APM_4STA3_TA

Apprentissage statistique supervisé et non supervisé

Christine Keribin
Université Paris-Saclay

29 janvier 2025

Avant-propos

L'apprentissage est un ensemble de méthodes mathématiques et informatiques permettant d'acquérir des connaissances à partir de données souvent très complexes, c'est à dire d'y *apprendre* des informations à partir de procédures informatiques automatisées. L'apprentissage statistique - *statistical learning* - met en général l'accent sur la *modélisation* des phénomènes sous-jacents et leur compréhension, en utilisant les mathématiques pour fonder les procédures. L'apprentissage machine (automatique) - *machine learning* - en est le versant informatique où prime la performance de prédiction. L'intelligence artificielle - *AI*- s'attache à reproduire l'intelligence humaine à partir de programmes et matériels spécifiques. L'intersection entre ces disciplines peut être importante, en général autour de l'objectif de *prédiction*. La science des données (*Data Science*) est un mariage entre statistique, informatique et optimisation pour permettre de traiter les quantités gigantesques de données produites (*big data*). De façon plus récente, et pour le grand public et les média, IA est souvent devenu synonyme de techniques de *Réseaux de neurones*, une technique spécifique d'apprentissage, ce qui peut porter à confusion.

En 1997, Tom Mitchell a défini le *Machine Learning* de la façon suivante[1] : *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

On y retrouve les ingrédients fondamentaux du *Machine Learning* : l'expérience E (qui génère des données), la tâche T que l'on veut conduire sur ces données (qui mènera à différentes approches d'apprentissage), et la mesure de performance P (qui d'un point de vue statistique, pourra s'exprimer en fonction de risque associé au résultat proposé par la méthode). Les procédures donnent des résultats d'autant meilleurs qu'il y a plus de données. Autrement dit, le fait d'avoir beaucoup de données (pertinentes pour la tâche) permet d'avoir des résultats d'autant plus précis.

Il y a plusieurs approches au *Machine Learning* : apprentissage supervisé, non supervisé, semi-supervisé ou par renforcement par exemple. Récemment, les modèles génératifs et les grands modèles de langue (Large Language Model, LLM) ont envahi notre univers.

Nous introduirons dans ce cours les principes et méthodes de base en apprentissage non supervisé et supervisé, en laissant les problèmes spécifiques des grands volumes de données d'une part ou les méthodes plus avancées comme les réseaux de neurones d'autre part, pour les cours de deuxième année de master.

1. `http://www.cs.cmu.edu/~tom/`

— En apprentissage *supervisé*, une variable réponse Y est étudiée à partir de variables explicatives X à des fins de description ou de prédiction. Les techniques de régression paramétrique font partie des méthodes d'apprentissage supervisé. Son champ d'action s'est très largement diversifié : arbres et forêts aléatoires, SVM, réseaux de neurones, etc, passant d'un point de vue modélisation statistique à un point de vue optimisation.
La régression multiple est supposée être connue. Cette méthode fondamentale porte en elle les ingrédients de l'apprentissage, même si on réserve traditionnellement le vocable de *machine learning* à des méthodes plus évoluées. Nous développons dans un premier temps la méthodologie statistique dans le cadre de la régression paramétrique : non linéaire, logistique et classification supervisée, régression pénalisée, aussi bien d'un point de vue théorique (estimateurs du minimum de contraste, étude asymptotique), méthodologique (choix de modèle, mesure de performance) que pratique. Dans un second temps, nous présenterons un point de vue optimisation avec les techniques de régularisation et méthode à vecteurs supports.
— En apprentissage *non supervisé*, le problème est beaucoup moins bien posé : il s'agit de "découvrir" des items intéressants dans une population, sans bien savoir quels types d'items, ni leur nombre, ni quelles observations les présentent. On peut dégager deux types de problématiques : l'analyse exploratoire multidimensionnelle, permettant de réduire la dimensionalité des données et la classification non supervisée, (*clustering* en anglais, à ne pas confondre avec la classification supervisée), découvrant des groupes d'observations homogènes entre elles et très différentes entre les groupes. Dans ce cadre, nous présenterons la statistique exploratoire multidimensionnelle (techniques d'analyse factorielle : ACP, AFC, ACM), puis nous aborderons différentes techniques de *clustering* en anglais.

Les cours seront illustrés par des TPs avec le logiciel R sur des jeux de données provenant de différents domaines.

Table des matières

Première partie

Apprentissage non supervisé

Chapitre 1

De l'analyse descriptive à l'analyse multidimensionnelle

Dans une étude statistique, les *individus* (ou *unités statistiques*) sont décrits par des caractéristiques. Une variable collecte pour chaque individu la valeur observée d'une caractéristique déterminée. Elle a un domaine de définition $\mathcal{X}$, ensemble des valeurs que peut prendre l'observation pour un individu. Une variable est donc un vecteur à valeurs dans $\mathcal{X}^n$ si n est le nombre d'individus de l'échantillon. Une variable peut être quantitative (elle représente le résultat numérique d'une mesure) ou qualitative (elle représente une qualité non numérique : CSP, couleur, mode de déplacement...)

La statistique *descriptive* propose des graphiques et indicateurs adaptés à chaque type de variable, mais elle s'avère insuffisante sur des jeux de données multivariés. Nous présentons ici quelques rappels puis introduisons les thématiques développées dans la première partie du cours.

Variable qualitative. Elle représente une qualité du phénomène observé et prend ses valeur dans un ensemble fixé discret (numérique ou non). On parle de variable *nominale* ou *catégorielle*. Les valeurs autorisées sont appelées *niveau*, et la variable est un *facteur*. C'est le cas par exemple de la CSP.

- résumés numériques : table de comptage qui détermine le nombre d'observations dans chaque niveau du facteur (`table` dans R)
- visualisation : diagramme en bâtons (`barplot`) ou camembert (`pie`)

Variable quantitative. Elle représente une caractéristique mesurable (température, poids, solde bancaire, ...). Si la variable est entière, on parle de variable quantitative *discrète* (ex : nombre d'appels à un standard en une heure).

- résumés numériques :
  - tendance centrale : moyenne (`mean`), médiane (sépare la plage de variation en deux parties de même effectif, `median`), mode
  - dispersion : étendue (ou amplitude, `range`) ; quantiles (subdivision de la plage de variation en intervalles d'effectifs égaux, `quantile`) ; variance (moyenne des carrés de l'écart à la moyenne, `var`) et écart type (racine carrée de la variance, `sd`)
- visualisation : histogramme (`hist`, subdivision de la plage de variation de la variable en intervalles, la hauteur de chaque rectangle s'interprète comme une densité) ou diagramme en bâtons pour les variables discrètes

Statistique bivariée. Lorsque deux variables sont observées sur les mêmes individus, il est possible d'en faire une analyse conjointe, qui dépend de la nature de chacune d'elle :

- deux variables qualitatives : tables de contingence (`table`), diagrammes en barres et test du χ^2 d'indépendance (`chisq.test`)
- deux variables quantitatives : représentation par un nuage de points (`plot`), comparaison de moyenne (`t.test`), de variance (`var.test`), liaison (linéaire) entre les variables (`cor.test`), régression simple (`lm`)
- Une variable explicative qualitative et une variable à expliquer quantitative : boxplot de la variable à expliquer en fonction de la variable explicative, anova (`lm`)

Statistique multivariée. Il n'est plus possible de réprésenter graphiquement un individu dans l'espace à K dimensions formé à partir de $K > 3$ variables. Lorsqu'il y a plus de deux variables, les réprésentations graphiques combinent les représentations de toutes les paires de variables (`pairs`), mais la projection sur les plans canoniques n'est pas forcément celle qui conserve le plus d'information. L'analyse exploratoire multidimensionnelle va permettre de définir des plans qui déformeront le moins la représentation des données. Elle dépend du type des variables (qualitative, quantitative).

Les données se présentent donc sous la forme de (grands) tableaux X, les lignes représentant les unités statistiques ou individus ($i = 1, \ldots, I$), et les colonnes les variables ($k = 1, \ldots, K$). La cellule au croisement de la ligne i et de la colonne k contient la valeur x_{ik} que la variable k prend pour l'individu i.

$$x_{ik} = x_i[k] = X_k[i]$$

Nous avons vu que dans le cas d'une, deux ou trois variables, les visualisations conjointes sont simples, mais elles deviennent beaucoup plus complexes pour plus de trois variables. La statistique exploratoire (ou *analyse de données*) propose des outils pour explorer et visualiser de façon simplifiée ces grands tableaux, tout en conservant le maximum d'information.

- *multidimensionnelle* s'oppose à *unidimensionnelle* : on ne se focalise plus sur *une variable*, mais on représente l'unité statistique dans son ensemble par un *ensemble de variables* qui peut être très grand (réponses à un questionnaire, votes d'un député sur différents articles de loi, précipitations suivant les mois, notes d'un étudiant, etc...)
- *exploratoire* s'oppose à *inférentielle* : les données ne sont pas supposées être extraites d'une population dont on essaierait de définir des propriétés caractéristiques, mais étudiées pour elles-mêmes : il n'y a pas de définition d'un modèle sous-jacent, mais on essaie de déceler des liaisons entre les variables, des ressemblances ou différences entre les individus. Si on infère à une population plus générale dont l'échantillon serait issu, ce sera sans disposer de bornes de risque.

On parle aussi d'analyse *factorielle* : il s'agit de détecter des facteurs (combinaisons linéaires de variables) sous-jacents qui permettent d'interpréter la forme du nuage. Pour construire ces facteurs, les proximités entre éléments individus sont mesurées par la distance euclidienne. Il est également possible de définir une distance entre les variables. Deux nuages, l'un représentant les lignes et l'autre les colonnes, sont construits et représentés graphiquement, faisant *voir* des regroupements, oppositions, tendances, impossibles à discerner sur un grand tableau, même après un examen prolongé. Nous verrons d'ailleurs que les représentations en ligne et colonne sont fortement liées et permettent d'affiner l'interprétation. Les objectifs de l'analyse factorielle sont donc :

- l'exploration et l'interprétation : comparer des individus, des variables, comprendre le lien entre individus et variables, grouper.

— la réduction de la dimension : expliquer les mêmes données par des facteurs pertinents en moins grand nombre que le nombre de variables initiales tout en perdant le moins d'information possible.

Ces méthodes sont fondées sur des principes mathématiques de géométrie et d'algèbre linéaire, mais leur implémentation diffère suivant les types de variables considérées :

— ACP : variables quantitatives uniquement
— AFC : deux variables qualitatives à I et K niveaux respectivement
— ACM : strictement plus de deux variables qualitatives.

Classification non supervisée Si les méthodes précédentes peuvent permettre de faire apparaître sur des visualisations pertinentes des regroupements d'individus ou de variables, elles ne sont pas conçues pour segmenter les observations.

Les méthodes de *classification non supervisée* ou *clustering* permettent de partitionner des individus en groupes qui se ressemblent en n'ayant aucune information a priori sur les groupes. Elles font partie des méthodes d'apprentissage non supervisé, par opposition aux méthodes supervisées (ou méthode de classification, ou discrimination) qui déterminent une règle de *classement* à partir d'un échantillon d'apprentissage pour lequel le groupe d'appartenance est connu.

Nous présenterons la méthode des K-moyennes, la classification ascendante hiérarchique et les modèles de mélange.

Bibliographie Pagès (2005), Cornillon et autres (2008), Husson et al. (2016), Escofier et Pagès (2008), Lebart et al. (1995)

Chapitre 2

Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) est la technique d'analyse factorielle adaptée à un jeu de données où toutes les variables sont *quantitatives*, mais pouvant être de nature (très) différente. Par exemple :

- individu=ville ; variables = vitesse du vent, pluviométrie, ensoleillement, etc... x_{ik} contient la valeur de la variable k pour la ville i
- individu= produit ; variables = descripteur sensoriel (acidité, amertume, ..) x_{ik} mesure la note donnée par un jury au produit i concernant le descripteur k
- individu=entreprise ; variable= indicateur économique (taux de chômage, PIB, balance commerciale, etc)
- individu=enquêté ; variable=niveau d'accord sur les questions
- individu=gène ; variable=expression (différentielle) du gène suivant une condition de stress

2.1 Objectifs

Toute étude exploratoire multidimensionnelle commence par une étude préliminaire univariée, permettant d'examiner chaque variable, par exemple en calculant

$$\text{la moyenne : } \bar{X}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik} \text{ et l'écart type } s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}.$$

Dans l'analyse exploratoire, contrairement à la régression par exemple, il n'y a pas de variable particulière à expliquer en fonction des autres. C'est le profil sur l'ensemble des variables qu'on souhaite comparer d'un individu à l'autre : il faut donc définir la proximité (*distance*) entre individus et il y a $I(I-1)/2$ comparaisons possibles, où I peut être très grand.

De même, on souhaite pouvoir comparer les variables entre elles : peut-on avoir plus d'information que de le faire deux à deux ?

Enfin, à partir de la distance entre les individus, on pourra déterminer une *typologie* des individus (groupe d'individus homogènes du point de vue de leurs ressemblances) et mettre en évidence les profils de réponse type qu'on tentera d'interpréter en liaison avec des caractéristiques individuelles profondes, qui s'interpréteront à partir des variables.

Étude des individus

Chaque ligne x_i du tableau de données X est un individu, appelé également *profil ligne*, considéré comme un point de $\mathbb{R}^K$: les coordonnées sont les valeurs (numériques) observées pour chacune des K variables quantitatives. L'ensemble des I lignes du tableau de données forme le *nuage des individus* N^I.

La *distance euclidienne* entre deux profils ligne mesure la ressemblance entre deux individus

$$d^2(i,i') = \sum_k (x_{ik} - x_{i'k})^2 = ||\overrightarrow{M_i M_{i'}}||^2 \tag{2.1}$$

où M_i est un point de $\mathbb{R}^K$ de coordonnées x_i. Deux individus sont proches s'ils répondent de façon identique à chaque variable. Analyser la variabilité entre les individus, revient à étudier l'ensemble des distances interindividuelles, c'est à dire la forme du nuage des individus. Si chaque individu est affecté d'un poids $p_i \in]0;1[$ tels que $\sum_i p_i = 1$, le barycentre G des points M_i affecté des coefficients p_i est le centre de gravité G du nuage. G est de coordonnées $(\bar{X}_1, \ldots, \bar{X}_K)$.

Définition 1. *L'*inertie *du nuage de points* $(M_1, \ldots M_I)$ *affectés des poids* $p_i \in]0;1[$ *tels que* $\sum_i p_i = 1$ *est définie par*

$$Inertie = \sum_i p_i ||\overrightarrow{GM_i}||^2.$$

L'inertie peut être interprétée comme la variabilité des distances inter-individuelles :

$$\sum_{i,j} p_j p_j ||\overrightarrow{M_i M_j}||^2 = 2 \sum_i p_i ||\overrightarrow{GM_i}||^2.$$

L'étude du nuage des individus passe par

- la détermination de directions dans laquelle la globalité du nuage des individus sera le mieux représentée
- faire apparaître des groupes d'individus
- proposer un explication du jeu de données : lier les typologies d'individu avec des caractéristiques des variables.

L'exemple jouet suivant est extrait de Pagès (2005). On considère un nuage de $\mathbb{R}^3$ comprenant $I = 6$ individus dont on étudie simultanément $K = 3$ propriétés physiques.

nom	*Longueur*	*Largeur*	*Poids*
A	0	5	0
B	1	4	1
C	2	3	2
D	3	2	2
E	4	1	1
F	5	0	0
mean	2.5	2.5	1
sd	1.708	1.708	0.816

La figure 2.1 représente le nuage, et sa projection sur différents plan. Dans cet exemple de $\mathbb{R}^3$, il est facile de remarquer que tous les points sont dans un même plan, et la projection orthogonale sur ce plan ne déforme pas le nuage. Les individus `A` et `F` s'opposent en longueur et largeur, tandis que `C` et `D` s'opposent à `A` et `F` en poids. L'ACP va permettre de trouver automatiquement les directions de projection permettant de déformer le nuage le moins possible.

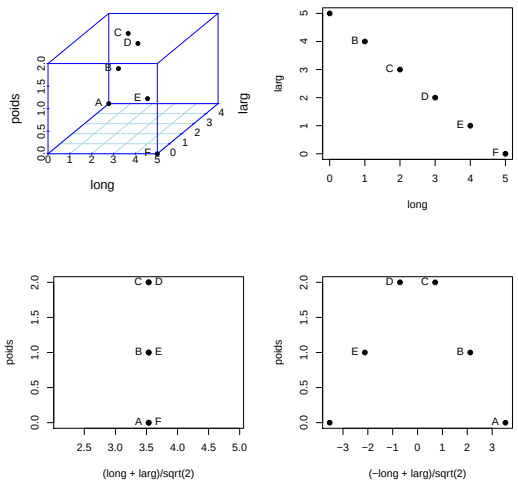


FIGURE 2.1 – ACP exemple jouet : le nuage en 3D et projeté orthogonalement sur différents plans

Transformation des données

Avant d'étudier le nuage des individus, on lui fait en général subir deux transformations :

— le vecteur des moyennes $(\bar{X}_k)$ est un point de $\mathbb{R}^K$ situé au centre de gravité G_I du nuage. L'opération de *centrage* $x_{ik} - \bar{X}_k$ déplace le centre de gravité à l'origine O du repère. Cette transformation ne modifie pas la forme du nuage, et sera *toujours effectuée*.

— La *réduction* $(x_{ik} - \bar{X}_k)/s_k$ où s_k est l'écart type de la variable k modifie la forme du nuage en harmonisant sa variabilité dans toutes les directions de base. C'est *indispensable* quand les variables ne s'expriment pas dans les mêmes unités, *souhaitable* pour donner le même poids à toutes les variables. Lorsque cette opération est effectuée, on parle d'*ACP normée*.

Tous les individus ont a priori le *même poids* $\frac{1}{I}$. On peut généraliser à un poids quelconque, c'est l'*ACP pondérée*.

Étude des variables

Une variable est une colonne X_k du tableau X et peut être considérée comme un point dans l'espace vectoriel $\mathbb{R}^I$. L'ensemble des variables forme le *nuage des variables* N_K. La proximité entre deux variables peut se définir par leur liaison (linéaire), c'est à dire leur *corrélation*. Il y a $K(K+1)/2$ couples à examiner, et comme dans le cas du nuage des individus, l'ACP va apporter un outil de synthèse :

— déterminer des directions dans laquelle la globalité du nuage des variables sera le mieux représentée

— faire apparaître des groupes de variables interprétables ; dans un groupe, les variables (fortement corrélées) sont considérées comme des mesures d'un même *facteur* sous-jacent

— proposer un indicateur synthétique : résumer un ensemble de variables à un petit nombre

On pourrait discuter de restreindre l'étude de la liaison à celle de la liaison linéaire. Si cette

dernière est très utilisée, c'est parce que
- elle est souvent observée
- on ne regarde pas une seule liaison, mais l'ensemble des liaisons entre les variables prises deux à deux, ce qui permet d'enrichir l'étude
- on peut commencer en linéaire, puis étendre si cela ne suffit pas.

Pour l'étude du nuage des variables centrées,
- dans le cas équipondéré, chaque individu apporte un poids de $1/I$. Ainsi, à chaque fois que la dimension associée à un individu apparaît, elle doit faire intervenir le coefficient $1/I$. Dans $\mathbb{R}^I$, le produit scalaire s'écrit

$$\forall u \in \mathbb{R}^I, \forall v \in \mathbb{R}^I, \ < u, v >_I = \sum_i \frac{1}{I} u_i v_i \tag{2.2}$$

- Pour deux variables *centrées* X_k et $X_{k'}$, le cosinus de l'angle $\theta_{kk'}$ formé par les vecteurs qui les représentent est égal au coefficient de corrélation empirique entre ces deux variables

$$\begin{aligned} < X_k, X_{k'} >_I &= \frac{1}{I} \sum_{i=1}^{I} x_{ik} x_{ik'} = cov(X_k, X_{k'}) \\ &= s_k s_{k'} \ cor(X_k, X_{k'}) = ||X_k||_I ||X_{k'}||_I \cos(\theta_{kk'}) \end{aligned}$$

Avec ce produit scalaire dans $\mathbb{R}^I$,
- deux variables non corrélées sont orthogonales.
- Si toutes les variables sont centrées réduites, elles sont toutes de rayon 1, et leur extrémité appartient l'hypersphère de $\mathbb{R}^I$ de rayon 1.
- Le centrage des variables ne déplace pas l'origine des axes des variables.

Ainsi, la représentation dans un plan de $\mathbb{R}^I$ des variables se trouve circonscrite dans le cercle de rayon 1, appelé *cercle des corrélations*, et l'angle entre deux variables est un moyen de visualisation de leur corrélation.

Dans l'exemple jouet, la corrélation entre longeur et largeur vaut -1, les variables sont anti-corrélées. Ces deux variables sont orthogonales au poids. Les trois variables sont donc parfaitement représentée dans un espace vectoriel de dimension 2, engendré par le vecteur normé v_1 colinéaire à (-long+larg), et le vecteur normé v_2 colinéaire au poids. Les coordonnées des variables initiales dans ce sous-espace sont :

coordonnées des variables sur l'axe	v_1	v_2
longueur	-1	0
largeur	1	0
poids	0	1

Longueur et largeur s'opposent sur le premier axe qui représente un facteur sous-jacent. Le poids est parfaitement corrélé au deuxième axe et décorrélé (orthogonal) à longueur et largeur. La mise en relation du graphe des individus et des variables permet de préciser l'interprétation :
- point de vue des individus : l'axe 1 oppose les individus A et F qui présentent des valeurs très différentes pour les deux premières variables
- point de vue des variables : l'axe 1 oppose `longueur` et `largeur`, corrélées négativement et décorrélées de la variable `poids`
- En mettant les deux représentations en perspective,

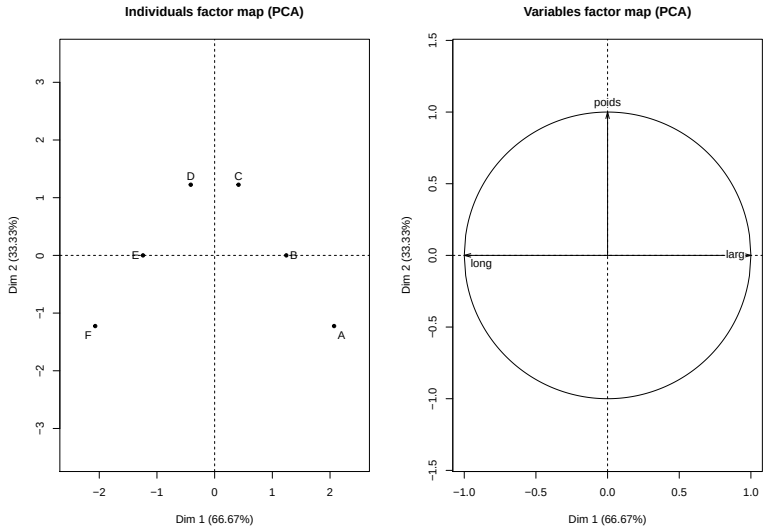


FIGURE 2.2 – ACP exemple jouet : nuage des individus et cercle des corrélations

Meilleure représentation des nuages

Des constatations précédentes découle l'étude de deux nuages :
- le nuage N_I des individus (points dans $\mathbb{R}^K$) où la variabilité des individus est mesurée par distance euclidienne inter-individus,
- le nuage N_K des variables (vecteurs dans $\mathbb{R}^I$) où la liaison entre variables est mesurée par leur corrélation (angle inter variables).

L'ACP permet
- de comparer les individus profils lignes,
- de comparer les variables profils colonnes,
- de permettre une représentation dans un espace de faible dimension approchant le mieux possible chacun des deux nuages,
- de relier les deux représentations, pour faciliter l'interprétation des individus en en liaison avec des caractéristiques individuelles profondes (*facteurs*), interprétées à partir des variables,

et ceci pour une dimension rendant la visualisation directe impossible.

Dans la suite, on se place dans le cadre de l'ACP normée.

