

$$\bar{X}_{kj} = \text{moyenne de la } j^e \text{ variable pour les individus appartenant au } k^e \text{ groupe.}$$

Considérons une nouvelle observation $\mathbf{X}_0 \in \mathbb{R}^p$. Pour classer ce nouvel individu dans un groupe de la population,

— on calcule son score $f(\mathbf{X}_0) = \mathbf{a}^\top \mathbf{X}_0$

— on l'assigne au groupe k_0 qui lui ressemble le plus, c'est-à-dire le groupe tel que

$$\left|\mathbf{a}^\top \mathbf{X}_0 - m_{k_0}\right| = \min_{1 \leq k \leq q} \left|\mathbf{a}^\top \mathbf{X}_0 - m_k\right|.$$

En appliquant cette règle à l'échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ lui-même, on peut estimer les risques de mauvaise classification avec la *matrice de confusion* :

	Classement			
Vrai groupe	Groupe 1	Groupe 2	$\cdots$	Groupe q
Groupe 1	p_{11}	p_{12}	$\cdots$	p_{1q}
Groupe 2	p_{11}	p_{12}	$\cdots$	p_{1q}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
Groupe q	p_{q1}	p_{q2}	$\cdots$	p_{qq}

Cas à 2 groupes

On peut montrer que le vecteur propre de l'analyse discriminante dans le cas où il n'y a que deux populations peut être défini ainsi :

$$\mathbf{a} = \mathbf{S}^{-1}\mathbf{C} = \sqrt{\frac{n_1 n_2}{n}}\, \mathbf{S}^{-1} \left(\tilde{\mathbf{X}}_1 - \tilde{\mathbf{X}}_2\right),$$

où

$$\mathbf{C} = \sqrt{\frac{n_1 n_2}{n}}\, (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2) \quad \text{et} \quad \mathbf{B} = \mathbf{C}\mathbf{C}^\top$$

et $\tilde{\mathbf{x}}_i$, $i = 1, 2$ sont les moyennes des caractérisiques $\mathbf{x}$ dans chaque groupe.

Supposons que

$$m_1 = \mathbf{a}^\top \tilde{\mathbf{x}}_1 > \mathbf{a}^\top \tilde{\mathbf{x}}_2 = m_2.$$

Alors, on classe un individu dans le premier groupe si

$$\mathbf{a}^\top \mathbf{x} > \bar{m} = \frac{m_1 + m_2}{2} = \mathbf{a}^\top \left(\frac{\tilde{\mathbf{x}}_1 + \tilde{\mathbf{x}}_2}{2} \right).$$

$$\Leftrightarrow (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)^\top \mathbf{S}^{-1} \mathbf{x} > (\tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2)^\top \mathbf{S}^{-1} \left(\frac{\tilde{\mathbf{y}}_1 + \tilde{\mathbf{x}}_2}{2} \right).$$

(Le facteur $\sqrt{n_1 n_2 / n}$ divise les deux côtés de l'inégalité. C'est pour cette raison qu'il n'y apparaît pas.)

4.3.2 Exemple

EXEMPLE **4.1** *49 hommes âgés ont été déclarés soit*

- *en bonne santé mentale (Groupe I) ou*
- *séniles (Groupe II)*

à la suite d'examens psychiatriques poussés. Les mêmes sujets ont été soumis à 4 tests standard beaucoup moins coûteux :

Test	*Groupe I* ($n_1 = 37$)	*Groupe II* ($n_2 = 12$)
Information	*12.57*	*8.75*
Similitudes	*9.57*	*5.33*
Arithmétique	*11.49*	*8.50*
Complétion de dessins	*7.97*	*4.75*

Estimation de Σ *Dans cette étude, on a trouvé*

$$\frac{\mathbf{S}}{n} = \begin{pmatrix} 11.2553 & 9.4042 & 7.1489 & 3.3830 \\ & 13.5318 & 7.3830 & 2.5532 \\ & & 11.5744 & 2.6170 \\ & & & 5.8085 \end{pmatrix}.$$