2.2 Ajustement du nuage des individus N_I dans $\mathbb{R}^K$

Un individu i est formé par les valeurs observées sur les K variables. C'est une ligne du tableau X, représentée comme un point $x_i \in R^K$, muni de la distance euclidienne

$$d^2(x_i, x_{i'}) = \sum_k (x_{ik} - x_{i'k})^2.$$

On suppose que chaque individu a le même poids $p_i = 1/I$ dans l'analyse. On se place dans le cadre de l'ACP normée, les variables X_k sont donc centrées et réduites. Ainsi, le centre du repère O est le centre d'inertie (barycentre) G du nuage. Il s'agit de trouver une représentation dans un espace de faible dimension (un pour un axe, ou deux pour un plan), telle que la projection du nuage sur cet espace soit la moins déformée possible. La projection étant une application affine, elle conserve les barycentres : le barycentre des points projetés est la projection du barycentre. Ainsi, le centre de gravité est invariant par la transformation recherchée et le nuage projeté est centré autour de son centre de gravité.

On commence donc par la recherche d'une direction de $\mathbb{R}^K$ telle que la projection orthogonale sur cette direction déforme le moins possible le nuage : ie, les distances entre les points initiaux et leurs projections sont les plus faibles possibles, ou les distances projetées sont les plus grandes possibles (Pythagore).

Ajustement suivant une direction

On définit
- $u \in \mathbb{R}^K$, vecteur unitaire de la direction cherchée
- $M_i = (x_{ik})_{k=1,\dots,K}$ un point du nuage N_I de $\mathbb{R}^K$
- H_i la projection de M_i sur u. Soit $OH_i = X[i,]u = \sum_j x_{ik}u_k$; le vecteur de toutes les distances signées est calculé par

$$Xu = \begin{pmatrix} OH_1 \\ \vdots \\ OH_n \end{pmatrix}$$

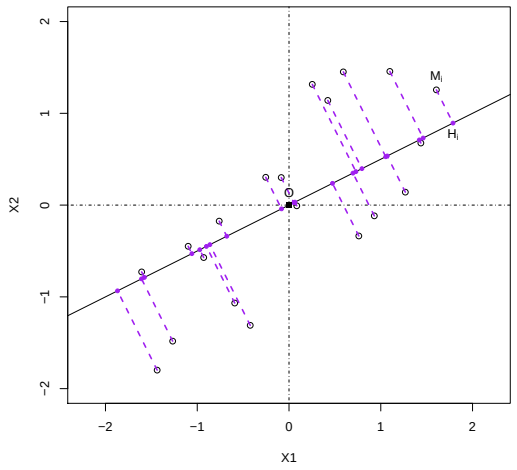


FIGURE 2.3 – Projection orthogonale du nuage sur une direction

Le problème devient : trouver u rendant maximum

$$Inertie(u) = \frac{1}{I}\sum_i OH_i^2 = \frac{1}{I}u'X'Xu = u'Cu$$

où $1/I$ est le poids de chaque individu et C la matrice de corrélation. Ce critère (*inertie de la projection*) s'interprète alors comme une variance car le nuage est centré (c'est la variance de la projection Xu). Il peut être exprimé en terme de distances interindividuelles : rendre maximum $\sum_{i,i'} d^2(H_i, H_{i'})$. En effet, O étant le barycentre des n individus dans le nuage recentré on a $\sum_i \overrightarrow{OH}_i = 0$ et :

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{I} d^2(H_i,H_j) &= \sum_i \sum_j ||H_iH_j||^2 \\
&= \sum_i \sum_j ||OH_i||^2 + ||OH_j||^2 - 2 < \overrightarrow{OH}_i, \overrightarrow{OH}_j > \\
&= 2I \sum_i ||OH_i||^2 - 2 < \sum_i \overrightarrow{OH}_i, \sum_j \overrightarrow{OH}_j > \\
&= 2I \sum_i ||OH_i||^2 \\
&= 2I^2 Inertie(u)
\end{aligned}$$

Théorème 1. *La direction qui maximise l'inertie projetée du nuage des individus est la direction associée à la plus grande valeur propre de la matrice de corrélation.*

Preuve. On cherche à maximiser $u'Cu = \sum_{ij} c_{ij}u_iu_j$ sous la contrainte $u'u = 1$, où $C = \frac{1}{I}X'X$ est la matrice de corrélation quand le nuage est centré-réduit. Il s'agit de la recherche d'un extrémum de formes quadratiques sous contraintes quadratiques. On annule les dérivées du Lagrangien $L = u'Cu - \lambda(u'u - 1)$ où λ est un multiplicateur de Lagrange. Or,

$$\frac{\partial u'Cu}{\partial u} = 2Cu$$

d'où

$$2Cu - 2\lambda u = 0$$

soit :

$$Cu = \lambda u.$$

u est un vecteur propre normé de C associé à la plus grande valeur propre λ, qui s'interprète comme l'inertie du nuage projeté. En effet, en multipliant par u',

$$\textit{Inertie projetée} = u'Cu = \lambda$$

L'inertie projectée λ sur la direction u a bien été maximisée. ◊

Appelons λ_1 cette valeur propre et u_1 le vecteur propre normé associé. Cette direction n'est pas en général celle de la droite de régression linéaire.

Ajustement suivant un plan

Le même principe prévaut pour une représentation plane : trouver un plan P tel que $\frac{1}{I}\sum_i OH_i^2$ soit maximum, où H_i est maintenant la projection sur P.

Or u_1 appartient forcément à P. On cherche alors u_2 orthogonal à u_1 et maximisant la projection, en annulant les dérivées du Lagrangien suivant : $L_2 = u_2'Cu_2 - \lambda_2(u_2'u_2 - 1) + \mu_2u_2'u_1$. La condition d'extremum pour u_2 donne :

$$2Cu_2 - 2\lambda_2u_2 - \mu_2u_1 = 0$$

En multipliant par u_1' et en utilisant les contraintes

$$2u_1'Cu_2 - \mu_2 = 0.$$

Or, $Cu_1 = \lambda_1 u_1 \Leftrightarrow u_1'C = \lambda_1 u_1'$, d'où $\mu_2 = 0$. Il reste donc à résoudre

$$Cu_2 = \lambda_2 u_2.$$

u_2 est le second vecteur propre relatif à la seconde plus grande valeur propre de $C = \frac{1}{I}X'X$.

Reconstruction du nuage

Ainsi, on peut construire une suite de vecteurs $u_1, \ldots, u_s$, orthogonaux, qui maximisent l'inertie du nuage projeté sur un espace de dimension s : pour chaque s, u_s maximise $\frac{1}{I}\sum_i (OH_i^s)^2$ sous la contrainte $u_s \perp u_t$ pour tout $t < s$.

u_s est ainsi vecteur propre de la matrice des corrélations $C = \frac{1}{I}X'X$ associé à la valeur propre λ_s de rang s. Les valeurs propres sont rangées en ordre décroissant et représentent l'inertie de la projection du nuage sur l'axe concerné :

$$Cu_s = \lambda_s u_s;\ \lambda_s = \frac{1}{I}\sum_i (OH_i^s)^2;\ u_s \perp u_t \text{ et } \lambda_s < \lambda_t \text{ pour tout } t < s$$

Les calculs de diagonalisation sont faits par les logiciels. Lorsque l'on utilise tous les vecteurs propres, la reconstruction du nuage est complète :

$$X = \sum_s X u_s u_s'$$

L'ACP peut ainsi être vue comme un changement de base dans $\mathbb{R}^K$, dans lequel les premiers vecteurs de la base jouent un rôle particulier et vont être conservés.

Propriété 1. *L'inertie totale du nuage des individus N_I vaut K en ACP normée.*

En effet,

$$\text{Inertie de } N_I \ = \sum_i \frac{1}{I} d^2(O, M_i) = \sum_i \frac{1}{I}\sum_k (x_{ik})^2 = K = \sum_k \frac{1}{I}\sum_i (x_{ik})^2 = \sum_k ||X_k||^2$$

Définition 2. *Les vecteurs propres normés u_s sont appelés vecteurs principaux et engendrent les axes principaux. On appelle composante principale associée à l'axe principal u_s, le vecteur $F_s = Xu_s$ des projections des individus sur u_s. Les u_s vus comme forme linéaire du dual de $\mathbb{R}^K$ sont appelés facteurs principaux.*

Note : en ACP normée, les vecteurs principaux et facteurs principaux ont les mêmes valeurs. Mais ce n'est pas le cas en ACP non normée où les facteurs principaux, définissant les coefficients de la combinaison linéaire des variables X_k permettant de calculer les composantes principales, ont des valeurs différentes de celles des vecteurs principaux normés.

Aide à l'interprétation

Des indicateurs permettent de quantifier les interprétations :

— Le *pourcentage d'inertie* associé à un axe est le rapport entre l'inertie du nuage projeté et l'inertie totale

$$\frac{\sum_i \frac{1}{I}(OH_i^s)^2}{\sum_i \frac{1}{I}(OM_i)^2} = \frac{\lambda_s}{\sum_s \lambda_s}$$

qui vaut λ_s/K si l'ACP est normée. Il mesure :
  — la *qualité de représentation des données* ou variabilité exprimée par l'axe. En ACP normée, si $\lambda_s < 1$, l'axe v_s représente moins de données qu'une variable isolée
  — l'importance relative des axes. Du fait de l'orthogonalité, les pourcentages somment à 1.

— La qualité de représentation d'un individu ou *cos carré*. On applique l'idée précédente à un individu :

$$\frac{(OH_i^s)^2}{(OM_i)^2} = \cos^2(\theta_i^s)$$

— La *contribution d'un individu* à un axe, part de chaque individu dans l'inertie associée à un axe :

$$\frac{\frac{1}{I}(OH_i^s)^2}{\lambda_s} \times 100$$

Lorsqu'un individu contribue fortement à la construction d'un axe factoriel, il est fréquent que les résultats d'une nouvelle ACP construite sans cet individu changent de façon substantielle. Les facteurs peuvent changer, et de nouvelles oppositions entre individus apparaître

— les *individus remarquables* : ils sont loin de l'origine. On calcule $(OM_i)^2$.

La qualité de représentation est bien une information distincte de la contribution : il peut y avoir des points bien représentés sur un axe, mais peu contributifs, par exemple parce qu'ils sont près de l'origine ; des points contributifs à un axe, mais mal représentés, parce qu'ils sont loin de l'origine dans toutes les directions. Un point proche de l'origine indique qu'il est proche des caractéristiques moyennes représentées par les directions considérées.

On peut regarder ces informations sur un axe, ou sur un plan en sommant les contributions (ou les qualités de représentations) des axes considérés, grâce aux propriétés des projections orthogonales.

2.3 Ajustement du nuage des variables N_K dans $\mathbb{R}^I$

On rappelle que les K variables X_k sont centrées réduites et le produit scalaire est défini comme la corrélation entre les variables. Comme les variables sont normées, leurs extrémités appartiennent à l'hypersphère de rayon 1. Chaque variable X_k étant normée, son poids est 1, et l'inertie est la somme des inerties de chaque variable

$$Inertie = \sum_k ||X_k||^2_{\mathbb{R}^I} = \sum_k \sum_i \frac{1}{I} x_{ik}^2$$

L'étude du nuage des variables suit la même démarche que celle du nuage des individus. On recherche dans $\mathbb{R}^I$ une suite d'axes orthogonaux telle que l'inertie projetée de N_K soit maximum. On note

— v_s, un vecteur unitaire de la direction cherchée de rang s
— $X_k = (x_{ik})_{i=1,\dots,I}$ une variable du nuage N_K dans $\mathbb{R}^I$
— T_k^s la projection de X_k sur v_s

En ACP normée,

$$T_k^s =< X_k, v_s >_{\mathbb{R}^I} = ||X_k|| \cos(v_s, X_k) = \text{cor}(v_s, X_k)$$

est la corrélation entre la variable X_k et la nouvelle variable v_s, combinaison linéaire des variables initiales. On recherche donc une variable synthétique v_s normée qui est la plus corrélée avec chacune des autres variables. Le critère d'inertie projeté s'écrit

$$\text{Inertie projetée} = \sum_k (T_k^s)^2 = \sum_k [\cos(v_s, X_k)]^2 = \sum_k [\text{cor}(X_k, v_s)]^2 \text{ maximum}$$

où v_s est la nouvelle variable la plus corrélée à l'ensemble des K variables initiales. Les variables synthétiques fournissent des plans dans lesquels on peut représenter les variables et interpréter leurs intercorrélations. Les extrémités des variables étant initialement sur l'hypersphère, leur projection sur un plan est contenu dans le cercle des corrélations, cf figure 2.2. Dans le cas de l'exemple jouet, l'une des variables synthétiques est de direction `largeur-longueur`. Pour trouver v_s, remarquons que

$$\text{cor}(X_k, v_s) = \frac{1}{I} \sum_i x_{ik} v_s[i] = \frac{1}{I} [X_k]' v_s; \quad k = 1, \ldots; K$$

Il s'agit donc de trouver v_s tel que $\sum_k \text{cor}(X_k, v_s)^2 = v' \frac{1}{I^2} XX'v$ soit maximum sous les contraintes $||v||_{R^I} = 1$ et $v_s \perp v_t$ pour $s < t$. Le lagrangien pour la recherche de la première direction s'écrit

$$L(v, \mu) = v' \frac{XX'}{I^2} v - \mu(\frac{v'v}{I} - 1)$$

Il s'agit donc de rechercher les valeurs propres et vecteurs propres d'une matrice de dimension $I \times I$, en général beaucoup plus grande que C.

Recherche des directions propres

Les axes v_s sont solutions de

$$\frac{\partial L}{\partial v} = \frac{2}{I} \left(\frac{XX'}{I} v - \mu v \right) = 0$$

Ce sont les vecteurs propres de la matrice des produits scalaires entre les individus $\frac{1}{I} XX'$, de taille $I \times I$, qui peut être déduite des vecteurs propres de la matrice de corrélation C.

Théorème 2. *La recherche des directions propres maximisant la corrélation avec les variables amène à rechercher les valeurs propres et vecteurs propres de XX^T :*

- *les valeurs propres sont* les mêmes *que celles calculées dans l'espace des individus ;*
- *les vecteurs propres se déduisent également de ceux du nuage des individus :*

$$v_j = \frac{1}{\sqrt{\lambda_j}} X u_j = \frac{1}{\sqrt{\lambda_j}} \text{F}_j$$

- *L'inertie du nuage des variables vaut K.*

Preuve. — les valeurs propres sont les mêmes que celles du nuage des individus. En effet, dans le nuage des individus, on a

$$\frac{1}{I} X'X u_s = \lambda_s u_s; \quad \frac{1}{I} XX' v_s = \mu_s v_s$$

En multipliant la première égalité par X, on obtient que Xu_s est vecteur propre de $\frac{1}{I} XX'$ associé à la valeur propre λ_s.

— les vecteurs propres s'obtiennent à partir des vecteurs propres du nuage des individus. En effet, le carré de la norme de $Xu_s \in R^I$ vaut

$$||Xu_s||^2 =< Xu_s, Xu_s >= u_s' \left(\frac{X'X}{I} u_s \right) = \lambda_s u_s' u_s = \lambda_s$$

d'où

$$||\frac{u_s' X'}{\sqrt{\lambda_s}}||^2_{R^I} = \frac{1}{I} \frac{u_s' X'}{\sqrt{\lambda_s}} \frac{X u_s}{\sqrt{\lambda_s}} = 1$$

$$v_s = \frac{1}{\sqrt{\lambda_s}} X u_s = \frac{1}{\sqrt{\lambda_s}} F_s; \quad F_s[i] = OH_i^s$$

où on reconnaît la composante principale F_s, ensemble des projections des individus sur l'axe u_s.

— Le critère au maximum vaut l'inertie projetée sur la direction (rappel : le produit scalaire dans $\mathbb{R}^I$est normalisé par $\frac{1}{I}$)

$$L(\lambda_s, v_s) = \sum_k (T_k^s)^2 = \frac{1}{I} v_s' \frac{XX'}{I} v_s = \lambda_s \frac{v'v}{I} = \lambda_s$$

L'inertie totale du nuage des variables

$$\text{Inertie de } N_K \ = \sum_k ||X_k||^2 = K = \sum_{k=1}^{K} \sum_{s=1}^{K} (T_k^s)^2 = \sum_s \lambda_s$$

vaut K en ACP normée.

$\Diamond$

Définition 3. *Soit F_s est la* composante principale *d'ordre s. La composante principale normée $v_s = F_v/\sqrt{\lambda_s}$ est appelée* axe factoriel *d'ordre s. C'est un axe principal dans l'espace des variables.*

Remarque : Lorsqu'une variable n'est pas normée, sa longueur est égale à son écart-type. En ACP non normée, chaque variable est donc affectée d'un poids égal à sa variance s_k^2. La projection T_k^s de la variable X_k sur l'axe v_s est donc $s_k \cos(v_s, X_k)$ et le critère s'écrit

$$\sum_k (T_k^s)^2 = \sum_k s_k^2 [\text{cor}(X_k, v_s)]^2$$

Aide à l'interprétation

On a donc remplacé $X_1, \ldots, X_K$ variables corrélées par de nouvelles variables $F_1, \ldots, F_k$, appelées composantes principales, non corrélées entre elles, combinaisons linéaires des précédentes, et de variance maximum. Elles sont normalisées en facteurs, qui sont des variables synthétiques formant les axes principaux de l'espace des variables. Des indicateurs permettent de quantifier les interprétations :

— le *pourcentage d'inertie associé à un axe* est le rapport entre l'inertie du nuage des variables projeté et l'inertie totale

$$\sum_k \frac{(T_k^s)^2}{\sum_k ||X_k||^2} = \frac{\lambda_s}{\sum_s \lambda_s}$$

qui vaut λ_s/K si l'ACP est normée. Il mesure

- la *qualité de représentation des variables* ou variabilité exprimée par l'axe. En ACP normée, si $\lambda_s < 1$, l'axe v_s représente moins d'information qu'une variable isolée
- l'importance relative des axes. Du fait de l'orthogonalité, les pourcentages somment à 1.

— La qualité de représentation d'une variable ou *cos carré* :

$$\frac{(T_k^s)^2}{||X_k||^2} = \cos^2(\theta_k^s) = (T_k^s)^2$$

La qualité de représentation d'une variable se lit directement sur le graphique en regardant la proximité de sa projection avec le cercle des corrélations.

— Les variables étant toutes de poids 1 en ACP normée, leur contribution est proportionnelle au cos carré et en général non affichée.

$$\frac{(T_k^s)^2}{\lambda_s} = \frac{\cos^2(\theta_k^s)}{\lambda_s}$$

Ces informations permettent de quantifier les informations graphiques portées sur le cercle des corrélations :

- une variable dont l'extrémité est proche du cercle des corrélations est bien représentée
- deux variables bien représentées et orthogonales sont décorrélées.
- une variable bien représentée et une variable mal représentées sont décorrélées
- on ne peut rien dire de deux variables mal représentées.

2.4 Relations entre les représentations individus et variables

Elles sont étroitement liées, c'est une des caractéristiques essentielles de l'ACP. et nous avons déjà vu des liens entre les deux. Gardons cependant à l'esprit que même si la formulation des problèmes est similaire dans les deux nuages, l'interprétation est différente dans chacun :

- Dans le nuage des individus, l'origine est l'individu moyen. Le critère d'inertie s'interprète comme une variance et les points projetés doivent être le plus éloignés possible les uns des autres
- Dans le nuage des variables, l'origine n'est pas la variable moyenne. Les points projetés ne sont pas nécessairement dispersés, mais le plus éloignés possible de l'origine.

Nous allons voir dans cette section que des relations de dualité permettent cependant d'établir une interprétation conjointe des deux nuages.

Propriété 2. *Les deux nuages ont la même inertie totale*

En effet,

$$\begin{aligned}\text{Inertie de } N_I &= \sum_i \frac{1}{I} d^2(O, M_i) = \sum_i \frac{1}{I} \sum_k (x_{ik})^2 \\ &= \sum_K ||X_k||^2 = \sum_k \sum_i \frac{1}{I} (x_{ik})^2 \\ &= \text{Inertie de } N_K\end{aligned}$$

et vaut K en ACP normée.

Propriété 3. *Si deux axes suffisent à représenter le nuage des individus, alors deux axes suffisent pour représenter le nuage de variables*

$$\lambda_s = \sum_i \frac{1}{I}(OH_i^s)^2 = \sum_k (T_k^s)^2$$

Relations de dualité

On a déjà vu que la composante principale F_s, vecteur des coordonnées de la projection des individus sur l'axe u_s du nuage des individus dans $\mathbb{R}^K$, définit une nouvelle variable (dans $\mathbb{R}^I$), et on a montré que

$$F_s = \sqrt{\lambda_s} v_s$$

— F_s est colinéaire à l'axe principal de rang s dans l'espace $\mathbb{R}^I$ des variables
— sa variance est λ_s, sa norme $\sqrt{\lambda_s}$.

La direction de $\mathbb{R}^K$ qui représente le mieux le nuage des individus correspond à une combinaison linéaire des variables initiales, variable synthétique qui représente au mieux le nuage des variables dans $\mathbb{R}^I$.

De même, G_s, vecteur des coordonnées des projections des variables sur l'axe $v_s \in R^I$ de norme λ_s est colinéaire au vecteur propre u_s de $C = \frac{1}{I} X'X$

$$G_s = \sqrt{\lambda_s} u_s = \frac{1}{I} X' v_s$$

D'où les relations de *dualité* ou de *transition*

$$F_s[i] = \sqrt{\lambda_s} v_s[i] =< X[i,], u_s >_{R^K} =< x_i, \frac{G_s}{\sqrt{\lambda_s}} >_{R^K} = \frac{1}{\sqrt{\lambda_s}} \sum_k x_{ik} G_s[k]$$

$$G_s[k] = \sqrt{\lambda_s} u_s[k] =< X[,k], v_s >_{R^I} =< X_k, \frac{F_s}{\sqrt{\lambda_s}} >_{R^I} = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{1}{I} x_{ik} F_s[i]$$

Interprétation : Un individu est situé du côté des variables pour lequel il prend de fortes valeurs : $G_s[k]$ est la corrélation au coefficient multiplicatif $\frac{1}{\sqrt{\lambda_s}}$ près entre la variable X_k et les projections F_s des points sur u_s. Si $G_s[k]$ est grand, alors si $X_k[i]$ est grand, $F_s[i]$ est grand.

Remarque $G_s[k] = \text{cor}(X_k, v_s) = \text{cor}(X_k, F_s)$. En effet, F_s est centrée (par projection d'un nuage centré), et de norme $\sqrt{\lambda_s}$.

Effet taille Si toutes les variables sont fortement corrélées, cela indique la présence d'un effet taille dans les données : un individu plus grand, est plus lourd, souvent plus large etc... Ainsi, le premier axe distingue les petits des grands. Les axes suivants permettront de mettre en lueur d'autres caractéristiques.

2.5 Illustration logicielle

Dans cette partie, l'ACP est illustrée sur l'exemple jouet présenté en introduction. Le jeu de données est centré et réduit pour former le dataframe `X`. Il existe plusieurs package qui permettent de traiter l'ACP. Dans un premier temps, nous ferons les calculs "à la main" pour dérouler la démarche. On commence par la diagonalisation de la matrice de corrélation :

```
> C=cor(X)                    #calcul de la matrice de corrélation
      long larg poids
long     1   -1     0
larg    -1    1     0
poids    0    0     1
> E=eigen(C,symmetric=TRUE) # diagonalisation
> lambda=E$values           # valeurs propres
> Us=E$vectors              # vecteurs propres (normés)
```

La matrice de corrélation C possède 3 valeurs propres $\lambda_s = 2, 1, 0$, de somme totale 3, qu'on ordonne en ordre décroissant. Les vecteurs propres associées $Cu_s = \lambda_s u_s$ sont enregistrés dans la matrice de passage `Us`

$u_1(\lambda_1 = 2)$	$u_2(\lambda_2 = 1)$	$u_3(\lambda_3 = 0)$
−0.7071068	0	0.7071068
0.7071068	0	0.7071068
0.0000000	1	0.0000000

Les coordonnées des individus dans la nouvelle base se calculent comme un changement de base ou comme produit scalaire des coordonnées de l'individu avec les axes u_s successifs :

```
CP=X%*%Us
```

Suivant la direction de lecture du tableau CP

- `CP[i,]` : nouvelles coordonnées de l'individu i après changement de base ;
- `CP[,s]` : composante principale d'ordre s

coord de l'individu i sur l'axe	1	2	3
A	2.07	−1.22	0
B	1.24	0.00	0
C	0.41	1.22	0
D	−0.41	1.22	0
E	−1.24	0.00	0
F	−2.07	−1.22	0

Les individus sont exactement représentés dans le plan formé par u_1, u_2. Les composantes sont bien centrées $\sum_i F_s[i] = 0$ mais pas normées $\lambda_s = \sum_i \frac{1}{I} F_s[i]^2$ (rappel, la norme dans $\mathbb{R}^I$ affecte un poids $1/I$ à chaque individu).

```
## on vérifie les propriétés des composantes principales
apply(CP,2,mean)  # centrées
apply(CP,2,function(x){mean(x^2)}) #[1] 1.9968667 0.9922667 0.0000000

## calcul des axes dans l'espace des variables
Vs=CP[,-3];
Vs=Vs%*%diag(1/sqrt(lambda[-3]))
apply(Vs,2,function(x){mean( x^2 )}) # [1] 1 1
```

Pour les axes v_s dans l'espace des variables, il suffit de normer les composantes principales :

$$v_s = \frac{F_s}{\sqrt{\lambda_s}};\ \sum_i \frac{1}{I}(v_s[i])^2 = 1$$

Les coordonnées des variables dans les nouveaux axes sont obtenues

$$G_s[k] = \frac{1}{I} \sum \frac{x_{ik} F_s[i]}{\sqrt{\lambda_s}} = \text{cor}(X_k, F_s)$$

```
# trois façons identiques de calculer les coordonnées de la première variable
mean(CP[,1]*X[,1])/sqrt(lambda[1]) ; cor(CP[,1],X[,1]) ; mean(Vs[,1]*X[,1])

# pour toutes les variables (les deux premiers axes suffisent
CQ=cor(X,CP)[,-3]     # vue corrélation,
# CP n'est pas normée, c'est la corrélation qui s'en charge

round(t(X)%*%Vs/I,2)  # vue changement de base
```

Les résultats de l'ACP ont déjà été présenté en figure 2.1

- point de vue des individus : l'axe 1 oppose les individus A et F qui présentent des valeurs très différentes pour les deux premières variables
- point de vue des variables : l'axe 1 oppose `longueur` et `largeur`, corrélées négativement et sont décorrélées de la variable `poids`
- de part la dualité, la coordonnée de la variable X_k sur v_s représente la corrélation entre ces deux variables. La forte valeur de l'individu A pour F_1 suggère une forte valeur de cet individu pour les variables corrélées positivement à F_1 (largeur) et une valeur faible pour les variables corrélées négativement à F_1 (longueur). Le premier axe oppose les individus trapus (A) aux longilignes (F). Le deuxième axe est corrélé avec le poids (et oppose les lourds aux légers)
- une valeur propre nulle : les données sont parfaitement représentées dans un plan
- les axes sont définis au sens près : les graphiques ne sont pas identiques à ceux de Pagès, mais l'interprétation qui en découle est la même
- Même si la dualité permet d'utiliser l'espace des variables pour poser des interprétations sur l'espace des individus, la représentation simultanée des deux nuages parfois proposée est factice, car représentant des éléments d'espaces différents. Il faut juste interpréter les directions

Le calcul des indicateurs peut se faire de la façon suivante :

```
##### cos2
round( diag( 1/apply(X,1,function(x){sum(x^2)}) ) %*% CP[,-3]^2 , 2)

##### contribution des individus à la variance de chaque axe
round ( 100*  CP[,-3]^2 %*% diag(1/lambda[-3])/I,2  )
```

	Cosinus carrés		Contributions	
	$axe1$	$axe2$	$axe1$	$axe2$
A	0.74	0.26	35.71	25.00
B	1.00	0.00	12.86	0.00
C	0.10	0.90	1.43	25.00
D	0.10	0.90	1.43	25.00
E	1.00	0.00	12.86	0.00
F	0.74	0.26	35.71	25.00

Les calculs précédents sont bien évidemment intégrés dans des fonctions toutes faites des logiciels de statistique et data mining. Par exemple, la fonction `PCA` du package `FactoMineR` s'utilise facilement :

```
library(FactoMineR)
par(mfrow=c(1,2)) # partage l'écran
res.pca=PCA(df)   # effectue l'ACP et trace les représentations
res.pca
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 6 individuals, described by 3 variables
*The results are available in the following objects:

   name               description
1  "$eig"             "eigenvalues"
2  "$var"             "results for the variables"
3  "$var$coord"       "coord. for the variables"
4  "$var$cor"         "correlations variables - dimensions"
5  "$var$cos2"        "cos2 for the variables"
6  "$var$contrib"     "contributions of the variables"
7  "$ind"             "results for the individuals"
8  "$ind$coord"       "coord. for the individuals"
9  "$ind$cos2"        "cos2 for the individuals"
10 "$ind$contrib"     "contributions of the individuals"
11 "$call"            "summary statistics"
12 "$call$centre"     "mean of the variables"
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w"      "weights for the individuals"
15 "$call$col.w"      "weights for the variables"
```

2.6 Individus ou variables supplémentaires

Les éléments contribuant à la création des axes factoriels sont appelés *actifs*. Il peut être intéressant de superposer à la représentation des éléments actifs d'autres individus ou d'autres variables, appelés *supplémentaires* ou *illustratifs*

— variable illustrative quantitative $X_{k'}$, pour $k' > K$: elle se représente dans le cercle des corrélations après le changement de variables correspondant (ou en calculant sa corrélation avec les facteurs.)

$$G_s(X_{k'}) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{I} \sum_{i \text{ actifs}} x_{ik'} F_s[i] = \text{cor}(X_{k'}, F_s)$$

— individu illustratif : chacune de ses coordonnées est centrée-réduite suivant les moyennes et écarts-types calculés sur les individus actifs. Puis on utilise les formules de changement de base

$$F_s[i'] = \frac{1}{\sqrt{\lambda_s}} \sum_{k \text{ actives}} x_{i'k} G_s[k]$$

— variable illustrative qualitative : on ajoute un individu par niveau de facteur, et on calcule ses coordonnées comme moyenne de individus actifs ayant ce niveau de facteur.

2.7 Méthodologie d'étude

1. Acquérir le jeu de données, comprendre la définition des observations et des variables.
2. Définir l'objectif de l'étude. Identifier les individus actifs, les individus illustratifs ; les variables actives, les variables illustratives.
3. Effectuer une analyse uni et bi-variée.
4. Choisir de normer ou non les variables.
5. Lancer les calculs (`PCA` de `FactoMineR` par exemple) en prenant en compte les points 2 et 4.
6. Analyser l'éboulis des valeurs propres. Choisir un nombre d'axe d'étude.
7. Interpréter le cercle des corrélations dans les différents axes principaux. Examiner la qualité de représentation des variables, interpréter leur position. Qualifier les axes.
8. Examiner le nuage des individus dans les différents plans principaux. Repérer les individus contributifs, les individus bien/mal représentés.
9. Interpréter simultanément les individus en fonction de facteurs principaux.

   Rappel : les variables et individus supplémentaires permettent d'aider à l'interprétation, mais ils ne contribuent pas à leur formation.
10. Certains individus extrêmement contributifs peuvent "tirer à eux" la définition d'un axe principal, et masquer des relations qui pourraient être intéressantes. Dans ce cas, il est conseillé de les enlever, puis de vérifier si leur retrait change –ou pas– l'analyse.

Remarques

— des groupes d'individus (ou des groupes de variables) peuvent apparaître naturellement à l'oeil lors de la visualisation des plans principaux. Mais la méthode ne permet pas d'en faire un clustering automatique, celui n'est que manuel à cette étape, voir le chapitre suivant.

— la propriété d'inertie (maximisation de l'inertie projetée sur le plan) et de discrimination (partage de l'espace contenant deux sous-populations identifiées) sont différents. Ainsi, si un tel séparateur plan existe, il faut parfois pousser plus loin dans l'ordre des axes propres pour visualiser cette séparation.

2.8 Décomposition en valeurs singulières

Dans cette section, on fait le lien entre l'ACP et la décomposition en valeurs singulières. On commence par écrire la reconstruction de la matrice de données en fonction des vecteurs propres normés de chacun des nuages.

Soit $\mathbf{X}$ la matrice de données quantitatives de n lignes (observations, individus) et p colonnes (variables). On prendra les notations suivante : x_i est une ligne de $\mathbf{X}$, c'est un vecteur dont la transposée x_i' appartient à $\mathbb{R}^p$: $\mathrm{x}_i' \in \mathbb{R}^p$. On notera $X_i = \mathrm{x}_i'$.

Soit r le rang de la matrice $\mathbf{X}$. On range les r valeurs propres non nulles de la matrice de corrélation $(\mathbf{X}'\mathbf{X})/n$ par ordre décroissant, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$: ce sont les inerties projetées sur les différents axes principaux (vecteurs propres associées à ces valeurs propres) $u_1, \ldots, u_r$. La formule de reconstruction s'écrit pour une observation

$$\mathrm{x}_i = \sum_j < X_i, u_j > u_j' = \sum_j \mathrm{x}_i u_j u_j'$$

et pour l'ensemble du tableau :

$$\mathbf{X} = \sum_{j=1}^{r} F_j u_j' = \sum_{j=1}^{r} \sqrt{\lambda_j} v_j u_j'$$

où $F_j = \mathbf{X}u_j$ est la j-ème composante principale. En utilisant les formules de dualité entre le nuage des individus et celui des variables, on a $F_j = \sqrt{\lambda_j} v_j$, où v_j sont les vecteurs propres normés (suivant la norme du nuage des variables) associés aux valeurs propres λ_j correspondantes :

$$u_i u_j = \delta_{ij}; \;\; v_j' v_j / n = \delta_{ij}.$$

Soit U la matrice dont les colonnes sont les vecteurs u_j et V celle dont les colonnes sont les vecteurs v_j. On peut ainsi écrire

$$\mathbf{X} = V \begin{pmatrix} \sqrt{\lambda_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{\lambda_r} \end{pmatrix} U' = \tilde{V} \begin{pmatrix} \sqrt{\lambda_1/n} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{\lambda_r/n} \end{pmatrix} U'$$

où $\tilde{V}'\tilde{V} = Id_r$.

Cette décomposition est appelée décomposition en valeurs singulières de la matrice $\mathbf{X}$.

2.8.1 Construction de la SVD

Dans l'introduction, nous avons défini les matrices U et V en étant homogènes avec les notations utilisées pour l'ACP. Ce ne sont pas celles classiquement utilisées pour définir la SVD, nouvelles notations que nous introduisons maintenant. Prendre $\mathbf{X} = A'$ pour faire la correspondance.

Définition 4. *Toute matrice $A \in \mathbb{R}^{p \times n}$ de rang $r \leq \min(n, p)$ peut être décomposée en un produit de trois matrices*

$$A = U \Sigma V'$$

où U est la matrice d'un changement de base orthonormée de $\mathbb{R}^p$, V celle d'un changement de base orthonormée de $\mathbb{R}^n$, Σ une matrice de dimension $p \times n$ telles que :

— Σ *est la matrice diagonale par bloc, dont tous les coefficients sont nuls, sauf ceux du premier bloc* $\Sigma_{1:r;1:r} = D$*, matrice diagonale de coefficients diagonaux* $d_j = \sqrt{\lambda_j}$.
Ces d_j *sont les racines carrées des valeurs propres non nulles* $\lambda_1 \geq \ldots \geq \lambda_r > 0$ *de la matrice* AA'*, classées par ordre de valeur décroissante.*
Ce sont aussi les valeurs propres non nulles de la matrices $A'A$ *;*
— *les r premières colonnes de U sont les vecteurs propres* u_j *associés aux valeurs propres* λ_j *non nulles de* AA'
— *les r premières colonnes de V sont les vecteurs* $v_j = A'u_j/\sqrt{\lambda_j}$*, vecteurs propres de* $A'A$ *associés aux valeurs propres non nulles* λ_j.

Note On désigne parfois la SVD sous une forme plus compacte :

$$A = U_{1:r} D V'_{1:r} = \sum_{j=1}^{r} d_j u_j v'_j,$$

appelée *SVD réduite.*

Définition 5. *Les valeurs* d_j *sont appelées valeurs singulières de la matrice A, les* u_j *sont les vecteurs singuliers à gauche ; les* v_j *sont les vecteurs singuliers à droite.*

Propriété : on a aussi $Av_j/\sqrt{\lambda_j} = u_j$ pour toute valeur propre non nulle λ_j.

Preuve. AA' est une matrice de $\mathbb{R}^{p\times p}$ symétrique. Par le théorème de représentation spectrale, elle est diagonalisable, toutes ses valeurs propres sont réelles, son rang $r \leq \min(n,p)$. Elle est semi-définie positive : $\forall y \in \mathbb{R}^p, y'AA'y = ||Ay||^2 \geq 0$.

Il existe une base orthonormale $(u_1,\ldots,u_r,\ldots,u_p)$ de $\mathbb{R}^p$ vecteurs propres associés aux valeurs propres classées par valeurs décroissantes $\lambda_1 \geq \ldots\ldots \geq \lambda_p$. Soit $\Delta = diag(\lambda_j)$. Dans cette base

$$AA' = U\Delta U' = U_{1:r} D^2 U'_{1:r} = \sum_{j=1}^{r} \lambda_j u_j u'_j \ .$$

On écrit maintenant la décomposition de chaque colonne $A_j \in \mathbb{R}^p$, $j = 1,\ldots,n$ de la matrice A dans cette base :

$$A_i = < A_i, u_1 > u_1 + \ldots + < A_i, u_r > u_r + \ldots < A_i, u_p > u_p$$

soit

$$A = U \begin{pmatrix} u'_1A_1 & \ldots & u'_1A_i & \ldots & u'_1A_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u'_pA_1 & \ldots & u'_pA_i & \ldots & u'_pA_n \end{pmatrix} = U \begin{pmatrix} u'_1A \\ \vdots \\ u'_pA \end{pmatrix} = UU'A$$

Soit $\lambda_j \neq 0$. Soit le vecteur $v_j = A'u_j/\sqrt{\lambda_j}$. Il est normé :

$$||v_j||^2 = u'_jAA'u_j/\lambda_j = u_j\lambda_j/\lambda_j = 1$$

et il est vecteur propre de la matrice $A'A$ associé à la valeur propre λ_j :

$$A'Av_j = A'AA'u_j/\sqrt{\lambda_j} = A'(\lambda_j u_j)/\sqrt{\lambda_j} = \lambda_j v_j.$$

Les r premières lignes de matrice $U'A$ correspondent aux vecteurs lignes $\sqrt{\lambda_j}v'_j$.

Si $\lambda_j = 0$, les $(u_j)_{j=r+1,\dots,p}$ forment une base du noyau de AA' et on a $AA'u_j = 0$, d'où $u_j'AA'u_j = 0 = ||A'u_j||^2$. Ainsi, les $p-r$ dernières lignes de $U'A$ sont nulles, ce qui amène à la représentation réduite de la SVD :

$$A = U_{1:r}DV_{1:r}' .$$

On retrouve ce résultat de façon alternative en remarquant que $Im(AA')$ est engendré par la famille libre $\{u_1, \dots, u_r\}$. Comme $Im(AA') = ImA$, cette famille est également une base de $Im(A)$ et $U_{1:r}U_{1:r}'$, projection de $\mathbb{R}^p$ sur $Im(AA')$, est également la projection sur $Im(A)$. Ainsi $A = U_{1:r}U_{1:r}'A = U_{1:r}DV_{1:r}'$.

A et A' étant de même rang, on peut choisir une base de $p-r$ vecteurs du noyau de A' pour compléter $\{v_1, \dots, v_r\}$ en V. ◊

2.8.2 Lien avec les normes matricielles

Les normes classiques sur les matrices peuvent s'écrire en fonction des valeurs singulières $d_j = \sqrt{\lambda_j}$ de la matrice.

Norme de Frobenius associée au produit scalaire de Frobenius $< A, B >_F = \sum_{i,j} A_{ij}B_{ij}$:

$$< A, A >_F^2 = \sum_{i=1}^{p}\sum_{j=1}^{n} A_{ij}^2 = trace(A'A) = \sum_j d_j^2$$

puisque les valeurs singulières sont les racines carrés des valeurs propres de $A'A$.

Norme opérateur définie par :

$$||A||_2 := \sup_{x\neq 0} = \frac{||Ax||_2}{||x||_2} = \sup_{||x||\leq 1} ||Ax||_2 = d_1.$$

En effet, $Ax = \sum_j d_j u_j v_j' x$, d'où $||A||_2^2 = \sum_j d_j^2 < v_j, x >^2 \leq d_1^2||x||^2 \leq d_1^2$, avec égalité si $x = v_1$.

2.8.3 Comparaison SVD / décomposition en valeurs propres

La décomposition en valeurs singulières $A = U\Sigma V'$ et la décomposition en valeurs propres $A = P\Delta P^{-1}$ définissent deux compositions de trois applications linéaires : un changement de base dans l'espace de départ, une mise à l'échelle (inflation, déflation) de chaque nouveau vecteur, puis un changement de base dans l'espace d'arrivée.

La décomposition en valeurs singulières généralise la décomposition en valeurs propres dans deux directions :

- d'une part à des matrices non carrées $n \neq p$
- d'autre part à des matrices carrées pour lesquelles il n'est pas possible de trouver une base de vecteurs propres de $\mathbb{R}^{p=n}$.

En particulier,

- la SVD existe toujours contrairement à la décomposition en valeurs propres
- U et V sont des matrices orthogonales (i.e. des rotations), contrairement à P qui ne l'est pas nécessairement.

— U et V ne sont généralement pas inverses l'une de l'autre, définissant des rotations dans des espaces différents
— Les valeurs singulières sont réelles et positives, alors que les valeurs propres peuvent être complexes ou négatives

Pour des matrices symétriques non négatives, les deux décompositions sont similaires.

2.8.4 Application à la réduction de dimension

La SVD réduite permet de reconstruire exactement la matrice A avec un minimum d'information : les r valeurs singulières et les r vecteurs singuliers à gauche, dont on déduit les r vecteurs singuliers à droite (ou le contraire). Il est possible de définir une SVD *tronquée*, de rang $k < r$, en n'utilisant que k valeurs singulières et vecteurs singuliers associés. La reconstruction donne alors une approximation de la matrice A par la matrice $A^{(k)}$ de rang k

$$A^{(k)} = \sum_{j=1}^{k} d_j^2 u_j u_j' = \sum_{j=1}^{k} d_j u_j v_j'.$$

On aurait pu choisir une autre approximation de rang k, par exemple $\sum_{j=r-k+1}^{r} d_j^2 u_j u_j'$, mais elle est moins bonne que la précédente, comme le montre dans le théorème suivant

Théorème 3 (Eckart-Young)**.** *Soit A une matrice de $\mathbb{R}^{p\times n}$ de rang r et B une matrice de $\mathbb{R}^{p\times n}$ de rang $k \leq r$. Alors*

$$A^{(k)} = \sum_{j=1}^{k} d_j^2 u_j u_j' = \sum_{j=1}^{k} d_j u_j v_j' = \arg \min_{B;rang(B)=k} ||A - B||_2$$

et $||A - A^{(k)}||_2 = d_{k+1}$

Le théorème précise que de toutes les projections de la matrice A sur l'espace des matrices de rang $k \leq p$, la SVD minimise l'erreur (calculée sous la norme spectrale) entre A et n'importe quelle approximation de A de rang inférieur ou égal à k. Elle est dans ce sens optimale par rapport à la mesure d'erreur de la norme opérateur.

Preuve. En effet, $A - A^{(k)} = \sum_{j=k+1}^{r} d_j u_j v_j'$, d'où $||A - A^{(k)}||_2 = d_{k+1}$. Supposons qu'il existe une autre matrice B de rang inférieur ou égal à k dont l'approximation soit strictement meilleure : $||A - B||_2 < ||A - A^{(k)}||_2$.

Par hypothèse, le noyau de B est de dimension supérieure ou égale à $n - k$. Pour tout $x \in Ker\ B$, on a $Bx = 0$ et

$$||Ax||_2 = ||(A - B)x||_2 \leq ||A - B||_2\ ||x||_2 < ||A - A^{(k)}||_2\ ||x||_2 = d_{k+1}||x||_2.$$

Par ailleurs, il existe des $x \in \mathbb{R}^n$ tels que leurs images Ax vérifient $||Ax||_2 > d_{k+1}||x||_2$. Ils appartiennent à l'ensemble E des antécédents des vecteurs non nuls de l'espace vectoriel engendré par les vecteurs singuliers à droite $v_1, \ldots, v_{k+1}$, qui est de dimension $k+1$. Comme $(Ker\ B) \cap E = \{0\}$, on a $dim(Ker\ B) + dim(E) \geq n - k + k + 1 = n + 1$, ce qui est impossible. Donc $B = A^{(k)}$. $\Diamond$

Chapitre 3

Analyse Factorielle des Correspondances

L'AFC est une technique qui porte sur les tableaux de contingence. Un *tableau de contingence* croise les modalités de deux variables qualitatives ayant respectivement I et J niveaux, observées sur des individus, et enregistre le nombre d'individus ayant répondu simultanément la modalité i sur la première variable et j sur la seconde.

Le tableau de travail est donc un résumé du nombre de couples de réponses individuelles sur les deux variables ($n \times 2$ données initiales) et se présente sous forme d'une table $X = (x_{ij})$ de dimension $I \times J$. On peut calculer les *marges des lignes* $x_{\bullet j}$, les *marges des colonnes* $x_{i\bullet}$, la somme de toutes les cellules du tableau étant égale au nombre d'individus n :

$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij}; \quad x_{\bullet j} = \sum_{i=1}^{I} x_{ij}; \quad x_{\bullet\bullet} = \sum_{i=1}^{I} x_{i\bullet} = \sum_{j=1}^{J} x_{\bullet j} = n$$

				total
	x_{11}	$\dots$	x_{1J}	$x_{1\bullet}$
	$\vdots$	x_{ij}	$\vdots$	$x_{i\bullet}$
	x_{I1}	$\dots$	x_{IJ}	$x_{I\bullet}$
total	$x_{\bullet 1}$	$x_{\bullet j}$	$x_{\bullet J}$	n

Exemples :

- analyse du croisement couleur des yeux, couleur des cheveux
- analyse de résultats sportifs : individus=médailles, pour lesquelles on observe le couple (pays, discipline)
- association département / tranche d'âge de la population
- association type de média / type d'information
- association alimentation/CSP
- etc...

Le tableau manipulé est numérique, on pourrait utiliser une technique d'ACP, mais celle-ci ne prend pas en compte la symétrie du problème.

De plus, dans le cadre de l'étude du croisement de deux réponses, il est important de regarder la valeur d'une cellule par rapport à sa marge en ligne ou en colonne : si x_{ij} est élevé, cela ne

veut pas forcément dire que la modalité i attire particulièrement la modalité j : c'est peut-être que chacune des modalités i et j est globalement (sur l'ensemble des individus) plus importante.

La question est donc de savoir si, dans la population ayant répondu i à la première question, il y a proportionnellement plus (ou moins) de personnes répondant j à la deuxième variable que la taux moyen de réponse observée à la modalité j dans l'ensemble de la population :

$$\frac{x_{ij}}{x_{i\bullet}} > \frac{x_{\bullet j}}{n} ?$$

et réciproquement, s'il y a plus de personnes de la modalité j qui répondent i à la première variable que ne le ferait un individu moyen (sur l'ensemble des modalités j.)

$$\frac{x_{ij}}{x_{\bullet j}} > \frac{x_{i\bullet}}{n} ?$$

L'exemple suivant est proposé par Husson et Pagès (2009) : Des personnes ont dû répondre à deux questions à trois niveaux

— Q_1 : Quelle est pour vous la famille idéale : les deux conjoints travaillent également (H=F), le mari a un métier plus absorbant (H>F) ; seul le mari travaille (H)

— Q_2 : Quelle est l'activité convenant le mieux à une mère de famille quand les enfants sont en âge scolaire : rester au foyer (foyer), travailler à mi-temps (mi), travailler à plein-temps (plein)

$Q_1 \backslash Q_2$	$foyer$	mi	$plein$	$total\ Q_1$
$H = F$	13	142	106	261
$H > F$	30	408	117	555
H	241	**573**	94	908
$Total\ Q_2$	284	1123	317	1724

On pourrait croire que la modalité mi attire la modalité H. Mais n'est-ce pas dû au fait que ces deux réponses sont séparément majoritaires ? A suivre... Les tableaux de contingence ne peuvent donc être étudiés qu'en gardant à l'esprit les marges qui leur correspondent. L'AFC permet

— de comparer les profils lignes entre eux

— de comparer les profils colonnes entre eux

— de caractériser des profils de lignes qui se ressemblent, par des modalités de colonnes.

Il s'agit donc d'interpréter des liaisons entre les modalités des lignes et les modalités colonnes, c'est à dire de mettre en correspondance, d'où le nom *analyse des correspondances*.