On trouve

$$\mathbf{C}^* = \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2 = (3.82, 4.24, 2.99, 3.22)^\top$$

et

$$\mathbf{C} = \sqrt{\frac{37 \times 12}{49}}\, \mathbf{C}^*.$$

En principe, on a

$$\mathbf{a} = \mathbf{S}^{-1}\mathbf{C},$$

mais on peut tout aussi bien utiliser

$$\mathbf{a} = n\mathbf{S}^{-1}\mathbf{C}^*$$

si ça nous tente, puisque $\mathbf{a}$ *n'est défini qu'à une constante multiplicative près.*

En effectuant les calculs décrits précédemment, on trouve

$$m_1 = \mathbf{a}^\top \tilde{\mathbf{x}} y_1 = 5.97 \quad \text{et} \quad m_2 = \mathbf{a}^\top \tilde{\mathbf{x}}_2 = 3.54,$$

ce qui conduit à déclarer un individu sénile si

$$\mathbf{a}^\top \mathbf{x} > \mathbf{a}^\top \left(\frac{m_1 + m_2}{2} \right) = 4.755.$$

Tableau sommatif

	Diagnostic clinique		*Totaux*
	"OK"	*"sénile"*	
Classé "OK"	*29*	**4**	*33*
Classé "sénile"	**8**	*8*	*16*
Totaux	*37*	*12*	*49*

Probabilités d'erreur

Taux global	$12/49 \approx 24.5\%$
Taux chez les "OK"	$8/37 \approx 21.6\%$
Taux chez les "séniles"	$4/12 \approx 33.3\%$

4.4 Approche alternative visant à minimiser la probabilité d'erreur de classement

4.4.1 Problème et notation

Au lieu d'utiliser la fonction discriminante de Fisher, on peut créer une procédure de classification qui minimise les risques d'erreur de classement. Pour simplifier les calculs, on se limite au cas où il n'existe que 2 groupes dans la population, disons Π_1 et Π_2. Sous certaines circonstances, cette méthode sera équivalente à l'utilisation de la fonction discriminante de Fisher.

Dans le cas où il n'existe que deux groupes dans la population, la matrice de confusion est de la forme suivante :

	Classement	
Vrai groupe	Positif	Négatif
Positif	Vrais Positifs (VP)	Faux Négatifs (FN)
Négatif	Faux Positifs (FP)	Vrais Négatifs (VN)

Sensibilité (*sensitivity*)

Probabilité qu'un individu qui appartient au groupe positif soit classé dans le groupe positif :

$$\frac{VP}{VP+FN}.$$

Spécificité (*specificity*)

Probabilité qu'un individu qui appartient au groupe négatif soit classé dans le groupe négatif :

$$\frac{VN}{VN+FP}.$$

Ce qui est difficile : avoir de hautes valeurs pour ces deux probabilités simultanément !

Parfois, un coût peut être associé au mauvais classement d'un individu. Le tableau suivant présente la notation liée au coût d'un mauvais classement :

	Classement		
Vrai groupe	Π_1	Π_2	Probabilité *a priori*
Π_1	0	$C(2\|1)$	q_1
Π_2	$C(1\|2)$	0	q_2

Pour $i \in \{1, 2\}$,

$$q_i = \mathrm{P}(\mathbf{X} \in \Pi_i)$$

et

$$C(3-i|i) = \text{coût associé au classement de l'individu ayant les caractéristiques } \mathbf{X} \text{ dans le groupe } \Pi_{3-i} \text{ alors que } \mathbf{X} \in \Pi_i.$$

Le choix du groupe d'appartenance d'un individu est basé sur une règle de décision. En fait, on aimerait déterminer une région $R_1 \subset \mathbb{R}^p$ telle que

$$\boxed{\mathbf{X} \in R_1 \Leftrightarrow \mathbf{X} \in \Pi_1.}$$

On pourrait aussi définir $R_2 = \mathbb{R}^p \setminus R_1$.