3.1 Indépendance

Plutôt que d'étudier la liaison entre les deux variables (qui est en général acquise), il s'agit de définir comment se positionnent les données par rapport à la situation d'indépendance :

$$P(Q_1 = i, Q_2 = j) = P(Q_1 = i)P(Q_2 = j).$$

Pour étudier l'indépendance/dépendance, on transforme le tableau en tableau de fréquences $f_{ij} = \frac{x_{ij}}{n}$. Ce nouveau tableau définit une mesure de probabilité sur l'ensemble produit $I \times J$. Ses marges (ou *probabilités marginales*) sont définies par :

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij}; \quad f_{\bullet j} = \sum_{i=1}^{I} f_{ij}; \quad f_{\bullet\bullet} = \sum_{i=1}^{I} f_{i\bullet} = \sum_{j=1}^{J} f_{\bullet j} = 1$$

D'où le tableau des fréquences relatives (en %) :

x_{ij}/n	$foyer$	mi	$plein$	marge des lignes
$H = F$	0.75	8.24	6.15	15.14
$H > F$	1.74	23.67	6.79	32.19
H	13.98	33.24	5.45	52.67
marge des colonnes	16.47	65.14	18.39	100

Il y a indépendance quand

$$f_{ij} = f_{i\bullet} f_{\bullet j}$$

— les lignes sont proportionnelles à la marge des colonnes avec le facteur $f_{i\bullet}$
— les colonnes sont proportionnelles à la marge des lignes avec le facteur $f_{\bullet j}$

Étudier la liaison revient à comparer le tableau des effectifs observés $x_{ij} = nf_{ij}$ avec les effectifs théoriques $nf_{i\bullet}f_{\bullet j}$ s'il y avait indépendance. Ou autrement dit : la connaissance du niveau de réponse à la question Q_1 des personnes ayant répondu j à Q_2 donne-t-il plus d'information que le profil moyen des réponses ?

Exemple (suite) : on calcule le tableau des effectifs théoriques et des fréquences théoriques (en %)

$x_{i\bullet}x_{\bullet j}/n$	$foyer$	mi	$plein$	
$H = F$	43	170	48	261
$H > F$	91.4	361.5	102.1	555
H	149.6	591.5	167	908
	284	1123	317	1724

$f_{i\bullet}f_{\bullet j}$	$foyer$	mi	$plein$	
$H = F$	2.5	9.8	2.8	15.1
$H > F$	5.3	21.0	5.9	32.2
H	8.7	34.3	9.7	52.7
	16.5	65.1	18.4	100

En fait, les deux modalités mi et H ont tendance à se repousser légèrement : on observe 33.24% des réponses en (H,mi), c'est inférieur à la valeur théorique de 34.3% s'il y avait indépendance ! La forte valeur de cette cellule dans le tableau des observations est donc bien due au fait que les deux modalités prises séparément sont fréquentes.

Le test du $Khi - deux$ permet l'indépendance de deux variables qualitatives observées sur les mêmes individus :

$$\begin{aligned} Khi^2 &= \sum_{ij} \frac{(\text{effectif observé - effectif théorique})^2}{\text{effectif théorique}} \\ &= \sum_{ij} \frac{(x_{ij} - x_{i\bullet}x_{\bullet j}/n)^2}{x_{i\bullet}x_{\bullet j}/n} = n \sum_{ij} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} \end{aligned}$$

Sous l'hypothèse d'indépendance, la statistique Khi^2 suit une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté. Ici, la statistique observée vaut $Khi^2 = 233 > q_{\chi^2(4)}(0.95) = 9.48$, pour une loi du χ^2 à 4 degrés de liberté : on rejette donc l'hypothèse d'indépendance. Plus que de savoir si on est en situation d'indépendance, on cherchera à définir les ressemblances (dissimilarités) entre lignes (resp. colonnes), et on s'attachera à regarder

— si deux lignes s'attachent trop (ou trop peu) aux mêmes colonnes (par rapport à la situation d'indépendance, ie, par rapport au profil moyen ligne = marge des colonnes) ;
— si deux colonnes s'attachent trop (ou trop peu) aux mêmes lignes (par rapport à la situation d'indépendance, ie, par rapport au profil moyen colonne= marge des lignes).

On cherche donc des lignes (réciproquement des colonnes) qui s'écartent le plus possible de la situation d'indépendance, c'est à dire dont la répartition s'écarte le plus du profil moyen ; et on

regroupe les lignes (resp. colonnes) qui se ressemblent le plus. On caractérise alors ces groupes de lignes (resp.colonnes) par les modalités de colonnes (resp. de lignes) auxquelles le groupe s'associe trop ou trop peu.

3.2 Nuage des profils lignes

Comme dans l'ACP, nous sommes en présence de deux nuages de points, celui des lignes et celui des colonnes, et nous avons vu qu'il est intéressant de normaliser les lignes par leur somme et les colonnes par leur somme. Nous aurons donc deux nuages à étudier, pour lesquels, comme en ACP, il faut définir la normalisation, le poids de chaque individu et la distance entre les individus. Mais la définition de ces opérations est différente de celle opérée en ACP.

Différences par rapport à l'ACP

- Transformation en profil. Les *profils lignes* sont définis par les vecteurs lignes $f_{ij}/f_{i\bullet}$, et définissent I points dans un espace de dimension J : c'est la loi conditionnelle définie par i sur l'ensemble des colonnes. Le profil-ligne moyen est la marge des colonnes G_ℓ de coordonnées $f_{\bullet j}/1 = f_{\bullet j}$.
- Affectation d'un poids $f_{i\bullet}$ à chaque profil ligne. Une modalité est d'autant plus influente à profil égal qu'elle est fréquente. Le profil moyen G_ℓ est ainsi le *barycentre* de N_I affecté de ces poids :
$$\sum_i f_{i\bullet} \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$$
- Définition de la distance. Chaque colonne est affectée d'un coefficient $1/f_{\bullet j}$. La distance entre deux profils est alors la distance du χ^2 dont le carré se calcule de la façon suivante :
$$d^2_{\chi^2}(i,i') = \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 ; \quad < u_1, u_2 >_{R^J} = \sum_j \frac{1}{f_{\bullet j}} u_{1j} u_{2j}$$

La distance du χ^2 jouit de la propriété d'*équivalence distributionnelle* : si deux colonnes proportionnelles sont cumulées en une seule, alors les distances entre les profils lignes ne sont pas modifiées. La proportionnalité stricte est peu souvent observée, mais on utilise cette propriété de façon pragmatique : le regroupement de modalités dont les profils sont presque proportionnels ne modifie pas sensiblement les résultats d'une AFC.

Remarque En ACP normée, les lignes sont affectées d'un poids $1/n$ identique tandis que celui affecté à chaque colonne pour le calcul de la distance est constant égal 1 puisque les variables sont normées.

Le tableau des profils lignes de l'exemple est donc :

	foyer	*mi*	*plein*		*poids*
$H = F$	0.050	0.544	0.406	1.000	0.151
$H > F$	0.054	0.735	0.211	1.000	0.322
H	0.265	0.631	0.104	1.000	0.527
G_ℓ	0.165	0.651	0.184	1.000	1.000

On a vu qu'un des buts de l'AFC est de déterminer les individus éloignés du profil moyen. L'inertie de l'individu i, définie comme produit du poids par le carré de la distance à G_ℓ, se

calcule donc par

$$f_{i\bullet} d^2_{\chi^2}(i, G_\ell) = f_{i\bullet} \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 = \sum_j \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}$$

On retrouve, au terme n près la contribution de la ligne au Khi^2, d'où le nom de distance du χ^2. Examiner la dispersion du nuage N_I autour de son centre de gravité G_ℓ revient bien à examiner l'écart entre les données et le modèle d'indépendance.

Remarque : l'inertie totale

$$\mathcal{I} = \sum_i f_{i\bullet} d^2_{\chi^2}(i, G_\ell) = Khi^2/n$$

mesure l'*intensité de la liaison*, tandis que Khi^2 mesure la *détection statistique* de cette intensité : s'il y a peu de points, on ne détecte pas forcément une forte liaison ; s'il y a beaucoup de points, on peut détecter une liaison, même faible.

C'est une autre grande différence avec l'ACP : dans cette dernière, l'inertie est égale au nombre de variables : elle dépend donc du format des variables (nombre...), mais pas des données elles-mêmes (valeurs).

3.3 Nuage des profils colonnes

On échange i en j et tout est pareil ! La distance est définie par

$$d^2_{\chi^2}(j, j') = \sum_i \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2$$

et l'inertie vaut

$$f_{j\bullet} d^2_{\chi^2}(j, G_c) = \sum_i \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}.$$

Examiner l'écart entre les données et le modèle d'indépendance revient à examiner la dispersion du nuage N_J autour de son centre de gravité G_c.

	$foyer$	mi	$plein$	G_c
$H = F$	0.046	0.126	0.334	0.151
$H > F$	0.106	0.363	0.369	0.322
H	0.849	0.510	0.297	0.527
	1.000	1.000	1.000	1.000
$poids$	0.165	0.651	0.184	1.000

3.4 Ajustement des nuages

On procède comme pour l'ajustement du nuage des individus dans l'ACP. Ayant placé l'origine du repère au centre de gravité des profils lignes G_ℓ [1], on calcule une succession d'axes orthogonaux qui maximise l'inertie projetée. Soit H_i^s la projection du profil i sur l'axe u_s, on cherche u_s qui rend maximum

$$\sum_i f_{i\bullet} (OH_i^s)^2$$

tout en étant orthogonal aux axes précédents.

1. Ceci n'est en fait pas nécessaire, voir la propriété 4

— L'inertie projetée λ_s mesure la part de la liaison (au sens de l'inertie) exprimée par cet axe.
— la distance entre un profil ligne et G_ℓ exprime la contribution de l'écart à l'indépendance de ce profil
— la proximité entre deux profils lignes exprime une même façon de s'écarter du profil moyen
— l'éloignement de part et d'autre de G_ℓ exprime deux façons opposées de s'écarter du profil moyen : les modalités avec lesquelles i s'associe le plus sont celles avec lesquelles i' s'associe le moins.

Mise en oeuvre des calculs

Calculons OH_i, où H_i est la projection d'un profil ligne sur l'axe u :

$$OH_i =< OM_i, u >= \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} u_j \right) = \sum_j \frac{f_{ij}}{f_{i\bullet}} \frac{u_j}{f_{\bullet j}}$$

Soit $D_I = diag(f_{i\bullet})$, $D_J = diag(f_{\bullet j})$, $F = (f_{ij})$. Le profil ligne s'exprime par $\frac{f_{ij}}{f_{i\bullet}} = (D_I^{-1}F)[i,]$. On veut maximiser l'inertie totale suivante :

$$\sum_i f_{i\bullet} \left[\sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} u_j \right) \right]^2 = \left(u' D_J^{-1} F' D_I^{-1} \right) \quad D_I \quad \left(D_I^{-1} F D_J^{-1} u \right) = u' D_J^{-1} F' D_I^{-1} F D_J^{-1} u$$

donc à maximiser le Lagrangien :

$$L = u' D_J^{-1} S u - \lambda (u' D_J^{-1} u - 1).$$

Soit u_s **le vecteur propre normé** $u_s' D_J^{-1} u_s = 1$ **d'ordre** s **associé à la valeur propre** λ_s **de la matrice** $S = F' D_I^{-1} F D_J^{-1}$, on a :

$$F' D_I^{-1} F D_J^{-1} u_s = \lambda_s u_s$$

et la s-ième composante principale est :

$$F_s =< D_I^{-1} F[i,], u_s >= D_I^{-1} F D_J^{-1} u_s \tag{3.1}$$

Sa moyenne pondéré est nulle $\sum_i f_{i\bullet} F_s[i] = 0$, et sa variance (inertie) vaut

$$\left(u_s' D_J^{-1} F' D_I^{-1} \right) D_I \left(D_I^{-1} F D_J^{-1} u_s \right) = u_s' D_J^{-1} \lambda_s u_s = \lambda_s \geq 0$$

D'un point de vue technique, on n'est pas obligé de centrer le nuage. La somme des coordonnées d'un profil étant égale à 1, le nuage appartient à un sous espace de dimension $I - 1$. La première valeur propre est alors associée à l'axe factoriel OG_ℓ, orthogonal au nuage N_I et sa valeur propre vaut 1 : la projection de tous les points sur cette direction se trouve en G_ℓ et $d_{\chi^2}(0, G_\ell) = 1$.

Dans le nuage des profils colonne, on trouve que v_s est le vecteur propre d'ordre s de la matrice $T = F D_J^{-1} F' D_I^{-1}$

$$F D_J^{-1} F' D_I^{-1} v_s = \lambda_s v_s$$

et

$$G_s = D_J^{-1} F' D_I^{-1} v_s$$

centré de variance λ_s également.

— Comme dans l'ACP, les inerties projetées sont les valeurs propres d'une matrice dont les vecteurs propres correspondants sont les axes principaux. Les valeurs propres sont comprises entre 0 et 1 de part la forme spécifique de la matrice, voir la propriété 6.
— Le nuage a été préalablement centré, c'est à dire que le nuage des profils lignes $f_{ij}/f_{i\bullet}$ a été remplacé par

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

qui peut être vu comme le résidu par rapport à la situation d'indépendance.
Il y a une valeur propre nulle correspondant à la direction orthogonale à l'hyperplan du nuage.
— Si on n'a pas centré nuage, la plus grande valeur propre vaut 1, et on l'écarte, voir la propriété 4 en fin de chapitre.
— le nombre maximum d'axes d'inertie utilisables : $min(I-1, J-1)$

Dans le cas de l'exemple, la valeurs propres valent 1, 0.117, 0.018 (nuage non centré), d'inertie totale 0.135. On retrouve cette inertie par $Khi^2/n = 233.43/1724$. On peut représenter la projection du nuage dans les premiers axes principaux.

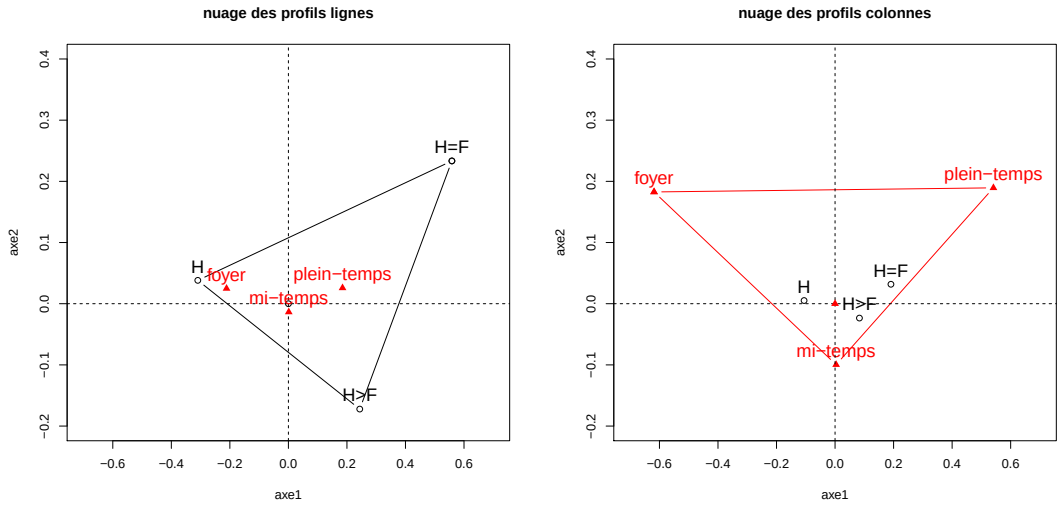


FIGURE 3.1 – Représentation barycentrique

Remarque : dans le nuage des profils lignes (par exemple), le profil ligne moyen $G_l = (f_{\bullet 1}, \ldots, f_{\bullet J})'$ est au barycentre des profils ligne ; cette propriété se conserve dans la projection : c'est l'origine des plans principaux. De plus, le barycentre des projections des profils lignes affectés des coefficients du profil moyen colonne est également au centre du repère :

$$G_c' F_s = G_c' D_I^{-1} F D_J^{-1} u_s = (1 \ldots 1)_I F D_J^{-1} u_s = (1 \ldots 1)_J u_s = \sum_i u_s[j] = 0$$

Le principe est strictement identique pour le nuage N_J en adaptant les poids et distance. Dans le cas de l'exemple, on peut interpréter de la façon suivante :
— nuage des profils colonnes : le premier axe (attitude des femmes vis à vis du travail féminin) oppose *travailler à plein temps* et *rester au foyer*. Cette opposition se traduit en terme de

différences de profils de réponse sur les questions des lignes, et marque l'écart du tableau à l'indépendance (ou explique la liaison).
— nuage des profils lignes : on trouve le premier axe comme un indicateur en faveur du travail féminin
— *Travailler à mi-temps* est proche du profil moyen : ce n'est pas très informatif, parce que le choix de cet item ne donne pas d'idée sur ce qu'aurait pu être répondu à la question 1 : distribution conditionnelle à *travailler à mi-temps* proche du profil moyen G_c

Représentation barycentrique

On peut aussi ajouter les modalités de l'autre variable en points supplémentaires. Dans l'espace du nuage des profils lignes, on ajoute les barycentres des profils lignes projetés, affectés des coefficients des profils colonnes $f_{ij}/f_{\bullet j}, i = 1, \ldots, I$ (soit J points supplémentaires) : un point barycentre indique comment une modalité de l'autre variable (représentée par le barycentre) s'associe globalement avec les modalités de la variable représentée
— dans le nuage des points lignes : la réponse *foyer* s'associe presque exclusivement avec la réponse *Homme seul travaille* : ie, dans la colonne *foyer*, le poids de la réponse *Homme seul travaille* est très fort (85%)
— dans le nuage des points colonnes : la réponse *Homme seul travaille* s'associe fortement avec la réponse *mi-temps* (63%) ; la réponse *deux conjoints travaillent* s'associe à part égale aux réponses *mi-temps* (54%) et *plein-temps* (40%)

3.5 Dualité entre les facteurs dans les deux nuages

Chacune des analyses (nuage des profils lignes, nuage des profils colonnes) amène à une représentation optimale du point de vue de l'inertie projetée. Mais ce sont des représentations du même tableau (prises d'un point de vue différent), elles sont donc naturellement liées par des relations de dualité
— On note que $FD_J^{-1}u_s$ est vecteur propre de $T = FD_J^{-1}F'D_I^{-1}$ et sa norme vaut $\sqrt{\lambda_s}$ d'où

$$v_s = \frac{1}{\sqrt{\lambda_s}} FD_J^{-1}u_s$$

— De même, $F'D_I^{-1}v_s$ est vecteur propre de $S = F'D_I^{-1}FD_J^{-1}$ et sa norme vaut $\sqrt{\lambda_s}$ d'où

$$u_s = \frac{1}{\sqrt{\lambda_s}} F'D_I^{-1}v_s \tag{3.2}$$

La dualité prend forme de façon suivante :

1. L'inertie des deux nuages est identique, égale à Khi^2/n et mesure l'écart à l'indépendance, qui est le même si on prend le point de vue des profils lignes, ou des profils colonnes
2. L'inertie projetée λ_s sur l'axe d'ordre s est la même dans les deux nuages. C'est une propriété caractéristique des axes factoriels

$$\lambda_s = \sum_i f_{i\bullet}(OH_i^s)^2 = \sum_j f_{\bullet j}(OH_j^s)^2$$

3. Il y a des relations de dualité entre les coordonnées des profils lignes et des profils colonnes sur les axes de même rang

$$F_s[i] = \frac{\sqrt{\lambda_s}}{f_{i\bullet}} v_s[i] = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i\bullet}} G_s[j]$$

$$G_s[j] = \frac{\sqrt{\lambda_s}}{f_{\bullet j}} u_s[j] = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{\bullet j}} F_s[i]$$

Dans le nuage des profils ligne (resp. colonne), la projection sur l'axe s d'un profil ligne (resp.colonne) est au barycentre des colonnes (resp. lignes), affectées des coefficients $\frac{f_{ij}}{f_{i\bullet}}$(resp. $\frac{f_{ij}}{f_{\bullet j}}$).

Interprétation simultanée quasi-barycentrique

Contrairement à l'ACP, la représentation simultanée est possible car les profils ligne et les profils colonnes sont intrinsèquement de même nature, même inertie sur les axes, et relation duale entre les coordonnées.

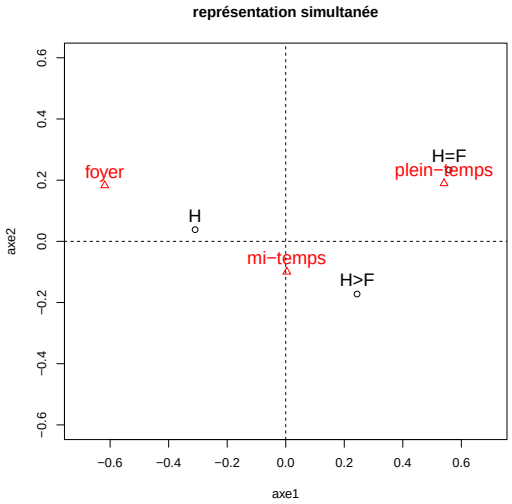


FIGURE 3.2 – Représentation simultanée quasi-barycentrique : $\lambda_s = (0.11, 0.018)$, soit $1/\sqrt{\lambda_s} = (3, 7.45)$

— dans le nuage des lignes : un profil ligne i est au barycentre des profils colonnes affectés des coefficients $\frac{f_{ij}}{f_{i\bullet}}$, et au coefficient $1/\sqrt{\lambda_s}$ sur chaque axe.
— dans le nuage des colonnes : un profil colonne j est au barycentre des profils lignes affectés des coefficients $\frac{f_{ij}}{f_{\bullet j}}$, et au coefficient $1/\sqrt{\lambda_s}$ sur chaque axe.

Cette propriété est appelée *quasi-barycentrique* (barycentrique au coefficient $1/\sqrt{\lambda_s}$ près). Ce coefficient est un coefficient de dilatation car $0 \leq \lambda_s \leq 1$ d'après la forme de la matrice, ie : le quasi-barycentre est plus éloigné de l'origine que le barycentre correspondant, et d'autant plus que la valeur propre est petite (donc que l'intensité de la liaison est faible). Ici $\lambda_s = (0.11, 0.018)$, soit $1/\sqrt{\lambda_s} = (3, 7.45)$.

La représentation simultanée quasi-barycentrique est pratique parce que

— elle résume toute l'information sur un seul graphique
— les deux axes y jouent le même rôle
— elle rapproche les modalités qui se correspondent (contrairement à la représentation barycentrique qui contracte les modalités de la deuxième variable autour du centre d'inertie)

mais il faut faire attention à la dilatation quand on fait de l'interprétation...

3.6 Interprétations

En résumé, on mesure
— l'*intensité de la liaison* par la valeur propre (liaison d'autant plus forte que la valeur propre est proche de 1) : en effet, la valeur propre sur un axe la part de l'intensité (inertie) de la liaison prise par cette axe. On peut vouloir "normaliser" cette valeur par l'inertie totale du nuage (qui est la même dans N_I et dans N_J) : c'est le *pourcentage d'inertie* associé à l'axe.
— la *nature de la liaison* entre les variables par les proximités des profils lignes / profils colonnes dans la relation quasibarycentrique.

La représentation simultanée permet aussi de qualifier les axes au regard des variables.

Inerties associées aux axes

Toutes les valeurs propres sont inférieures à 1 : les barycentres de N_I (resp. N_J) associés aux poids des profils colonnes (resp. lignes) sont à l'intérieur de N_I (resp. N_J). Une valeur propre λ_s proche de 1 indique une structure proche d'un partitionnement où les réponses à une question sont complètement liées aux réponses de l'autre question. Les valeurs propres expriment l'inertie expliquée sur chacun des axes. Elles se somment par orthogonalité. Il est donc possible de regarder la part d'un axe dans l'explication de l'inertie totale :

$$\frac{\lambda_s}{\sum_s \lambda_s} = \frac{\lambda_s}{\mathcal{I}}$$

Remarques :
— Ce n'est pas parce qu'on a 100% d'inertie sur un plan que le phénomène indique une grande liaison... Cela veut juste dire que le nuage n'est pas déformé dans ce plan
— L'intensité de la liaison s'exprime à travers l'inertie totale $\mathcal{I} = \sum_s \lambda_s = \chi^2/n$. Cette inertie totale vaut au maximum $\inf(I-1, J-1)$, ie dans le cas où toutes les valeurs propres non nulles valent 1 : pour chaque niveau de la question ayant le plus de modalités, le niveau de la réponse à l'autre question est automatiquement défini. En rapportant la valeur observée $\mathcal{I}$ à sa valeur maximale $\inf(I-1, J-1)$, on obtient un indicateur statistique appelé *V de Cramer*

$$V = \left(\frac{\mathcal{I}}{\inf(I-1, J-1)}\right)^{1/2}$$

variant de 0 à 1. Il joue un rôle analogue à celui du coefficient de corrélation. Face à plusieurs variables qualitatives, on peut éditer une matrice des V de Cramer.

Contribution d'un point à l'inertie de l'axe

$$\frac{f_{i\bullet}(OH_i^s)^2}{\lambda_s} \times 100$$

Les points n'ont en général pas les mêmes poids en AFC. On ne peut donc les lire sur le graphique comme en ACP (où elle est à peu près proportionnelle à la distance à l'origine), d'où l'intérêt de ces calculs en AFC.

Qualité de représentation sur un axe

$$\frac{(OH_i^s)^2}{(OM_i)^2} = \cos^2(OM_i, OH_i^s) = \cos^2(\theta_i^s)$$

Il faut avoir suffisamment de points pour que ce critère ait un sens ! La qualité de représentation est utilisée dans les situations suivantes

— trouver un plan dans lequel la modalité est la mieux représentée
— chercher un petit nombre de modalités pour illustrer la signification d'un axe.

3.7 Illustration logicielle

La fonction `CA` du package `FactoMineR` permet de faire une AFC à partir de la matrice X. L'appel de la fonction déclenche les calculs et la visualisation de la représentation simultanée quasi-barycentrique.

```
library(FactoMineR)
quiTravaille=c("H=F","H>F","H")
activite=c("foyer","mi-temps","plein-temps")
X=matrix(c(13,142,106,  30,408,117,  241,573,94 ),byrow=TRUE,nrow=3,ncol=3,
         dimnames=list(quiTravaille,activite))
res=CA(X)
```

Nous comparons ici ces résultats avec une démarche manuelle A partir de la matrice X, calcul de la matrice des fréquences relatives, des marges en lignes et en colonnes

```
n=sum(X)
Fq=X/n
Fj=apply(Fq, 2, sum) # de dimension J: profil moyen ligne  0.16 0.65 0.18
Fi=apply(Fq, 1, sum) # de dimension I: profil moyen colonne 0.15 0.32 0.53
```

Le tableau des fréquences théoriques et le χ^2 d'indépendance :

```
Fij=Fi%*%t(Fj)  # fréquences théoriques
n*sum ( (Fq-Fij)^2/Fij ) # chi^2 à la main
chisq.test(X)            # chi^2 par la fonction R
```

Le calcul des deux nuages :

```
sweep(Fq,1,Fi,FUN="/") # profils lignes
sweep(Fq,2,Fj,FUN="/") # profils colonnes
```

Dans le nuage des profils-lignes, diagonalisation de la matrice S

```
DIinv=diag(1/Fi)#  de dim IxI
DJinv=diag(1/Fj)# de dim JxJ
S=t(Fq)%*%DIinv%*%Fq%*%DJinv
lambda=eigen(S)$values[-1]

res$eig # avec CA: les valeurs propres et % d'inertie
```

La première valeur propre correspondant à la direction OG_l est retirée. On vérifie que le lien entre la somme des valeurs propres et la statistique du χ^2

```
sum(lambda)
chisq.test(X)$statistic/n
```

Calcul des vecteurs propres normés (attention aux poids D_J^{-1} dans définition de la distance)

```
us=eigen(S)$vectors[,-1] #
c=diag(t(us)%*%DJinv%*%us)
us=us%*%diag(1/sqrt(c))
diag(t(us)%*%DJinv%*%us) # bien normé maintenant
```

Le calcul des coordonnées, et l'affichage

```
Plignes= DIinv%*%Fq%*%DJinv%*%us # coordonnées
res$row$coord # avec CA

Plignes2=rbind(Plignes,Plignes[1,]) #pour tracer le dernier segment
plot(Plignes2[,1],Plignes2[,2],type="b",main="nuage des profils lignes",
     xlim=c(-0.7,0.7),ylim=c(-0.2,0.4),xlab="axe1",ylab="axe2")
text(Plignes[,1],Plignes[,2],quiTravaille,pos=3)
abline(v=0,lty=2);abline(h=0,lty=2)
points(0,0);#text(0,0,"profil moyen\n ligne")
```

L'ajout des barycentres exacts

```
Pbaryl=t( Plignes[,1:2]) %*% Fq %*%DJinv
points(Pbaryl[1,],Pbaryl[2,],pch=17,col=2)
text(Pbaryl[1,],Pbaryl[2,],activite,col=2,pos=3)
```

Principe identique pour le nuage des profils colonnes. On vérifie que l'on trouve les mêmes valeurs propres, et on utilise les relations de transitions pour définir l'orientation des vecteurs propres

```
T=Fq%*%DJinv%*%t(Fq)%*%DIinv
vs=Fq%*%DJinv%*%us%*%diag(1/sqrt(lambda))
Pcolonnes= t(DIinv%*%Fq%*%DJinv)%*% vs
res$col$coord # avec CA

Pbaryc=DIinv%*% Fq  %*% Pcolonnes[,1:2]

# on retrouve Pbaryl
Pcolonnes[,1:2]%*%diag(sqrt(lambda[1:2]))
```

Enfin, la représentation simultanée

```
plot(c(Plignes[,1],Pcolonnes[,1]),c(Plignes[,2],Pcolonnes[,2]),
     pch=rep(1:2,c(3,3)),col=rep(1:2,c(3,3)),
     main="représentation simultanée",xlab="axe1",ylab="axe2",
     xlim=c(-0.7,0.7),ylim=c(-0.6,0.6))
text(Pcolonnes[,1],Pcolonnes[,2],activite,col=2,pos=3)
text(Plignes[,1],Plignes[,2],quiTravaille,col=1,pos=3)
abline(h=0,lty=2);abline(v=0,lty=2)
```

Remarque : la fonction `CA` affiche nativement la représentation simultanée quasi-barycentrique. Pour les représentations exactes :

```
# lignes
plot(res,invisible="col") # projection des profils ligne
coord.col=sweep(res$col$coord,2,sqrt(lambda),FUN="*")
points(coord.col,pch=17,col=2)

#colonnes
res.col.coord # identique à Pcolonnes
plot(res,invisible="row") # projection des profils colonne
coord.row=sweep(res$row$coord,2,sqrt(lambda),FUN="*")
points(coord.row)
```

Les listes `res$row`, `res$col` contiennent les coordonnées, l'inertie, les contributions, les cos^2. Il est possible de prendre en compte des colonnes ou des lignes inactives supplémentaires.

3.8 Éléments supplémentaires

Il est possible d'ajouter des éléments supplémentaires dans le tableau de contingence initial. Pour calculer les coordonnées des lignes supplémentaires dans les axes principaux des lignes, on utilise la relation de transition

$$F_s[i_{sup}] = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{i_{sup}j}}{f_{i_{sup}\bullet}} G_s[j]$$

où $f_{\bullet j}$ est le profil-ligne moyen calculé sur le tableau sans élément supplémentaire :

- la proximité d'un point ligne supplémentaire avec un point ligne initial indique des profils de réponse similaire
- la proximité d'un point ligne supplémentaire avec un point colonne initial V_k indique que les individus ayant le profil ligne supplémentaire ont majoritairement choisi la réponse k à la deuxième question.

3.9 Quelques preuves de résultats d'AFC

Propriété 4. *Lorsque le nuage n'est pas recentré il possède une valeur propre qui vaut 1*

Preuve. les points du nuage sont dans un hyperplan de dimension $\min(J-1, I-1)$. La droite (affine) passant par O et le point moyen du nuage $G_l = (f_{\bullet j})$ est orthogonale (au sens de métrique du χ^2) à l'hyperplan d'équation $\sum_j y_j = 1$ contenant le nuage. En effet, soit Y un point du nuage,

$$< OG_l, G_lY >_{\chi_2} = \sum_j \frac{1}{f_{\bullet j}} f_{\bullet j}(y_j - f_{\bullet j}) = \sum_j y_j - \sum_j f_{\bullet j} = \sum_j y_j - 1 = 0$$

Ainsi, la projection orthogonale sur cette droite de direction $(f_{\bullet j})$ de n'importe quel point du nuage est G_l : $(\sum_j \frac{1}{f_{\bullet j}} y_i f_{\bullet j})[f_{\bullet j}] = [f_{\bullet j}]$. L'inertie projetée vaut donc

$$\sum_i f_{i\bullet}(OH_i^{(s)})^2 = \sum_i f_{i\bullet} OG_l^2 = OG_l^2 = \sum_j \frac{1}{f_{\bullet j}} (f_{\bullet j})^2 = 1$$

$\Diamond$

Propriété 5. *Relations quasi-barycentriques*

Preuve. D'après les relations de transition sur les vecteurs propres des deux nuages

$$v_s = \frac{1}{\sqrt{\lambda_s}} F D_J^{-1} u_s$$
$$u_s = \frac{1}{\sqrt{\lambda_s}} F' D_I^{-1} v_s$$

les composantes principales (projection des points-ligne (resp.colonnes) sur les directions propres) peuvent s'écrire

$$F_s = D_I^{-1} F D_J^{-1} u_s = \frac{1}{\sqrt{\lambda_s}} D_I^{-1} F D_J^{-1} F' D_I^{-1} v_s = \sqrt{\lambda_s} D_I^{-1} v_s$$
$$G_s = D_J^{-1} F' D_I^{-1} v_s = \frac{1}{\sqrt{\lambda_s}} D_J^{-1} F' D_I^{-1} F D_J^{-1} u_s = \sqrt{\lambda_s} D_J^{-1} u_s$$

d'où

$$F_s = \frac{1}{\sqrt{\lambda_s}} D_I^{-1} F G_s$$
$$G_s = \frac{1}{\sqrt{\lambda_s}} D_J^{-1} F' F_s$$

La projection des profils-ligne F_s sur une direction est au barycentre (au coefficient $1/\sqrt{\lambda_s}$ près) de la projection des profils colonne profils-colonne G_s affectés des poids $D_I^{-1} F$ ◊

Propriété 6. *Les valeurs propres sont de norme inférieure à 1*

Preuve. A partir des relations quasi-barycentriques, on peut écrire la j-ème composante de F_s sous la forme

$$\sqrt{\lambda_s} F_{si} = \sum_j \frac{f_{ij}}{f_{i\bullet}} G_{sj}$$

soit, en utilisant la propriété du barycentre, pour toute projection F_{si}

$$\min_j G_{sj} \leq \sqrt{\lambda_s} F_{si} \leq \max_j G_{sj}, \quad i = 1, \ldots, I$$

soit,

$$\max_i \sqrt{\lambda_s} F_{si} \leq \max_j G_{sj}$$

et de façon symétrique

$$\max_j \sqrt{\lambda_s} G_{sj} \leq \max_i F_{si}$$

Comme $\max_i F_{si}$ est positive (strictement, sauf cas dégénéré d'une vp nulle)

$$\max_i \lambda_s F_{si} \leq \max_j \sqrt{\lambda_s} G_{sj} \leq \max_i F_{si}$$

d'où le résultat

Remarque : On construit la projection de N_I, puis les barycentres de ces points projetés, associés des poids de profils colonnes : ces J points sont à l'intérieur des projections de N_I ; or ces J points sont à un coefficient près $1/\sqrt{\lambda_s}$ les projections du nuage N_J. De même les barycentres des projection de N_J sont à l'intérieur des projections de N_J, qui sont à un coefficient près $1/\sqrt{\lambda_s}$ les projections du nuage N_J. La représentation simultanée nous indique donc qu'il faut dilater les barycentres, donc que $\sqrt{\lambda_s} \leq 1$ pour la rendre possible ◊

Propriété 7. *Sur chaque axe, la moyenne des coordonnées des individus ligne pondérés par la masse est nulle*

Preuve.

$$\sum_{i=1}^{I} f_{i\bullet} F_{si} = \sum_{i} f_{i\bullet} \sum_{j=1}^{J} \frac{f_{ij}}{f_{i\bullet} f_{\bullet j}} u_{sj} = \sum_{j} u_{sj} = 0 \tag{3.3}$$

Comme les profils-ligne appartiennent au plan affine $\sum_j y_j = 1$, les vecteurs propres u_s (sauf celui associé à la valeur propre triviale égale à 1) appartiennent donc au plan vectoriel associé $\sum_j y_j = 0$. ◊

Propriété 8. *La distance euclidienne entre deux profils lignes représentés par leur coordonnées factorielles est égale à la distance du χ^2 des deux profils initiaux*

Preuve.

$$\begin{aligned}
\sum_{k=1}^{J} (F_{s_k i} - F_{s_k i'})^2 &= \sum_{k=1}^{J} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\bullet} f_{\bullet j}} u_{s_k j} - \frac{f_{i'j}}{f_{i'\bullet} f_{\bullet j}} u_{s_k j} \right)^2 \\
&= \sum_{k=1}^{J} \sum_{j=1}^{J} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 \left(\frac{u_{s_k j}}{f_{\bullet j}} \right)^2 \\
&= \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 \sum_{k=1}^{J} \frac{(u_{s_k j})^2}{f_{\bullet j}} \\
&= \sum_{j=1}^{J} \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2
\end{aligned}$$

en utilisant la normalisation de u_s : $u_s D_J^{-1} u_s = 1$. En effet, soit U la matrice formée par les u_{sk} colonne, alors

$$U' D_J^{-1} U = Id_J$$

Donc $U' D_J^{-1}$ l'inverse de la matrice U. On a donc

$$U' D_J^{-1} U = U U' D_J^{-1} = Id_J$$

soit

$$U U' = D_J$$

◊

Chapitre 4

Analyse des Correspondances Multiples

Comment généraliser l'AFC au cas où plus de deux questions sont mises en correspondance, comme dans le cas typique d'application du traitement de données d'enquête ? A chaque question, l'individu choisit une unique réponse (une modalité ou un niveau) : x_{ij} est la valeur de la modalité prise par l'individu i pour la variable j. Ce tableau initial de dimension $I \times J$, où I désigne le nombre d'individus et J le nombre de variables n'est pas exploitable directement car les variables sont qualitatives. A partir de ce tableau, on construit le tableau disjonctif complet Z qui transforme chaque variable qualitative en autant de variables binaires qu'il y a de modalités : les individus sont en ligne, et l'ensemble de toutes les modalités sur toutes les questions en colonne. Soit K_j le nombre de modalités de la variable qualitative j, ce tableau est donc de dimension $I \times K$, où $K = \sum_{j=1}^{J} K_j$ est le nombre total de modalités sur l'ensemble des variables. Il ne contient que des 0 ou 1 (d'où son nom anglais *one hot encoding*) : z_{ik} vaut 1 si l'individu a choisi la k-ième modalité dans la liste de toutes les modalités.

Ce tableau possède deux propriétés intéressantes. Soit $z_{\bullet k} = \sum_i z_{ik}$ le nombre d'individus ayant choisi la k-ième modalité. Les $(z_{\bullet k})$ sont une ligne résumée, marges des colonnes ; la marge des lignes est une colonne dont toutes les composantes sont égales au nombre de question, car

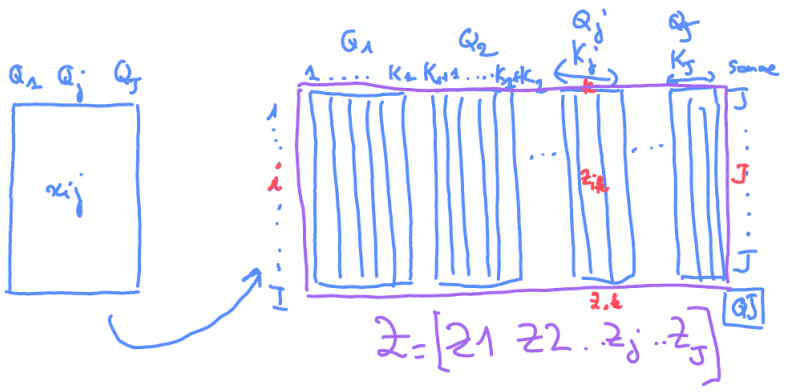


FIGURE 4.1 – Le tableau initial et le tableau disjonctif complet

chaque répondant a fait une unique réponse à chaque question : $z_{\bullet i} = \sum_k z_{ik} = K$. De plus, le tableau possède la particularité de marges suivant les modalités d'une question : $\sum_{k \in Q_j} z_{ik} = 1$ où Q_j désigne l'ensemble des indices des modalités de la question j.

L'objectif est de caractériser les individus vis à vis du profil de leurs réponses aux différentes questions d'une part, les questions vis à vis des individus qui formulent les mêmes réponses d'autre part : quels individus se ressemblent, quelles variables correspondent et comment mettre en relation les individus et les variables. Dans un sens, l'ACM procède à la fois de l'ACP et de l'AFC, et on peut la présenter sous ces deux aspects. Nous choisirons la méthodologie de l'AFC :

— transformation en profils lignes et colonnes
— définition d'un critère d'ajustement avec pondération des points par les profils marginaux (inertie)
— choix de la distance du Khi2.

Bibliographie : Escofier et Pagès (2008), Saporta (2006), Pagès (2011), Cornillon et al. (2018)

4.1 Distances et critère d'ajustement

On étudie les ressemblances entre les individus du point de vue de leurs réponses à l'ensemble des variables. Peut-on déterminer des groupes ayant répondu de la même façon aux questions ?

On transforme le tableau disjonctif en tableau de fréquence $F = Z/(IJ)$: ainsi, $\sum_{i,k} f_{ik} = 1$, comme en AFC (mais attention, les indices i et j ne représentent pas les même entités qu'en AFC).

La somme des coordonnées d'un individu ligne est la même quelle que soit la ligne : $\sum_k f_{ik} = f_{i\bullet} = 1/I$. Ainsi, la transformation en profil ligne $f_{ij}/f_{i\bullet}$ amène à définir un poids $p_i = 1/I$ identique pour chaque individu i, ce qui est à rapprocher de l'ACP équipondérée.

La somme des coordonnées d'un individu-colonne (modalité) vaut $m_k = \sum_i z_{ik}/(IJ) = p_k/J$ où $p_k = I_k/I$ la proportion d'individus ayant choisi la modalité k. On affecte les colonnes de ces poids m_k. Le profil moyen en ligne est $G_\ell = (m_1, \ldots, m_K)$.

On définit alors le critère d'ajustement et la distance du Khi2 :

— les individus sont affectés du poids moyen des lignes $m_i = 1/I$
— les variables sont affectées du poids moyen des colonnes $m_k = z_{\bullet k}/(IJ) = p_k/J$ [attention à la notation : m désigne un poids moyen et l'indice indique s'il s'agit d'une ligne (i) ou d'une colonne (k)].
— la distance du Khi2 entre deux individus lignes est définie par

$$d^2(i,i') = \sum_k \frac{1}{m_k}\left(\frac{f_{ik}}{1/I} - \frac{f_{i'k}}{1/I}\right)^2 = \sum_k \frac{1}{z_{\bullet k}/(IJ)}\left(\frac{z_{ik}}{J} - \frac{z_{i'k}}{J}\right)^2 = \frac{1}{J}\sum_k \frac{I}{z_{\bullet k}}\left(z_{ik} - z_{i'k}\right)^2$$

— le critère d'inertie est

$$Inertie = \sum_i m_i d^2(i, G_l)$$

avec

$$d^2(i, G_l) = \frac{I}{J}\sum_k (\frac{z_{ik}}{z_{\bullet k}} - 2z_{ik} + z_{ik}) = \frac{I}{J}\sum_k (\frac{z_{ik}}{z_{\bullet k}} - z_{ik})$$

en prenant en compte le fait que $z_{ik} = z_{ik}^2$. D'où l'inertie totale du nuage

$$Inertie = \frac{1}{IJ}(\sum_k z_{\bullet k} - \sum_{ik} z_{ik}) = \frac{K}{J} - 1$$

— La distance du Khi2 entre deux (profils) modalités est

$$d^2(k,k') = \sum_i \frac{1}{1/I}\left(\frac{z_{ik}}{z_{\bullet k}} - \frac{z_{ik'}}{z_{\bullet k'}}\right)^2$$

Avec cette définition, la distance entre deux individus est donc liée à la comparaison des individus modalité par modalité et à la rareté de la modalité :

— la distance entre deux individus prenant les mêmes modalités est nulle
— la distance entre deux individus ayant en commun un grand nombre de modalités est faible
— une distance est importante entre deux individus même s'ils ne diffèrent que sur une modalité, si celle-ci est prise par un individu, et rarement chez tous les autres

4.2 Calculs des axes et composantes principales

Du point de vue des calculs, ACM = AFC sur tableau disjonctif complet, même si le tableau disjonctif complet comporte des différences de nature et de propriétés avec un tableau de contingence :

— les valeurs ne sont que des 0 ou 1
— les colonnes peuvent être regroupées en paquet qui somment à 1
— la somme des nombres d'une même ligne est constante égale à J, nombre total de variables.

On déroule exactement les mêmes calculs qu'en AFC avec les notations suivantes :

$$F = \frac{1}{IJ}Z;\;\; D_J = \frac{1}{IJ}D;\;\; D_I = \frac{1}{I}I_n$$

où D est la matrice diagonale de dimension $K \times K$ de coefficients diagonaux $G_\ell = (z_{\bullet k})$; D_I est la matrice diagonale de dimension $I \times I$ de coefficients diagonaux $G_c = (1/I)$.

Il s'agit donc de diagonaliser la matrice la matrice $S = F' D_I^{-1} F D_p^{-1} = \frac{1}{J} Z' Z D^{-1}$.

On en déduit les relations entre les axes principaux des lignes et des colonnes d'ordre s, associés à la s-ème plus grande valeur propre de S :

$$u_s = \frac{1}{\sqrt{\lambda_s}} F' D_I^{-1} v_s;\;\; v_s = \frac{1}{\sqrt{\lambda_s}} F D_J^{-1} u_s$$

Les composantes principales F_s issues de la projection des profils lignes sur l'axe :

$$F_s = D_I^{-1} F D_J^{-1} u_s = \sqrt{\lambda_s} D_I^{-1} v_s = \frac{1}{J\sqrt{\lambda_s}} Z G_s$$

Les composantes principales issues de la projection des profils colonnes sur l'axe

$$G_s = D_J^{-1} F' D_I^{-1} v_s = \sqrt{\lambda_s} D_J^{-1} u_s = \frac{1}{\sqrt{\lambda_s}} D^{-1} Z' F_s$$

4.2.1 Relations de transition

On retrouve des relations de dualité entre les deux nuages :

— au coefficient $\sqrt{\lambda_s}$ près, l'individu i se trouve au point moyen du nuage des modalités qu'il a choisies

$$F_s[i] = \frac{1}{J\sqrt{\lambda_s}} \sum_{k \in R(i)} G_s[k]$$

où on note $R(i)$ l'ensemble des indices des modalités choisies par i

— avant dilatation, la modalité k se trouve au point moyen du nuage des individus qui l'ont choisie pour réponse

$$G_s[k] = \frac{1}{z_{\bullet k}\sqrt{\lambda_s}} \sum_{i \in I(k)} F_s[i]$$

où on note $I(k)$ l'ensemble des individus ayant choisi la modalité k

Mais en général, on ne fait pas la représentation barycentrique, les individus étant projetés trop près du centre de gravité du nuage.

4.2.2 Sous-nuage des modalités associées à une même variable

Le sous-nuage des modalités d'une même variable possède une propriété intéressante : il a même centre de gravité que le nuage total. En effet, les coordonnées du sous-nuage de la question j sont le colonnes $Z_j D_J^{-1}$ où Z_j est le tableau extrait de Z ne prenant que les colonnes correspondant aux modalités de la question j et les éléments diagonaux de $\frac{1}{I} D_j$ sont les masses des points K_j points de ce sous-nuage. En effet

$$\sum_{k \in Q_j} z_{ik} = 1$$

et la i-ème composante du centre de gravité du sous-nuage vaut

$$\sum_{k \in Q_j} \frac{d_{kk}}{I} \frac{z_{ik}}{d_{kk}} = \frac{1}{I}$$

qui est également la i-ème composante du centre de gravité de la marge de colonnes.