Comme on ne peut être parfait, on va fixer une zone de classement positif et une zone de classement négatif qui considèrent les probabilités et les coûts d'une mauvaise classification.

4.4.2 Probabilité et coût des mauvaises classifications

On pose

$$p_i(\mathbf{x}) = \text{densité des caractéristiques } \mathbf{X} \text{ d'un individu sachant qu'il provient de la population } \Pi_i$$

On peut ainsi calculer la probabilité de classer un individu de la population Π_i dans la population Π_{3-i} :

$$\begin{aligned} P(3-i|i) &\equiv P(\text{classé dans } \Pi_{3-i}|\text{vient de } \Pi_i) \\ &= P(\mathbf{X} \in \mathbf{R}_{3-i}|\text{vient de } \Pi_i) \\ &= \int_{\mathbf{R}_{3-i}} p_i(\mathbf{x})\, d\mathbf{x}. \end{aligned}$$

La probabilité d'erreur dépend

— des probabilités de mauvaise classification et

— des probabilités *a priori* d'observer un individu d'une population.

La probabilité de l'événement

« l'individu provient de la population Π_i et
il est classé dans la population Π_{3-i} »

est donnée par

$$q_i P(3-i|i).$$

Minimisation du coût espéré

Pour minimiser le coût espéré total

$$C(2|1)q_1 + \int_{R_1} \{C(1|2)q_2 p_2(\mathbf{x}) - C(2|1)q_1 p_1(\mathbf{x})\} d\mathbf{x},$$

il suffit de prendre

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p : \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \geq \frac{C(1|2)q_2}{C(2|1)q_1} \equiv \Omega \right\}.$$

En l'absence d'informations *a priori*, on suppose habituellement que

$$C(1|2) = C(2|1) \quad \text{et} \quad q_1 = q_2.$$

Dans ce cas, la règle de classification se simplifie à

$$\mathbf{x} \in \Pi_1 \Leftrightarrow \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} > 1 \Leftrightarrow \log \left\{ \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} \right\} > 0$$

4.4.3 Caractéristiques X sont de lois normales multivariées

Minimisons le coût espéré en ne supposant aucune information *a priori* et sous l'hypothèse de la normalité des caractéristiques :

$$\mathbf{p}_1 \sim \mathcal{N}_p\left(\mu_1, \mathbf{\Sigma}_1\right) \text{ et } \mathbf{p}_2 \sim \mathcal{N}_p\left(\mu_2, \mathbf{\Sigma}_2\right).$$

L'analyse discriminante linéaire suppose que $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_1 = \mathbf{\Sigma}$, alors que **l'analyse discriminante quadratique** relaxe cette contrainte. Pour la suite, nous nous concentrons sur l'analyse discriminante linéaire.

Sous les hypothèses précédentes, on a

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu_i\right)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)\right\},\ i = 1, 2.$$

Ainsi, $\log\{p_1(\mathbf{x})/p_2(\mathbf{x})\}$ devient

$$-\frac{1}{2}\left\{(\mathbf{x}-\mu_1)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_0) - (\mathbf{x}-\mu_2)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)\right\}.$$

On classe donc un individu dans la population Π_1 lorsque

$$-\frac{1}{2}\left\{\mu_1^\top \mathbf{\Sigma}^{-1}\mu_1 - \mu_2^\top \mathbf{\Sigma}^{-1}\mu_2 - 2\mathbf{x}^\top \mathbf{\Sigma}^{-1}(\mu_1-\mu_2)\right\} \geq 0,$$

ou encore, en factorisant cette dernière expression, quand

$$\left(\mathbf{x} - \frac{\mu_1+\mu_2}{2}\right)^\top \mathbf{\Sigma}^{-1}(\mu_1-\mu_2) \geq 0.$$

On remarque que la dernière expression est linéaire en $\mathbf{x}$, d'où le nom *d'analyse discriminante linéaire.*