4.2.3 Inertie du nuage des modalités

Avec la distance choisie, la distance d'une modalité au centre de gravité est d'autant plus grande que l'effectif est plus faible :

$$d^2(k, G_c) = \sum_i \frac{1}{1/I} \left(\frac{z_{ik}}{z_{\bullet k}} - \frac{1}{I} \right)^2 = \frac{I}{z_{\bullet k}} - 1$$

L'inertie de la modalité k vaut

$$Inertie(k) = m_k d^2(k, G_c) = \frac{z_{\bullet k}}{IJ} \left(\frac{I}{z_{\bullet k}} - 1 \right) = \frac{1}{J} \left(1 - \frac{z_{\bullet k}}{I} \right)$$

d'où l'inertie d'une modalité qui est d'autant plus grande que l'effectif est faible pour cette modalité. Il faut donc éviter les modalités rares qui font influencer fortement l'analyse ; le cas échéant, *regrouper les modalités de faible effectif.*

On note que l'inertie totale vaut $K/J - 1$, comme dans le nuage des individus.

On peut aussi déterminer l'inertie d'une question

$$Inertie(j) = \sum_{k \in Q_j} Inertie(k) = \frac{1}{Q}(K_j - 1)$$

où la même notation est employée pour l'inertie de la question et celle d'une de ses modalités, l'indice permettant de référer à l'une ou l'autre. On voir que l'inertie d'une question est fonction du nombre de modalités de sa réponse. Il vaut donc mieux *équilibrer le nombre de réponses possibles aux différentes questions.*

4.3 Interprétation

On a vu que dans la représentation barycentrique exacte, l'individu i est au centre de gravité des modalités qu'il a choisies ; la modalité k est au centre de gravité des individus qui la possède. Les représentations simultanées barycentriques exactes ne sont en général pas possibles (trop d'individus qui se retrouvent proches du centre du repère), on construit une représentation compromis

- en construisant le nuage des individus : deux individus sont proches s'ils ont de nombreuses modalités en commun.
- puis en positionnant les modalités en multipliant les coordonnées des barycentres sur l'axe de rang s par le coefficient $\sqrt{\lambda_s}$. Le barycentre de toutes les modalités d'une variable est au centre de gravité du nuage des individus, donc confondu avec l'origine des axes

Ceci évite d'avoir toutes les modalités concentrées au centre du graphique.

Compte tenu des distances et des relations barycentriques, on exprime

- la proximité entre deux individus entre termes de *ressemblance* de profil (ils ont choisi globalement les mêmes modalités
- la proximité entre modalités de variables différentes en terme d'*association* : ces modalités sont les points moyens des individus qui les ont choisies et sont proches parce qu'elles concernent des individus assez semblables.
- la proximité entre modalités d'une même variable en terme de *ressemblance* : par construction, les modalités d'une même variable s'excluent. Si elles sont proches, c'est parce que les individus qui les ont choisis se ressemblent.

Synthèse des variables qualitatives Un aspect de la synthèse par les méthodes factorielles est la mise en avant d'un petit nombre de variables synthétiques. Montrons que les facteurs F_s trouvés par l'ACM sont effectivement liés le plus possible à l'ensemble des variables initiales. Pour ce faire, on utilise le *facteur de corrélation*, qui mesure la liaison entre une variable numérique F_s et une variable quantitative j.
Une variable qualitative définit une partition de l'ensemble des individus en autant de classes qu'elle a de modalités. Par le théorème de Huygens

$$\text{var}(F_s) = \text{ inertie inter } + \text{ inertie intra}$$

L'inertie inter classe représentant l'inertie des centres de gravité des classes, tandis que l'inertie intra classe étant celle de chaque point au centre de gravité de la classe à laquelle il appartient. Le carré du rapport de corrélation est le rapport de l'inertie inter par l'inertie totale

$$\eta^2(F_s, j) = \frac{\text{inertie inter}}{\text{inertie totale}} = J \sum_{k \in Q_j} (\text{ inertie de la modalité } k \text{ projetée sur l'axe } v_s \ )$$

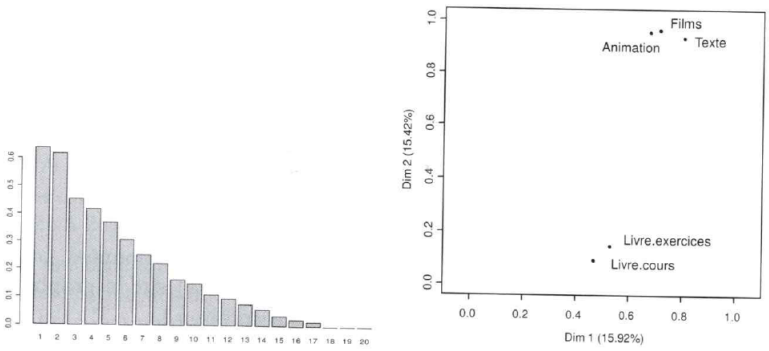


FIGURE 4.2 – Eboulis des valeurs propres (à gauche) ; carré des corrélations (à droite) (Pagès, 2011)

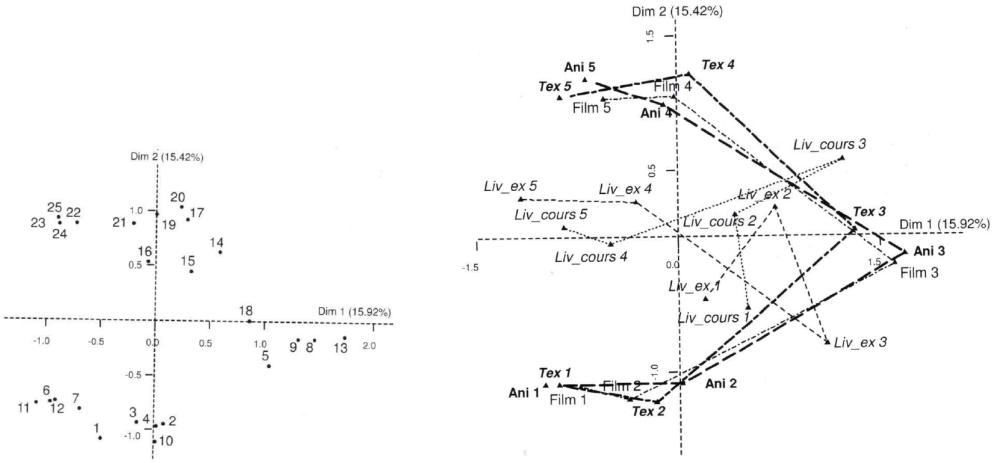


FIGURE 4.3 – Nuage des individus (à gauche) ; nuage des modalités(à droite) (Pagès, 2011)

L'ACM maximise l'inertie projetée sur l'ensemble des modalités. En regroupant les modalités par variables, l'ACM maximise donc $\sum_j \eta^2(F_s, j)/J$, moyenne des carrés des rapports de corrélation. On peut ainsi dans l'interprétation

- calculer la contribution d'une variable à l'inertie d'un facteur
- construire le graphique des carrés des liaisons : dans le plan (F_s, F_t), on construit les points $(\eta^2(F_s, j), \eta^2(F_t, j))$ pour chaque variable j. On montre que ce graphique s'interprète comme la projection d'un nuage dans lequel chaque point représente une variable, la proximité entre points-variables traduisant la ressemblance entre les partitions engendrées par les deux variables.

Un exemple Pagès (2011) détaille l'exemple d'un questionnaire sur l'utilité de $J = 5$ supports pédagogiques : trois issus d'un cours en ligne (texte, animations et films décrivant des logiciels) et deux livres (l'un de cours, l'autre d'exercices). Il y a cinq modalités par question de l'utilité de ces supports (de 1 = inutile à 5 = très utile), soit $K = 25$. $I = 25$ étudiants ont répondu au questionnaire. Les variables ont été considérées comme qualitatives, et une ACM est effectuée.

L'éboulis des valeurs propres en Figure 4.2 montre deux valeurs propres assez fortement plus grandes que les autres, mais peu différentes l'une de l'autre. Les pourcentages d'inertie associées à des pourcentages d'inertie 15.9% et 15.5% peuvent sembler faibles ; mais compte tenu du nombre de variables, il ne peut excéder 25% : si toutes les variables étaient identiques, alors il n'y aurait qu'un axe de 100% d'inertie en ACP, alors que l'ACM conduit à $K_j - 1 = 4$ axes qui se partagent équitablement 100% de l'inertie.

Les deux première valeurs propres valent 0.64 et 0.62 ; ce sont des moyennes de carrés de corrélation, et peuvent être considérées comme élevées. Si on décompose cette inertie selon les variables, on obtient peu tracer le carré de corrélation suivant deux axes factoriels. Les livres interviennent dans la formation du premier axe, les cours en ligne contribuent à la formation des deux axes.

Le nuage des individus (Figure 4.3 à gauche) présente trois groupes ; l'axe 1 en sépare deux, le premier axe oppose les points 18, 5, 8 et 13 aux autres.

Les modalités étant ordonnées, elles sont liées par des traits qui permettent de voir leur *trajectoire* (Figure 4.3 à droite).

L'ACM permet une représentation des individus mettant en avant leurs principales variabilités ; une représentation des variables montrant leurs associations remarquables :

- L'axe 2 sépare les étudiants ayant jugé les trois composants en ligne inutile (en bas) et ceux les ayant jugés très utiles (en haut).
- Cette opinion n'est pas liée à celle de l'utilité des livres qui s'exprime plus sur le premier axe (très utile - assez utile) ;
- le premier axe oppose les modalités moyennes aux modalités extrèmes concernant les cours en ligne.

4.4 Tableau de Burt

Pour traiter l'analyse des correspondances multiples, on pourrait dresser un hypertableau de contingence (autant de dimension que de variables) qui n'a pas d'intérêt pratique, à cause du nombre de cellules à effectifs nuls, croissant très rapidement avec le nombre de dimensions. En revanche, il est possible de construire un tableau symétrique, croisant deux à deux les variables qualitatives, de taille $K \times K$, $K = \sum_j K_j$

- sur la diagonale du tableau complet : matrices bloc de taille $K_j \times K_j$ diagonale. Sur la diagonale, on trouve I_k le nombre total de réponse à la modalité k
- au croisement de la variable j et de la variable j', on trouve le tableau de contingence de tailles $K_j \times K_{j'}$ associé à ces deux variables.

Le tableau de Burt est symétrique, et analogue à une matrice des corrélations, puisqu'il récapitule les liaisons entre les variables deux à deux. Il contient beaucoup moins d'information que l'hypertableau, et ne permet pas de reconstruire le tableau disjonctif complet.

- la $k$ième ligne du tableau de Burt = somme des lignes du tableau disjonctif qui présentent la modalité k : dans R^K, la profil de la modalité k du tableau de Burt se trouve au barycentre des profils des individus i du tableau disjonctif qui la possède
- AFC du tableau de Burt et celle du tableau disjonctif aboutissent au même résultat : analyse du nuage des barycentres = AFC du tableau de Burt
  - axe d'inertie du nuage des individus (disjonctif) = axe d'inertie du nuage de leur barycentres (Burt)
  - pour obtenir une représentation simultanée

Chapitre 5

Clustering

Les méthodes factorielles permettent de représenter des individus dans des plans privilégiés, et d'interpréter leurs caractéristiques en fonction de facteurs sous-jacents. Si on peut visuellement proposer de grouper les observations (ou les variables), ces regroupements restent empiriques. Les méthodes de *classification non supervisée* ou *clustering* permettent de partitionner des individus en groupes qui se ressemblent en n'ayant aucune information a priori sur les groupes. Elles font partie des méthodes d'apprentissage non supervisé, par opposition aux méthodes supervisées (ou méthode de classification, ou discrimination) qui déterminent une règle de *classement* à partir d'un échantillon d'apprentissage pour lequel le groupe d'appartenance est connu.

Le clustering pose deux problèmes : d'abord, définir une méthode de partitionnement de n observations en K groupes (clusters) sans chevauchement, c'est à dire trouver une partition des observations $C_1, \dots C_K$, qui peut être définie sur les obsevrations ou sur leurs indices

— $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, n\}$
— $C_k \cap C_{k'} = \varnothing$ pour tout $k \neq k'$

Deuxièmement, le choix du nombre K de groupes. On cherche donc K et $f : \mathcal{X} \to \{1, \dots, K\}$ telle $x_i \mapsto f(x_i) = k_i$ le numéro du groupe (ou classe, ou cluster) auquel x_i appartient.

Exemples

— classification automatique d'images sans échantillon préalablement labellisé
— segmentation de clientèle en marketing
— profilage : classification de courbes de consommation électrique
— classification de variables

Motivations

— explicatives : interprétation des groupes
— utilisation dans une étape ultérieure (souvent supervisée) : traitement différencié et particularisé des groupes ; stratification pour des sondages ; définition d'un individu moyen

Méthodes On cherche à regrouper des individus qui se ressemblent en eux, mais dissimilaires aux individus des autres classes (ie, avoir des classes bien séparées si possible). Il faut donc commencer à définir les notions de similarité et dissimilarité. Ensuite, il y a différentes approches : par partitionnement (méthode des K-moyennes, la classification ascendante hiérarchique), par modélisation probabiliste (modèles de mélange), par continuité ou spectrale. Nous abordons ici les deux premiers types.

5.1 Dissimilarité

Afin de regrouper les individus qui se ressemblent, il est nécessaire de définir un critère de dissimilarité

Définition 6. *Une matrice de* dissimilarité $D = (d_{ij})$ *est telle que* $d_{ii} = 0$, $d_{ij} \geq 0$ *et* $d_{ij} = d_{ji}$. *L'inégalité triangulaire* $d_{ik} \leq d_{ij} + d_{jk}$ *n'est pas requise.*

Si S est une matrice de similarité, $\max(S) - S$ est une matrice de dissimilarité. Dans le cas de *vecteurs* x_i de p composantes, la dissimilarité est construite comme la somme des dissimilarités de chacune des composantes :

$$\Delta(x_i, x_{i'}) = \sum_{j=1}^{p} d(x_{ij}, x_{i'j})$$

Quelques exemples de dissimilarité :

— le carré de la distance *euclidienne* : $d(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$
— la distance ℓ_1 : $d(x_{ij}, x_{i'j}) = |x_{ij} - x_{i'j}|$
— variables catégorielles : $d(x_{ij}, x_{i'j}) = \mathbb{1}(x_{ij} \neq x_{i'j})$

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

— Dans le cas de variables standardisées

$$\Delta(X_j, X_{j'}) = \sum_{i} (X_{ij} - X_{ij'})^2 = 2(1 - \mathrm{cor}(X_j, X_{j'}))$$

Dans la suite, on notera d la fonction de dissimilarité, qu'elle soit pour une composante ou pour un vecteur.

Définition 7. *On appelle dissimilarité de la classe* C_k *la somme des dissimilarités entre ses élements*

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} d(x_i, x_i')$$

où $|C_k|$ *désigne le cardinal de la classe* C_k. *La dissimilarité intra-classe est la somme des dissimilarité de chaque classe*

$$I_{intra} = \sum_{k=1}^{K} W(C_k)$$

5.2 Méthodes de partitionnement

Le problème de partitionnement revient à minimiser la dissimilarité intra classe I_{intra} sur l'ensemble des partitions possibles. Il cumule donc deux handicaps : d'une part, il n'est pas possible de calculer de façon exhaustive I_{intra} sur toutes les partitions possibles ; d'autre part, le calcul des dissimilarités deux à deux est coûteux. Dans le cas de la dissimilarité basée sur le carré de la distance euclidienne, le deuxième problème s'évanouit, grâce à la relation entre la somme des carrés des distances deux à deux des éléments d'un ensemble et son l'inertie par rapport à son centre de gravité. Nous nous plaçons dans ce cas dans la suite.

5.2.1 K-moyennes

Dans le cas d'une dissimilarité basée sur le carré de la distance euclidienne, le problème de minimisation s'écrit :

$$\min_{C_1,\ldots,C_k} \sum_{k=1}^{K} \sum_{i,j \in C_k} ||x_i - x_j||^2$$

Algorithme

Comme il y a K^n partitions possibles, ce problème de minimisation est NP complet, mais un algorithme très simple permet de converger vers un minimum *local*. Il utilise le fait que la dissimilarité à l'intérieur d'une classe est la somme des inerties des points de la classe par rapport à son centre de gravité.

Algorithme des K-moyennes ou centres mobiles
1- *Initialisation* : affecter un nombre de 1 à K à chaque observation
2- *Itérer* jusqu'à ce que les assignations soient stables :
(a) Pour chaque k, calculer le centre comme le point moyen des observations affectées à ce groupe
(b) Affecter chaque observation au nouveau centre qui est le plus proche

La convergence de cet algorithme est garantie :

— L'étape 2-a permet de réécrire l'inertie facilement à partir des centres g_k, moyenne des individus de la classe C_k. En effet, par un calcul que nous avons déjà rencontré en ACP,

$$\begin{aligned} W(C_k) =& \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \frac{1}{|C_k|} \sum_{i,i' \in C_k} ||x_i - x_{i'}||^2 \\ =& \frac{1}{|C_k|} \sum_{i,i' \in C_k} \left(||x_i - g_k||^2 + ||x_{i'} - g_k||^2 \right) + \frac{1}{|C_k|} < \underbrace{\sum_{i \in C_k} (x_i - g_k)}_{=0}, \underbrace{\sum_{i' \in C_k} (x_{i'} - g_k)}_{0} > \\ =& 2 \sum_{i \in C_k} ||x_i - g_k||^2 \end{aligned}$$

Notons que tout autre choix que g_k ne fait qu'augmenter la somme des carrés des déviations à g_k et ne peut convenir :

$$\begin{aligned} \sum_{i \in C_k} ||x_i - g_k||^2 =& \sum_i ||x_i - a + a - g_k||^2 \\ =& \sum_i ||x_i - a||^2 + |C_k| \, ||a - g||^2 + 2|C_k| < g_k - a, a - g_k) \end{aligned}$$

d'où

$$\sum_i ||x_i - a||^2 = \sum_i ||x_i - g_k||^2 + |C_k| \, ||a - g||^2 \geq I$$

Cette étape ne modifie pas la valeur du critère.

— A l'étape 2-b, la réallocation ne peut qu'améliorer le critère. En effet, soient deux classes $C_k^{(t)}$ et $C_{k'}^{(t)}$ définies à l'itération t. La somme de leur inertie intra s'écrit

$$I_{kk'}^{(t)} = \sum_{i \in C_k^{(t)}} ||x_i - g_k^{(t)}||^2 + \sum_{i \in C_{k'}^{(t)}} ||x_i - g_{k'}^{(t)}||^2.$$

Soit $C_{k \to k'}^{(t)}$ l'ensemble des observations de $C_k^{(t)}$ qui sont plus proches de $g_{k'}^{(t)}$ que de $g_k^{(t)}$. Alors :

$$\begin{aligned} I_{kk'}^{(t)} &= \sum_{i \in C_k^{(t)} \setminus C_{k \to k'}^{(t)}} ||x_i - g_k^{(t)}||^2 + \sum_{i \in C_{k \to k'}^{(t)}} ||x_i - g_k^{(t)}||^2 + \sum_{i \in C_{k'}^{(t)}} ||x_i - g_{k'}^{(t)}||^2 \\ &\geq \sum_{i \in C_k^{(t)} \setminus C_{k \to k'}^{(t)}} ||x_i - g_k^{(t)}||^2 + \sum_{i \in C_{k'}^{(t)} \cup C_{k \to k'}^{(t)}} ||x_i - g_{k'}^{(t)}||^2 \\ &\geq \sum_{i \in C_k^{(t)} \setminus C_{k \to k'}^{(t)}} ||x_i - \tilde{g}_k^{(t)}||^2 + \sum_{i \in C_{k'}^{(t)} \cup C_{k \to k'}^{(t)}} ||x_i - \tilde{g}_{k'}^{(t)}||^2 \end{aligned}$$

où $\tilde{g}_k^{(t)}$ et $\tilde{g}_{k'}^{(t)}$ sont les barycentres des ensembles $C_k^{(t)} \setminus C_{k \to k'}^{(t)}$ et $C_{k'}^{(t)} \cup C_{k \to k'}^{(t)}$ respectivement, puis qu'on a vu que l'inertie est minimisée quand elle est calculée par rapport au centre de gravité.

Ainsi, le critère ne peut jamais diminuer, jusqu'à l'obtention d'un *optimum local*

— l'optimum n'étant que local, il faut lancer plusieurs fois l'algorithme et prendre la solution donnant la plus petite valeur du critère
— Il faut fixer le germe du générateur pour une solution reproductible
— Cet algorithme construit les clusters en nombre K donné, mais ne détermine pas la valeur de K.

Cet algorithme est aussi appelé K-moyenne en français, ou moyennes mobiles pour prendre en compte les modification de centre de gravité à chaque itération.

Cet algorithme est illustré sur le jeu de données seeds [1] qui répertorie les caractéristiques géométriques, toutes quantitatives, de différents grains de blé : la surface (`area`), le périmètre (`perimeter`), la compacité (`compact`), la longueur (`length`), la largeur (`width`), l'asymétrie (`asym`), la longueur du sillon (`lgroove`).

5.2.2 Fuzzy kmeans

Pour chaque i et chaque k, on définit la variable d'assignation $z_{ik} = \mathbb{1}_{\{i \in C_k\}}$ qui vaut 1 si le point i est dans la classe k et 0 sinon. On a $\sum_k z_{ik} = 1$ et $\sum_i \sum_k z_{ik} = n$, le nombre total de points (individus). $z_i = (z_{i1}, \ldots, z_{iK})$ est la variable mutinomiale associée, on on s'autorise l'abus de notation $z_{ik} = 1 \Leftrightarrow z_i = k$.

Notons $\mu = (\mu_1, \ldots, \mu_K)$ et $z = (z_1, \ldots, z_K)$. On peut réécrire le problème de minimisation précédent sous la forme

$$\min_{z, \mu;\ \text{t.q.} \sum_k z_{ik} = 1} \sum_k \sum_i z_{ik} ||x_i - \mu_k||^2.$$

et on retrouve que $\mu_k = \sum_i z_{ik} x_i / \sum_i z_{ik} = g_k$ pour z donné.

1. http ://archive.ics.uci.edu/ml/datasets/seeds

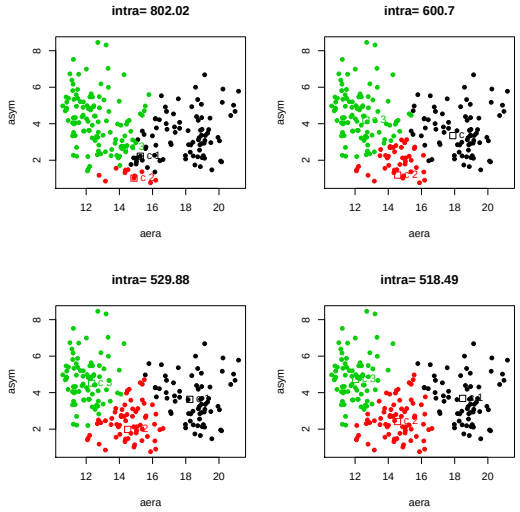


FIGURE 5.1 – Différentes étapes de l'algorithme des K-moyennes (K=3) représentées sur le premier plan principal

Nous avons vu que le fait que les z_{ik} soient contraints d'être entiers rend le problème NP complet. Afin de le simplifier, on peut le relaxer, c'est à dire mettre moins de contraintes. En particulier, au lieu de contraindre les z_{ik} à valoir exactement 0 ou 1, on peut décider d'être plus permissifs et leur permettre d'appartenir au segment $[0,1]$. On appelle $c_{ik} \in [0,1]$ ces variables ; elles doivent toujours vérifier la contrainte $\sum_k c_{ik} = 1$ qui est conservée.

Soit $m \in]1;+\infty[$. Le problème de minimisation sous contrainte suivant devient un problème facile

$$\min_{\mu;c;\ \text{tq}\ \sum_k c_{ik}=1} \sum_k \sum_i c_{ik}^m ||x_i - \mu_k||^2.$$

On écrit le Lagrangien avec n multiplicateurs de Lagrange λ_i :

$$\mathcal{L} = \sum_i \sum_k c_{ik}^m ||x_i - \mu_k||^2 - \sum_i \lambda_i \sum_{k'} (c_{ik'} - 1).$$

On dérive par rapport à μ_k et on trouve

$$\mu_k = \sum_i c_{ik}^m x_i / \sum_i c_{ik}^m. \tag{5.1}$$

On dérivant par rapport à c_{ik} :

$$m c_{ik}^{m-1} ||x_i - \mu_k||^2 - \lambda_i = 0$$

soit, en utilisant les contraintes,

$$c_{ik} = \left(\sum_{k'} \frac{||x_i - \mu_k||^{2/(m-1)}}{||x_i - \mu_{k'}||^{2/(m-1)}} \right)^{-1}. \tag{5.2}$$

La résolution est alors itérative, en itérant les mises à jour (5.1) et (5.2). A convergence, on affecte l'observation i à la classe de plus grand c_{ik}.

Cette méthode donne des classes compactes et est relativement efficace. Mais elle nécessite de définir une initialisation (en tirant K centres par exemple) ; elle peut être être piégée dans un mimimum local ; elle est influencée par les outliers et non adaptée aux classes non convexes ou peu compactes.

Remarque *Fuzzy* veut dire flou. La classification est dite floue, car la variable d'optimisation ne vaut plus exactement 0 ou 1, elle est moins nette, comme floutée.

5.2.3 Kernel Kmeans

Pour permettre de gérer des clusters qui ne sont pas linéairement séparables dans l'espace d'origine $\mathcal{X}$, on définit une fonction non linéaire ϕ de l'espace d'origine $\mathcal{X}$ vers un espace $\mathcal{Y}$, souvent de plus grande dimension qui, à tout point $x_i \in \mathcal{X}$ associe $\phi(x_i)$. On espère que les clusters deviennent linéairement séparables dans $\mathcal{Y}$. La difficulté est de définir $\mathcal{X}$!

Soit $m > 1$, on réécrit le problème de fuzzy kmeans avec les points projetés sans $\mathcal{Y}$

$$W_{fuzzy}(c) = \sum_i \sum_k c_{ik}^m ||\phi(x_i) - \bar{\phi}_k(x)||^2$$

On reprend les mêmes formules

$$\bar{\phi}_k(x) = \frac{\sum_i c_{ik}^m \phi(x_i)}{\sum_i c_{ik}^m}$$

avec

$$c_{ik} = \left(\sum_{k'} \frac{||\phi(x_i) - \bar{\phi}_k(x)||^{2/(m-1)}}{||\phi(x_i) - \bar{\phi}_{k'}(x)||^{2/(m-1)}} \right)^{-1}.$$

En remplaçant $\bar{\phi}_k(x)$ dans $||\phi(x_i) - \bar{\phi}_k(x)||^2$ et en développant, on voit que l'expression ne fait intervenir que des produits scalaires d'images $\phi(x_i)'\phi(x_j)$. Ainsi, il n'est pas nécessaire de définir ϕ en totalité, mais juste les produits scalaires entre les images, grâce à l'intermédiaire d'une fonction K qui, à $(x_i, x_j) \in \mathcal{X} \times \mathcal{X}$, associe :

$$K(x_i, x_j) = \phi(x_i)'\phi(x_j).$$

Cette fonction est appelée *noyau*, et ce principe est dit *astuce du noyau*.

Quelques exemples de noyau

— polynomial : $K(x, y) = (x'y + a)^b$
— gaussien : $K(x, y = \exp -(||x - y||^2)/(2\sigma^2)$
— sigmoïde : $K(x, y) = \tanh(ax'y + b)$

5.3 Classification Ascendante Hiérarchique

La Classification Ascendante Hiérarchique (CAH) a pour objectif de construire un arbre par agrégation successive de sous-groupes : les partitions sont successivement emboîtées et permettent une visualisation "en profondeur", contrairement à la visualisation "à plat" des K-moyennes. On dit que l'approche est ascendante, bottom-up ou agglomérative. C'est une méthode *hiérarchique*. L'arbre généré s'appelle un *dendogramme*.

Comme pour les K-moyennes, il faut définir une mesure de dissimilarité entre individus et entre classes.

Algorithme CAH
1- *Initialiser* : Choisir une mesure de dissimilarité sur les n individus et prendre la partition en n classes
2- *Boucler* pour $i = n, n-1, \ldots, 2$:
(a) Examiner toutes les paires de dissimilarités inter-cluster
↪ Identifier la paire de clusters le moins dissimilaires.
↪ Fusionner les deux clusters.
↪ La dissimilarité entre ces deux clusters indique la hauteur de la branche de l'arbre
(b) Calculer les nouvelles dissimularités inter-cluster parmi les $i-1$ clusters restants

Définition 8. *On appelle* saut *ou* linkage *la dissimilarité inter-classe.*

Complet Calcul de toutes les dissimilarités entre les observations de C_k et $C_{k'}$, et conserver la *plus grande*. Ce critère tend à produire des classes de diamètres égaux, mais est très sensible aux outliers.

Simple Calcul de toutes les dissimilarités entre les observations de C_k et $C_{k'}$, et conserver la *plus petite*. Ce critère est sensible à l'effet de chaîne. Il permet donc de bien prendre en compte des classes allongées ou sinueuses. Mais des classes reliées entre elles par des points isolés pourront se voir regroupées.

Moyenne Calcul de toutes les distances entre une observation de C_k et une observation de $C_{k'}$, et en calculer la *moyenne*. C'est un intermédiaire entre le linkage complet et le linkage simple.

Centre Dissimilarité entre les *centres* de C_k et $C_{k'}$. Est souvent utilisée en génomique, mais elle souffre d'un inconvénient majeur, dit d'inversion : deux clusters peuvent être fusionnés à une hauteur inférieure à n'importe lequel des deux clusters individuels. Elle est robuste aux outliers.

Ward saut=inertie inter-classe. C'est l'une de celles qui correspondent le mieux à l'objeectif de la classification. Elle tend à produire des classes sphériques et de mêmes effectifs, mais peu efficace quand les classes sont allongées.

Méthode de Ward

Dans le cadre euclidien, la variance est décomposée suivant le théorème de Huygens

$$\begin{aligned}&\text{Inertie totale } (C_k \cup C_{k'})\\ &= \underbrace{\text{Inertie } (C_k) + \text{Inertie } (C'_k)}_{\text{intra}} + \underbrace{\text{Inertie } (G_{C_k}, G) + \text{Inertie } (G_{C_{k'}}, G)}_{\text{inter}}\end{aligned}$$

En effet, si on appelle g_k le point moyen de la classe C_k à $|C_k|$ éléments, g_ℓ le point moyen de la classe C_ℓ à $|C_\ell|$ éléments, le barycentre de la classe issu du regroupement est

La méthode de Ward maximise la qualité de la partition obtenue :
— les individus sont homogènes à l'intérieur d'une classe (variance intra-classe faible)
— les individus sont différents d'une classe à l'autre (variance inter-classe importante)
— la qualité d'une partition est mesurée par le rapport inertie inter-classes / inertie totale

Propriété 9. *L'augmentation $\delta(k,\ell)$ d'inertie intra-classes due au regroupement de deux classes k et ℓ avec la méthode de Ward vaut*

$$\delta(k,\ell) = \frac{n_k n_\ell}{n_k + n_\ell} ||G_k G_\ell||^2$$

où n_k et n_ℓ sont les effectifs des classes k et ℓ et G_k et G_ℓ leurs centres de gravité respectifs.

Preuve. Soit G le centre de gravité du nuage avant le regroupement et G_k et G_ℓ les centres d'inertie des deux classes k et ℓ. Soit $G_{k\ell}$ le centre d'inertie des deux classes. L'inertie inter-classe de ces deux classes avant le regroupement vaut

$$\begin{aligned} Inter_{avant} =& n_k ||GG_k||^2 + n_\ell ||GG_\ell||^2 \\ =& \underbrace{n_k ||G_{k\ell}G_k||^2 + n_\ell ||G_{k\ell}G_\ell||^2}_{\text{inertie inter perdue lors du regroupement}} + \underbrace{(n_k + n_\ell)||GG_{k\ell}||^2}_{\text{Inertie inter après regroupement}} \\ & + 2 < \underbrace{n_k\ G_{k\ell}G_k + n_\ell\ G_{k\ell}G_\ell}_{=0}, GG_{k\ell} > \end{aligned}$$

Le double produit est nul par définition du centre d'inertie

$$OG_{k\ell} = \frac{n_k OG_k + n_\ell OG_\ell}{n_k + n_\ell}$$

des deux classes après le regroupement. Le saut d'inertie s'écrit donc

$$\begin{aligned} \delta(k,\ell) &= n_k ||OG_k - \frac{n_k OG_k + n_\ell OG_\ell}{n_k + n_\ell}||^2 + n_\ell ||OG_\ell - \frac{n_k OG_k + n_\ell OG_\ell}{n_k + n_\ell}||^2 \\ &= \frac{n_k}{(n_k + n_\ell)^2} ||n_\ell G_k G_\ell||^2 + \frac{n_\ell}{(n_k + n_\ell)^2} ||n_k G_k G_\ell||^2 \\ &= \frac{n_k n_\ell}{n_k + n_\ell} ||G_k G_\ell||^2 \end{aligned}$$

◊

On choisira le regroupement qui fait perdre le moins d'inertie inter-classe (ie, qui fait augmenter le moins l'inertie intra-classe). Les amas ont tendance à rester dispersés.

Définition 9. *L'augmentation d'inertie intra-classe δ_t à l'étape t est appelée* indice d'agrégation

L'augmentation d'inertie intra-classe δ_t a les propriétés suivantes :

— L'agrégation se fait suivant des classes de centre de gravité proches, ou de faibles effectifs
↪ classes de tailles homogènes

— $\delta_t \geq \delta_{t-1}$
↪ L'arbre ne présente pas d'inversion : un élément c de $C_k \cap C'_k$ ne s'aggrège pas un niveau inférieur à celui de l'agrégation entre C_k et C'_k.

— La somme de tous les indices est égale à l'inertie totale : $\sum_{t=1}^{n-1} \delta_t =$ inertie totale

CAH en pratique

La construction est ascendante, mais l'analyse se fait de façon descendante et permet un choix du nombre de classes en examinant l'allure générale de l'arbre et repérant les irrégularités dans le diagramme en bâton des niveaux des noeuds (règle du coude)

$$\min_{q_m \leq q \leq q_M} \frac{\delta_q}{\delta_{q+1}}$$

On peut préférer un nombre de classes peu élevé pour une vision synthétique, ou un peu plus élevé pour permettre une meilleure interprétation des classes. L'ACP peut servir aux représentations "à plat" et/ou pour réduire la dimension.

Sur la figure 5.2, le diagramme en barre des indices d'agrégration de l'exemple seeds suggère un nombre de clusters de 2, 4, ou éventuellement 7. On voit sur la figure 5.3 le découpage successif des partitions.

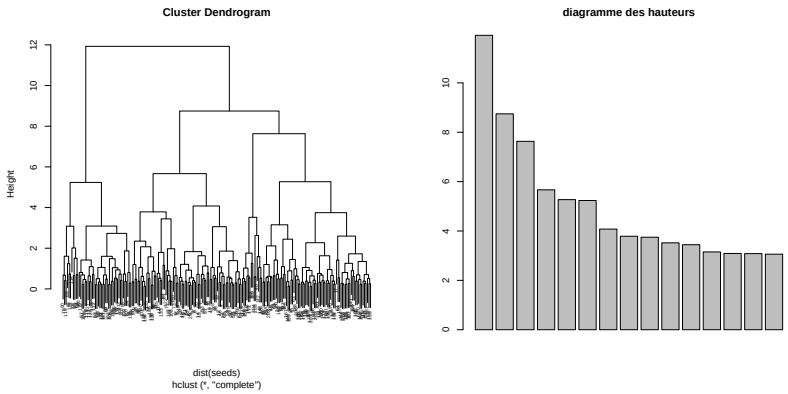


FIGURE 5.2 – Dendogramme et diagramme en barre des indices d'agrégation

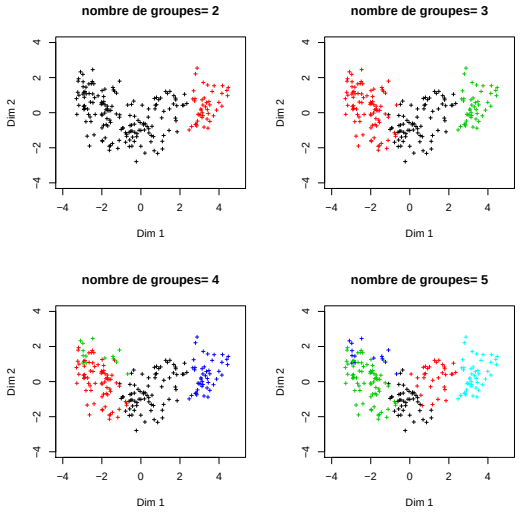


FIGURE 5.3 – Les groupes sélectionnés suivant la CAH suivant le premier plan principal

5.4 Méthode mixte

Les méthodes de partitionnement optimisent un critère global et peuvent traiter un grand nombre d'individus, mais ne peuvent fixer a priori le nombre de classes. D'un autre côté, l'issue

d'une CAH permet de choisir un nombre de classes. La combinaison des deux méthodes permet de tirer avantage des deux procédés.

Algorithme mixte
1- *Partitionner* avec un nombre de classes très élevé
↪ *stabilisation*
2- *CAH* en prenant comme élément à classifier les centres des classes obtenues à l'étape 1.
↪ *choix* d'un *nombre* restreint de *classes* K
3- *Partitionner* toutes les observations en K classes, en initialisant avec les centres des classes de l'étape 2-

5.5 Modèles de mélange

Les modèles de mélange posent un modèle *probabiliste* pour résoudre un problème de clustering. On suppose que :
— les observations $X = (X_1, \ldots, X_n)$ sont indépendantes et issues de K sous-populations (*composantes*) distinctes. La loi dans la sous-population k est $X_i|k \sim \varphi_k(x)$
— les indices (*labels*, allocations) d'appartenance d'un individu à une composante sont inconnus (*latents*, cachés)

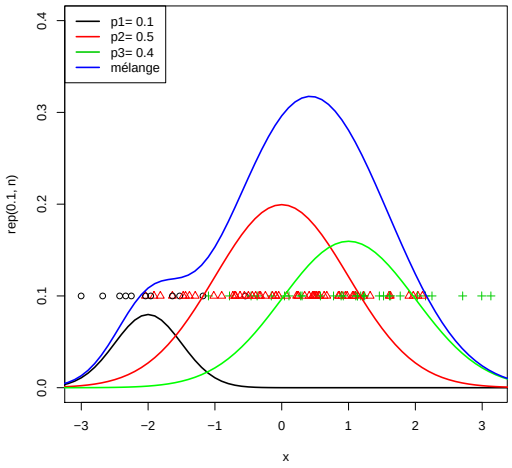


FIGURE 5.4 – Loi de mélange de trois gaussiennes univariées

En général, les composantes appartiennent à une même famille paramétrique et on note $f_k = \varphi(x; \theta_k)$, d'où la densité

$$f(x) = \sum_{k=1}^{K} \pi_k \varphi(x; \theta_k) = \sum_{k=1}^{K} \pi_k f_k(x)$$

Quand φ est la famille des lois gaussiennes, on parle de mélanges gaussiens, comme par exemple sur la figure 5.4, et seuls les paramètres des gaussiennes et les poids du mélange sont à identifier.

Identifiabilité

On rappelle que la famille paramétrique $\mathcal{F} = \{f(x;\alpha), \alpha \in \Omega\}$ indexé par $\alpha \in \Omega$ est identifiable si et seulement si f vue comme fonction de α est injective. Or, dans le cas des mélanges, il suffit d'intervertir l'ordre des classes pour faire tomber la définition : même si la classe $\mathcal{F}$ est identifiable, le mélange ne l'est pas, il y a même $K!$ permutations possible. On étend alors la définition d'identifiabilité, à celle d'identifiabilité à une permutation près.

La plupart des modèles de mélanges finis de lois continues sont identifiables, sauf les mélanges de la loi uniforme. Ce n'est pas le cas pour les lois discrètes : le mélange de lois binomiales ne sont pas identifiables si le nombre d'observations est trop faible par rapport au nombre de variables, ie $n < 2K - 1$.

Un autre problème d'identifiabilité se pose quand on sur-estime le nombre de composantes. On peut en effet écrire, pur $0 \leq \varepsilon \leq \pi_1$,

$$\pi_1 f(x;\alpha_1) + \pi_2 f(x;\alpha_2) = \varepsilon f(x;\alpha_1) + (\pi_1 - \varepsilon) f(x;\alpha_1) + (1-\pi_2) f(x;\alpha_2)$$

et ε n'est pas identifiable dans ce cas.

Estimation

Le paramètre à estimer est $\theta = (\pi, \alpha_1, \ldots, \alpha_K)$, où $\pi = (\pi_1, \ldots, \pi_k)$ représente le *poids* des composantes du mélange et α_k le paramètre de la composante k. L'estimation se fait en maximisant la vraisemblance, ou la log-vraisemblance, sous la contrainte $\sum_k \pi_k = 1$

$$\log L(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \varphi(x_i; \alpha_k) \right)$$

Soit λ un multiplicateur de Lagrange et soit $\mathcal{L}(\theta) = \log L(x_1, \ldots, x_n; \theta) + \lambda(1 - \sum_k \pi_k)$. Alors, on doit résoudre K équations

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_i \frac{\varphi(x_i, \alpha_k)}{f(x_i;\theta)} - \lambda = 0$$

En multipliant chacune d'elle par π_k, puis en les sommant, on obtient

$$\sum_k \sum_i \frac{\pi_k \varphi(x_i, \alpha_k)}{f(x_i;\theta)} = \lambda \sum_k \pi_k = \lambda = \sum_i \sum_k \frac{\pi_k \varphi(x_i, \alpha_k)}{f(x_i;\theta)} = n$$

Notant maintenant

$$t_{ik}(\theta) = \frac{\pi_k \varphi(x_i;\alpha_k))}{\sum_\ell \pi_\ell \varphi(x_i;\alpha_\ell)}$$

l'estimateur du maximum de vraisemblance des poids est $\widehat{\pi}_k = \sum_i t_{ik}(\widehat{\theta})/n$. Sans autre contrainte sur les α_k, on a :

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = \sum_i \frac{\pi_k}{f(x_i;\theta)} \frac{\partial \varphi(x_i;\alpha_k)}{\partial \alpha_k} = \sum_i \frac{\pi_k \varphi(x_i;\alpha_k)}{f(x_i;\theta)} \frac{\partial \log \varphi(x_i;\alpha_k)}{\partial \alpha_k} = \sum_i t_{ik}(\theta) \frac{\partial \log \varphi(x_i;\alpha_k)}{\partial \alpha_k} = 0$$

soit à résoudre

$$\sum_k \sum_i t_{ik}(\widehat{\theta}) \frac{\partial \log \varphi(x_i;\alpha_k)}{\partial \alpha_k}_{|\alpha_k = \widehat{\alpha_k}} = 0$$

On peut ainsi envisager une optimisation alternant le calcul de $t_{ik}(\theta^{(c)})$ pour une valeur de $\theta^{(c)}$ courante, puis l'optimisation de θ considérant la valeur de $t_{ik}(\theta^{(c)})$ précédente connue.

On peut constater la monotonie de cette procédure. Pour pouvoir la montrer, on utilise l'*idée* de modéliser les labels inconnus par une variable aléatoire latente $Z_i \sim \mathcal{M}(1, \pi)$: chaque observation provient conceptuellement d'une des composantes du mélange. Si on pouvait observer cette composante, ou pourrait accéder à la vraisemblance complète

$$p(x, z; \theta) = \prod_i \mathbb{P}(Z_i = k)\varphi(x_i; \alpha_k) = \prod_i \prod_k (\pi_k \varphi(x_i; \alpha_k))^{z_{ik}}$$

où z_{ik} est une variable binaire codant l'appartenance de l'observation i à la classe k. Ainsi

$$\log p(x, z; \theta) = \sum_i \sum_k z_{ik} \left(\log(\pi_k) \log(\varphi(x_i; \alpha_k))\right)$$

et l'estimation du maximum de vraisemblance de θ est maintenant aisée. L'algorithme EM considère les z_{ik} comme des données manquantes, et résout l'optimisation en estimant itérativement deux étapes

1. Expectation Step : gère l'addition des données manquantes non observées en calculant l'espérance de la log-vraisemblance des données complètes (observées et latentes) conditionnellement à la loi des observations sous le paramètre en cours $\theta^{(c)}$,

   $$Q(\theta|\theta^{(c)}) = \mathbb{E}[\log p(x, z; \theta)|x; \theta^{(c)}] = \sum_i \sum_k \mathbb{E}(z_{ik}|x_i; \theta^{(c)}) \left(\log(\pi_k) + \log(\varphi(x_i; \alpha_k)\right)$$

   où $\mathbb{E}(z_{ik|X;\theta^{(c)}}) = \mathbb{P}(Z_i = k|x_i; \theta^{(c)}) = t_{ik}^{(c)}$
2. Maximisation Step : maximise l'expression précédente en θ, d'où, en particulier,

   $$\pi_k^{(c+1)} = \frac{\sum_i t_{ik}^{(c)}}{n}$$

   Si Z était observable, on aurait $\widehat{\pi}_k = \sum_i z_{ik}/n$. $t_{ik}(\widehat{\theta})$ est l'estimation de la probabilité que l'observation i se trouve dans la classe k.

Il est possible de montrer que cet algorithme est monotone. En effet :

$$\begin{aligned}
\log p(x; \theta) &= \log p(x, z; \theta) - \log p(z|x; \theta) \\
&= \mathbb{E}[\log p(x, z; \theta)|x; \theta^{(c)}] - \mathbb{E}[\log p(z|x; \theta)|x; \theta^{(c)}] \\
&= Q(\theta|\theta^{(c)}) - H(\theta|\theta^{(c)})
\end{aligned}$$

Soit $\tilde{\theta} \in \arg\max_\theta Q(\theta|\theta^{(c)})$, alors $\tilde{\theta}$ fait augmenter la vraisemblance :

$$L(\tilde{\theta}) - L(\theta^{(c)}) = Q(\tilde{\theta}|\theta^{(c)}) - Q(\theta|\theta^{(c)}) + H(\theta^{(c)}|\theta^{(c)}) - H(\tilde{\theta}|\theta^{(c)}) \geq 0$$

car $H(\theta^{(c)}|\theta^{(c)}) - H(\tilde{\theta}|\theta^{(c)}) \geq 0$ est la dissemblance de Kullback entre les deux lois. Cette propriété est à la base de l'agorithme EM (Dempster et al. 1977), qui n'est pas spécifique des mélanges, mais peut être plus généralement utilisé pour des structures à données manquantes.

Algorithme EM
Initialiser, puis répéter jusqu'à convergence :
- ***E* Step** : calculer $Q(\theta|\theta^{(c)})$, *espérance* conditionnelle de la log vraisemblance complète
  ↪ calculer $p(z|x; \theta^{(c)})$ ou ses moments
- ***M* Step** : mettre à jour θ par *maximization* : $\theta^{(c+1)} = \arg\max_\theta Q(\theta|\theta^{(c)})$

L'algorithme EM converge vers un maximum local. Il est donc en général nécessaire de le relancer plusieurs fois avec différentes initialisations, puis de prendre la solution de plus grande vraisemblance.

EM gaussien

Dans le cas des mélanges gaussiens univariés, les étapes de l'algorithme EM s'écrivent de la façon suivante :

— ***E* Step** :

$$t_{ik} := \mathbb{P}(Z_i = k|X) = \frac{\pi_k^{(c)}\varphi(x_i;\theta_k^{(c)})}{\sum_\ell \pi_\ell^{(c)}\varphi(x_i;\theta_\ell^{(c)})}$$

— ***M* Step** :

$$\mu_k^{(c+1)} = \frac{\sum_{i=1}^n t_{ik}x_i}{\sum_{i=1}^n t_{ik}}$$

$$\sigma_k^{2(c+1)} = \frac{\sum_{i=1}^n t_{ik}(x_i - \mu_k^{(c+1)})^2}{\sum_{i=1}^n t_{ik}}$$

$$\pi_k^{(c)} = \frac{\sum_{i=1}^n t_{ik}}{n}$$

Après convergence, chaque observation i est affectée à une composante par la règle du Maximum A Posteriori (MAP)

$$\widehat{z}_i = \arg\max_{k=1,\dots,K} t_{ik}$$

Deux exemples de classification par modèle de mélange gaussien sont représentés sur les figures 5.5.