De façon équivalente, on suppose que $\mathbf{x} \in \Pi_1$ si

$$(\mathbf{x}-\mu_1)\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1) < (\mathbf{x}-\mu_2)\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2),$$

c'est-à-dire lorsque

$$D_1^2(\mathbf{x}, \mu_1) \leq D_2^2(\mathbf{x}, \mu_2),$$

où D_k^2 est la distance de Mahalanobis associée à la k^e population. Quand $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, l'expression ci-dessus ne se simplifie pas et demeure quadratique en $\mathbf{x}$, d'où le nom *d'analyse discriminante quadratique.*

On peut aussi déclarer que $\mathbf{x} \in \Pi_1$ si

$$\mu_1^\top \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\,\mu_1^\top \mathbf{\Sigma}^{-1}\mu_1 \geq \mu_2^\top \mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\,\mu_2^\top \mathbf{\Sigma}^{-1}\mu_2,$$

que l'on peut récrire (exercice)

$$(\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}\mathbf{x} > (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}\left(\frac{\mu_1 + \mu_2}{2}\right).$$

Si on estime μ_1 par $\tilde{\mathbf{x}}_1$, μ_2 par $\tilde{\mathbf{x}}_2$ et $\mathbf{\Sigma}$ par $\mathbf{S}$, on retrouve la fonction discriminante de Fisher.

REMARQUE **4.1** *L'analyse discriminante linéaire est donc équivalente à la méthode de Fisher sous les conditions suivantes :*

— *les caractérisiques* $\mathbf{X}$ *suivent des lois normales multivariées dans chaque population ;*

— *la matrice de variance de* $\mathbf{X}$ *est égale pour toutes les populations ;*

— *la probabilité de provenir de chaque population est la même.*

Sous ces conditions, la fonction discriminante de Fisher n'est pas seulement optimale parmi les fonctions linéaires, mais bien la fonction qui minimise globalement la probabilité d'erreur de classification.

4.4.4 Probabilités de mauvaises classifications

Sous les mêmes hypothèses, on peut calculer les probabilités d'erreur. Posons

$$Y = \left(\mathbf{X} - \frac{\mu_1 + \mu_2}{2}\right)^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2).$$

Comme $\mathbf{X} \sim \mathcal{N}_p(\mu_i, \mathbf{\Sigma})$, $i \in \{1, 2\}$, on a

$$Y \sim \mathcal{N}\left[(-1)^{i-1}\frac{\zeta^2}{2}, \zeta^2\right],$$

où

$$\zeta^2 = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2)\ldots$$

En effet, dans le cas $i = 1$, on a

$$Y = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1} \left(\mathbf{X} - \frac{\mu_1 + \mu_2}{2} \right).$$

On obtient donc

$$\mathrm{E}(Y) = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1} \left(\mu_1 - \frac{\mu_1 + \mu_2}{2} \right) = \frac{\zeta^2}{2}.$$

De même, on a

$$\mathrm{Var}(Y) = (\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1} \mathbf{\Sigma} \mathbf{\Sigma}^{-1} (\mu_1 - \mu_2).$$

On obtient donc

$$\mathrm{var}(Y) = \zeta^2.$$

Le cas $i = 2$ se démontre de la même manière.

Dans le cas particulier où

$$q_1 = q_2 \quad \text{et} \quad C(1|2) = C(2|1),$$

on classe $\mathbf{X}$ dans population Π_2 lorsque $Y < 0$. Donc,

$$P(2|1) = \mathrm{P}(Y < 0 | \mathbf{X} \in \Pi_1) = \Phi\left(\frac{-\zeta^2/2}{\zeta} \right) = \Phi\left(-\frac{\zeta}{2} \right)$$

et

$$P(1|2) = \mathrm{P}(Y > 0 | \mathbf{X} \in \Pi_2) = 1 - \Phi\left(\frac{\zeta^2/2}{\zeta} \right) = \Phi\left(-\frac{\zeta}{2} \right).$$

Généralisation au cas à q groupes

Pour déterminer à quel groupe $\mathbf{X}$ appartient, on calcule

$$D_k^2 = (\mathbf{X} - \mu_k)^\top \mathbf{\Sigma}^{-1} (\mathbf{X} - \mu_k),$$

qui est la distance de Mahalanobis entre $\mathbf{X}$ et la moyenne μ_k de la k^e population. On assigne $\mathbf{X}$ à la population Π_{k_0} lorsque

$$D_{k_0}^2 = \min\left(D_1^2, \ldots, D_q^2 \right),$$

ce qui est équivalent à maximiser

$$\mathrm{P}\left(\mathbf{X} \in \Pi_{k_0} \right) = \frac{e^{-D_{k_0}^2/2}}{e^{-D_1^2/2} + \cdots + e^{-D_q^2/2}}.$$

4.4.5 Estimation

Comme d'habitude, on estime μ_i par la moyenne de $\mathbf{x}$ dans la population i. On pourrait estimer $\mathbf{\Sigma}$ par

$$\hat{\Sigma} = \mathbf{S}/n,$$

mais cette estimation est biaisée car les observations ne viennent pas toutes de la même population. Une estimation sans biais de $\mathbf{\Sigma}$ est donnée par

$$\hat{\Sigma} = \mathbf{S}_{\text{pool}} = \frac{1}{n-q}\,\mathbf{W}$$

où

$$\mathbf{W} = \sum_{k=1}^{q} \sum_{i \in I_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)\,(\mathbf{x}_i - \bar{\mathbf{x}}_k)\,.^{\top}$$

4.5 Autres considérations

4.5.1 Statistiques associées à des tests d'égalité de moyennes

On a vu que

— la fonction discriminante de Fisher est $f(\mathbf{X}) = \mathbf{a}^{\top}(\mathbf{X} - \bar{\mathbf{X}})$;

— le vecteur $\mathbf{a}$ est le premier vecteur propre de la matrice $\mathbf{S}^{-1}\mathbf{B}$;

— la valeur propre associée à ce premier vecteur propre est $\lambda_1 \in (0, 1)$. Cette valeur propre représente le pouvoir discriminant de la fonction discriminante de Fisher.

Cette valeur propre est appelée la statistique de Roy.

Note : Le vecteur $\mathbf{a}$ est aussi le vecteur propre de $\mathbf{W}^{-1}\mathbf{B}$ associé à la première valeur propre, ξ_1, de cette matrice. (Exercice !)

Outre la statistique de Roy, d'autres statistiques sont utilisées pour tester l'égalité des vecteurs $\mu_1, \ldots, \mu_q$.

— Trace d'Hotelling-Lawley : $\text{trace}\left(\mathbf{BW}^{-1}\right) = \sum \xi_i$;

— Trace de Pillai : $\text{trace}(\mathbf{BS}^{-1}) = \sum \lambda_i$;

— Plus grande racine de Roy : $\xi_1 = \lambda_1/(1-\lambda_1)$;

— Lambda de Wilks : $\Lambda = \dfrac{|\mathbf{W}|}{|\mathbf{S}|} = \dfrac{1}{\mathbf{BW}^{-1}+1} = \prod \dfrac{1}{1+\xi_i}$.

Les logiciels donnent souvent la valeur de p associée à chacun de ces tests.

REMARQUE **4.2** *Si ces statistiques ne conduisent pas au rejet de*

$$H_0 : \mu_1 = \mu_2 = ... = \mu_q,$$

une analyse discriminante ne sera pas très prometteuse ...

Bibliographie

Bandyopadhyay, S. and Saha, S. (2013). *Unsupervised Classification*. Springer-Verlag, Berlin.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Wiley, New York.

Harville, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2014). *An Introduction to Statistical Learning : with Applications in R*. Springer Texts in Statistics. Springer New York.

Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. SpringerLink : Bücher. Springer New York.

Lichman, M. (2013). UCI machine learning repository.

Weiss, S. M. W., Indurkhya, N. and Zhang, T. (2015). *Fundamentals of Predictive Text Mining*. Springer, New York, 2nd edition edition.