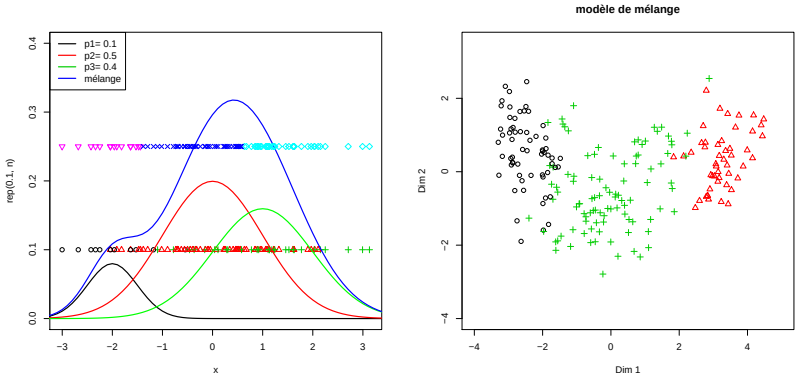


FIGURE 5.5 – L'allocation des observations suivant la règle du MAP : à gauche, dans le cadre d'un mélange univarié, les affectations sont matérialisées par des points alignés sur la ligne du haut ; à droite, dans le cas d'un mélange multivarié appliqué aux données seeds, les groupes sont représentés par différentes couleurs sur le premier plan principal.

Lien entre modèle de mélange Gaussien et Kmeans L'étape E de l'EM est une étape de classification floue : elle relaxe la recherche des $z_{ik} \in \{0,1\}$ à celle des $t_{ik} \in [0,1]$. Si, à chaque étape E, on affecte chaque observation i au cluster maximisant t_{ik}, on effectue une étape supplémentaire de classification. Ces classes étant connues, l'étape M revient à une étape de maximisation de la vraisemblance complète. Cet algorithme, variante de l'EM s'appelle CEM (pour *Classification EM*). Quand le modèle de mélange l'est sur des lois gaussiennes de même variance proportionnelle à l'identité, l'algorithme CEM donne le même résultat qu'un Kmeans. On voit ici l'explication statistique donnée à un problème posé au départ de façon déterministe.

Logiciels Plusieurs packages de R permettent d'estimer des modèles de mélange. On peut citer `Rmixmod` [2] ou `mclust` [3]. Le package `MixSim` [4] permet de simuler des données issues d'un modèle de mélange gaussien suivant différents degrés de séparation.

5.6 Evaluation d'une méthode de clustering

Alors, quelle méthode choisir ? Rappelons-l : comme le problème est non supervisé et donc mal posé, cette question est délicate.

On peut faire une évaluation technique à l'aide de critères essayant de définir la qualité d'une partition. Mais il faut aussi voir si le clustering obtenu est utile à l'utilisateur. La méthode à choisir dépendra de ce que l'on cherche, de la forme que l'on veut donner aux clusters, de la taille des données, etc. Une comparaison illustrative est faite dans la documentation du package scikit-learn de Python, reproduite sur la figure 5.6 :

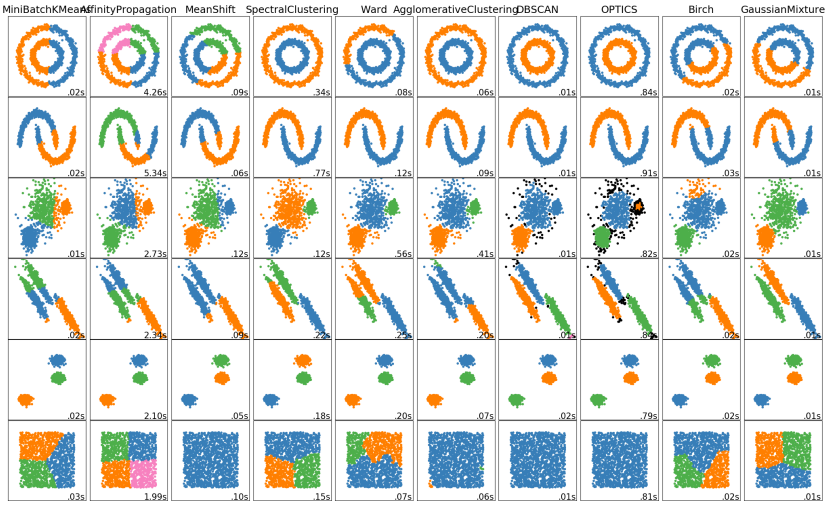


FIGURE 5.6 – Comparaison de différentes méthodes de clustering, package scikit-learn de Python, `https://scikit-learn.org/stable/modules/clustering.html`

Voici quelques critères techniques.

2. `https://CRAN.R-project.org/package=Rmixmod`
3. `https://CRAN.R-project.org/package=mclust`
4. `https://CRAN.R-project.org/package=MixSim`

5.6.1 Comparaison de partitions

Erreur empirique Soit z, $\widehat{z}$ les matrices de classification de deux partitions différentes. Il faut tenir compte du fait qu'en non supervisé, le numéro d'un cluster n'a pas d'importance en lui-même. C'est la répartition des données les unes par rapport aux autres qui en a. Il faut donc tenir compte de cette indépendance à la relabellisation des clusters.

Définition 10. *Soit $\mathcal{T}$ l'ensemble de partitions τ sur $\{1, \ldots, K\}$. Le taux d'erreur empirique entre les deux matrices de classification z et $\widehat{z}$ est défini par*

$$err(z, \widehat{z}) = \frac{1}{n} \min_{\tau \in \mathcal{T}} \sum_{i=1}^{n} \mathbb{1}_{z_i = \tau(\widehat{z}_i)}$$

Les partitions sont proches quand l'erreur est petite. Ce critère peut être pénible à calculer quand il y a de nombreux clusters dans la partition ; et il n'est pas adapté pour la comparaison de partitions ayant un nombre de clusters différent.

Indice de Rand (1971) Plutôt que de comparer les classements individu par individu, on mesure un accord/désaccord entre des paires d'individus, ce qui rend la définition indépendante de la numérotation des clusters et permet de ne pas limiter la comparaison aux partitions ayant le même nombre de clusters.

Définition 11. *On définit deux mesures d'accord (A et B) et deux mesures de désaccord (C et D) :*

- *A : le nombre de paires où les deux individus appartiennent à un même cluster dans la partition z et à un même cluster dans la partition $\widehat{z}$*
- *B : le nombre de paires où les deux individus appartiennent à deux clusters différents dans la partition z et à deux clusters différents dans la partition $\widehat{z}$*
- *C : le nombre de paires où les deux individus appartiennent à un même cluster dans la partition z et à deux clusters différents dans la partition $\widehat{z}$*
- *C : le nombre de paires où les deux individus appartiennent à deux clusters différents dans la partition z et à un même cluster dans la partition $\widehat{z}$*

Alors, l'indice de Rand est défini par

$$\text{rand}(z, \widehat{z}) = \frac{A + B}{A + B + C + D} \in [0; 1].$$

Autrement dit, l'indice de Rand est la proportion de paires en accord sur le nombre total de paires. Les partitions sont d'autant plus proches que l'indice de Rand est proche de 1.

Hubert et Arabie (1985) en ont proposé une version normalisée, l'ARI (Adjusted Rand Index), en particulier quand les partitions n'ont pas le même nombre de clusters.

Stabilité du résultat On considère que $\widehat{z}$ est l'estimation d'une hypothétique partition z, et on essaye d'estimer la variabilité de la moyenne de chaque classe par bootstrap. On tire ainsi B échantillons bootstrap $\mathbf{x}^b = (x_1^b, \ldots, x_n^b)$. Pour chaque tirage, on calcule la partition $\widehat{z}^b$ et le centre des classes $\mu^b = (\mu_1^b, \ldots, \mu_K^b)$, ce qui permet de déduire un intervalle de confiance du centre de chaque classe. Le point délicat dans cette procédure est le problème de relabellisation des classes dans chaque échantillon bootstrap.

Plots alluviaux Des outils graphiques permettent de comprendre visuellement les différences entre deux partitions : ce sont les plots alluviaux (alluvial plots). Voir par exemple une implémentation dans R [5]

5.6.2 Choix du nombre de clusters

Comment choisir le nombre de clusters ? Cette question n'est pas évidente : si on utilise la méthode Kmeans par exemple, choisir la partition réalisant le minimum de la variance intra classe amène à choisir autant de classes que d'individus (s'ils ont tous différents).

— détection d'un saut important du critère : *règle du coude* appliquée à la courbe des critères $W(k)$ successifs (Hartigan (1975)) :

$$\widehat{K} = \min_k \left\{ \frac{W(k)}{W(k-1)}(n-K-1) \geq 10 \right\}$$

— *statistique Gap* (Tibshirani et al., 2001) : on estime par méthode de Monte Carlo l'espérance du critère calculé à partir d'un échantillon de loi uniforme $\mathbb{E}_n^*(\log W(k))$

$$\text{Gap}(k) = \arg\max_k \frac{1}{B}\sum_{b=1}^{B} \ln W^{(b)}(k) - \ln W(k)$$

et on choisit un nombre de clusters tel que

$$\widehat{K} = \arg\min_k \{k : \text{Gap}(k) \geq \text{Gap}(k+1) - \text{standard deviation}(\{\ln W^{(b)}(k)\}_{b=1}^{B})\}$$

— une statistique basée sur le rapport entre l'inertie inter classe et l'inertie intra classe, quand le critère est l'inertie intra classe $W(k) = I_{intra}(k)$

$$\widehat{K} = \arg\max_k \frac{I_{inter}(k)/(k-1)}{W(k)/(n-k)}$$

— indice de *silhouette* (Rousseeuw, 1987) qui essaie de modéliser compacité et répartition des classes. Pour chaque point, son coefficient de silhouette est défini comme la différence entre la distance moyenne avec les points du même groupe (cohésion) et la distance moyenne avec les points des autres groupes voisins (séparation). Si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien, indiquant une mauvaise classification. En revanche, si la différence est positive, le point est en moyenne plus proche de son groupe que du groupe voisin, indiquant une bonne classification. Plus formellement, la silhouette du point x_i est définie par

$$\text{sil}(x_i) = \frac{b_i - a_i}{\max(a_i, b_i)}, \text{ où } a_i = \frac{1}{|C_k|-1}\sum_{x_\ell \in C_k} d(x_i, x_\ell), \;\; b_i = \min_{k' \neq k} \frac{1}{|C_{k'}|}\sum_{x_\ell \in C_{k'}} d(x_i, x_\ell)$$

et on choisit un nombre de clusters maximisant la silhouette moyenne :

$$\widehat{K} = \arg\max_k \frac{1}{n}\sum_i \text{sil}(x_i)$$

Dans le cas des modèles de mélange, l'hypothèse probabiliste permet de définir un critère BIC de choix du nombre de composantes et d'en montrer des propriétés de consistance (Keribin, 2000) sous certaines hypothèses.

Quoiqu'il en soit, il faut choisir un nombre *utile* de clusters.

5. `https://cran.r-project.org/web/packages/ggalluvial/vignettes/ggalluvial.html`

Deuxième partie

Apprentissage supervisé

Chapitre 6

Apprentissage supervisé

Nous nous plaçons ici dans un cadre d'explication ou de prédiction d'une variable aléatoire *réponse* Y en fonction d'une liste de variables *explicatives* $\mathbf{X} = (X_1, \ldots, X_p)$ observées sur des individus ou des objets. C'est le cas de très nombreuses situations, par exemple :

- On observe la taille d'alevins en fonction de leur âge. Quelle est la taille moyenne d'un alevin à un âge donné ? Quelle taille aura un alevin pris au hasard à un âge donné ?
- On dispose de la vitesse du vent, de la pluviométrie, d'une image satellite de Paris. Quelle sera la concentration d'ozone demain sur Paris ?
- Est-il possible de classer automatiquement un courriel (normal, indésirable) en fonction de certains mots ou ponctuations dans le corps du message ?
- Un organisme de crédit dispose de son portefeuille de clients (CSP, soldes, découverts, incidents). Peut-il accepter un nouveau crédit pour un client existant sans risque d'incident ? et pour un nouveau client ?
- Comment choisir les destinataires d'un mailing pour maximiser le taux de réponses positives, tout en minimisant le nombre d'envois ?
- La consommation d'alcool est-elle un facteur de risque pour le cancer du foie ?

Quelle est l'information utile ?

- Les événements observés fournissent une base de connaissance pour expliquer un phénomène ou prédire de nouvelles configurations
- Il s'agit de tirer partie des *corrélations* entre les phénomènes pour en prévoir d'autres.

De façon générale, le phénomène liant $\mathbf{X}$ à Y est inconnu, et l'objectif est d'essayer de l'induire en définissant une règle à partir des données observées. Définir cette règle, appelée *prédicteur*, c'est construire une fonction des variables explicatives qui représente au mieux la réalité observée sur un échantillon $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ (i.i.d. $\sim \mathbf{P}$) (*ajustement*), tout en étant capable de produire de bonnes prédictions sur un nouveau jeu de données de caractéristiques similaires (*généralisation*). Ce cadre est celui de l'apprentissage supervisé : *apprentissage* fait référence à un processus automatisable sur un ensemble de données pour en extraire de la connaissance ; *supervisé* indique ce processus possède des observations qui servent de référence pour construire les règles.

Le prédicteur peut être choisi à partir d'un modèle statistique, pour le modèle linéaire à erreur additive par exemple. La régression linéaire peut être étendue à des fonctions non linéaires des variables explicatives et à des cas non paramétriques où le modèle ne s'exprime pas avec un nombre fini de paramètre. L'apprentissage supervisé s'adapte aussi aux réponses binaires, catégorielles (on parlera alors de *classification supervisée*) ou aux spécificités des données de

comptage. La fonction de perte ou de coût (*loss*) définit le critère à optimiser pour trouver le prédicteur dans le modèle choisi : c'est par exemple le risque quadratique dans le cas de la régression.

Le but de l'apprentissage supervisé, c'est apprendre une règle pour construire un prédicteur (classifieur pour Y qualitatives, régresseur pour Y quantitative) $\widehat{h} \in \mathcal{F}$ à partir des données d'apprentissage $\mathcal{D}_n$ avec un coût $\mathcal{R}(\widehat{h})$ petit en moyenne ou avec grande probabilité par rapport à $\mathcal{D}_n$.

Le point de vue de ce cours est de fonder mathématiquement les modèles et méthodes, se concentrant sur les méthodes classiques d'apprentissage via la régression non linéaire (chapitre 7), la régression logistique et l'analyse discriminante (chapitre 8), la problématique spécifique de prédiction d'une variable qualitative (chapitre 9). Ces chapitres supposeront en grande partie que l'estimation se fait sous le modèle "génératif" ("vrai modèle"). Les problèmes de sur et sous-ajustement, de choix de modèles, de calcul des performances sont abordés au chapitre 10. Le dernier chapitre 11 aborde les méthodes de régularisation permettant une synthèse des chapitres précédents et une bonne introduction à des méthodes plus focalisées sur la capacité prédictive des modèles que leur capacité explicative.

Le présent chapitre introduit la notion de régression, présente les étapes de la démarche statistique et la problématique du choix de modèle, qui seront ensuite détaillés dans les chapitres suivants.

Bibliographie Azais et Bardet (2005), Pagès (2005), Rivoirard et Stoltz (2009), James et al. (2013).

6.1 Fonction de régression

Le principe des modèles de régression est d'approcher Y par une fonction $m(.)$ de la variable X, puis d'utiliser m à des fins de description, d'évaluation des contributions des différentes variables explicatives, ou de prédiction de la variable réponse à de nouvelles valeurs des variables explicatives. Il s'agit, par exemple, de prédire la concentration en polluant dans une zone donnée en fonction d'indices météorologiques, de décider si un message électronique est un courrier indésirable en fonction de la présence de mots ou caractères particuliers, ou de faire émerger des facteurs explicatifs aux défauts de circuits imprimés suivant les conditions de température et pression du processus de fabrication.

Comment définir la fonction m ? Soit h une fonction candidate. Son risque (quadratique) $R(h)$ associé à l'utilisation de $h(X)$ à la place de Y est appelé erreur quadratique (moyenne) d'estimation, c'est l'espérance du carré de l'erreur commise par l'approximation :

$$R(h) = \mathbb{E}[(Y - h(X))^2] = \int (y - h(x))^2 dP(x, y). \tag{6.1}$$

Théorème 4. *La fonction qui minimise le risque quadratique $R(.)$ est $m(X) = \mathbb{E}(Y|X)$, l'espérance de Y conditionnement à X*

$$\forall x, m(x) = \mathbb{E}(Y|X = x).$$

Elle est appelée ***fonction de régression****.*

Preuve. Soit $m(x) = \mathbb{E}(Y|X = x)$ Calculons $Y - h(X) = Y - m(X) + m(X) - h(X)$. Alors

$$\mathbb{E}(Y - h(X))^2 = \mathbb{E}(Y - m(X))^2 + \mathbb{E}(m(X) - h(X))^2 + 2\,\mathbb{E}[(Y - m(X))(m - h)(X)]$$

La dernière expression étant nulle grâce au théorème de l'espérance totale. D'où

$$R(h) = R(m) + \mathbb{E}(m(X) - h(X))^2 \geq R(m)$$

est minimum pour $h = m$. ◊

L'espérance conditionnelle $m(X)$ est la **meilleure approximation** (au sens du risque quadratique) de Y par une fonction de X. C'est la **projection orthogonale** de Y sur le sous-espace de Hilbert $\mathbb{L}^2(X)$ au sens du produit scalaire $< U, V >= \mathbb{E}(UV)$. On a donc :

$$\mathbb{E}((Y - \mathbb{E}(Y|X_1, \ldots, X_p))\, U) = 0, \ \ \forall\, U \in \mathbb{L}^2(X_1, \ldots, X_p).$$

Si X est déterministe, on travaille également à $X = x$ fixé et $m(x) = \mathbb{E}(Y|X = x)$ est à nouveau la meilleure approximation. Ainsi, les variables explicatives peuvent toujours être considérées comme constantes pour un individu ou un objet donné, qu'elles soient aléatoires (et l'analyse est conditionnelle aux valeurs observées), ou qu'elles soient déterministes et fixées a priori par le plan d'expérience. Réservant la notation majuscule aux variables aléatoires, nous noterons désormais les variables explicatives avec des lettres minuscules ; soit pour l'individu i, le vecteur ligne des valeurs des variables explicatives observées est $\mathrm{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ et nous ne ferons plus référence au conditionnement dans la suite.

Remarque D'autres choix de fonction de perte pourraient être possibles, certes plus difficiles à calculer, mais peut-être moins pénalisantes pour les grandes erreurs. Par exemple :
- *la perte en valeur absolue* : $l(Y, f(X)) = |Y - f(X)|$
- *la perte quadratique tronquée* : $l(Y, f(X)) = min\{(Y - f(X))^2, d^2\}$
- *la perte en IC* : $l(Y, f(X)) = 0$ si $|Y - f(X)| \leq d$; sinon, $l(Y, f(X)) = 1$: toutes les erreurs de l'intervalle sont tolérables, et celles en dehors des limites intolérables.

Le choix de la perte quadratique est historique, numérique, et la norme sous-jacente est hilbertienne.

Enfin, mentionnons la richesse du vocabulaire associé à la dénomination des variables :
- La variable *réponse* Y est parfois appelée endogène, variable à expliquer, dépendante.
- Les variables *explicatives* $X = (X_1, \ldots, X_p)$ sont parfois appelées exogènes, régresseurs, covariables, "indépendantes", facteurs (pour des variables qualitatives), *features* (en *machine learning*).

6.1.1 Détermination de $\mathbb{E}(Y|X_1, \ldots, X_p)$

Si la loi jointe de (Y, X) est connue, la calcul de la loi conditionnelle $\mathbb{P}(Y|X = x)$ est très simple. Dans le cas où $(Y, X_1, \ldots, X_p)$ est un vecteur discret, et pour $\mathbb{P}(X = x) \neq 0$,

$$\mathbb{E}(Y|X = x) = \sum_y y\ \mathbb{P}(Y = y|X = x), \text{ où } \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)}.$$

La loi de probabilité de Y conditionnellement à $\{X = x\}$ est la fonction $y \to \frac{\mathbb{P}(Y=y, X=x)}{\mathbb{P}(X=x)}$. Si $Y, X_1, \ldots, X_p$ admet une densité $f(x, y)$ dans $\mathbb{R}^{p+1}$, la densité conditionnelle $f_{Y|X=x}$ est définie par

$$f_{Y|X=x}(y) = f(x, y)/f_X(x) \text{ si } f_X(x) \neq 0,$$
$$f_{Y|X=x}(y) = 0 \text{ sinon.}$$

L'espérance de Y conditionnelle à $\{X = x\}$ est définie par $\mathbb{E}(Y|X = x) = \int y f_{Y|X=x}(y)dy$. Maintenant, l'application $e : x \mapsto e(x) = \mathbb{E}(Y|X = x)$ est une fonction réelle de la variable réelle. La fonction composée $e \circ X$ est une variable aléatoire réelle, notée $\mathbb{E}(Y|X)$, espérance conditionnelle de Y en X.

Remarque On a les relations suivantes :
— théorème de l'espérance totale

$$\mathbb{E}(Y) = \mathbb{E}_X[\mathbb{E}(Y|X)] \tag{6.2}$$

— théorème de la variance totale

$$\text{var}(Y) = \mathbb{E}_X[\text{var}(Y|X)] + \text{var}_X[\mathbb{E}(Y|X)] \tag{6.3}$$

Exemple Si $Y, X_1, \ldots, X_p$ est un vecteur gaussien, alors $\mathbb{E}(Y|X_1, \ldots, X_p)$ est la projection orthogonale de Y sur $V = vect\{1, X_1, \ldots, X_p\}$ et le meilleur prédicteur est une combinaison linéaire des X_j, soit

$$Y = \mathbb{E}(Y|X) + \varepsilon = a_0 + \sum_{j=1}^{p} a_j X_j + \varepsilon$$

avec ε gaussienne centrée indépendante des X_j.
Les variables gaussiennes sont de carré intégrable et appartiennent à $L^2(\mathbb{P})$. Soit $\Pi_V Y$ la projection orthogonale de Y sur V. Par définition de la projection orthogonale $< Y - \Pi_V Y, X_j >= 0$ pour tout j. Or, $< Y - \Pi_V Y, X_j >= \mathbb{E}(Y - \Pi_V Y|X_j)$, donc $Y - \Pi_V Y$ est décorrélé des X_j, et donc indépendant de V puisque le modèle est gaussien. Donc, pour toute fonction φ mesurable, par indépendance, $\mathbb{E}((Y - \Pi_V Y)\varphi(X)) = \mathbb{E}(Y - \Pi_V Y)\mathbb{E}(\varphi(X)) = 0$. Ainsi, $\Pi_V Y = \mathbb{E}(Y|X)$ et le résultat s'en suit.

Si la forme de la loi jointe n'est pas connue, il faut poser des hypothèses de **modélisation** sur la loi de $Y|X = x$. Le modélisateur pourra en particulier tenir compte d'une connaissance spécifique de la loi de la réponse, du comportement de l'espérance et de la variance de celle-ci. Une fois le modèle statistique choisi, il s'agira de l'utiliser pour inférer les paramètres inconnus. C'est ce qui est discuté dans les sections suivantes.

6.1.2 Hypothèses sur le lien entre Y et X

L'hypothèse que la loi conditionnelle $Y|X = x$ ne dépend de $X = x$ que par son espérance conditionnelle $m(x) = \mathbb{E}(Y|X = x)$, amène au modèle à bruit additif très simple

$$Y_{|X=x} = m(x) + \varepsilon$$

où
— l'espérance du bruit nulle : $\mathbb{E}(\varepsilon|X = x) = 0$,
— la variance du bruit indépendant de x : $\mathbb{E}(\varepsilon^2|X = x) = \sigma^2$,
Il reste bien sûr à déterminer la forme de $m(x)$, cf section 6.1.3.

La loi conditionnelle $Y|X = x$ peut également dépendre de $X = x$ par sa variance conditionnelle $\sigma^2(x) = \text{var}(Y|X = x)$

$$Y_{|X=x} = m(x) + \sigma(x)\varepsilon$$

Quand la loi conditionnelle appartient à une famille exponentielle (Bernoulli, Poisson, Binomiale, exponentielle, ...), la variance dépend de l'espérance $\sigma^2(x) = \sigma^2(m(x))$. La loi conditionelle n'est alors plus exprimée sous forme d'un modèle à bruit additif.

Enfin, le modélisateur doit réfléchir au choix du type de loi. Celui-ci pourra avoir une influence sur la méthode d'estimation des paramètres inconnus : dans les modèles additifs ne faisant pas d'a priori sur le type de loi, il est naturel d'utiliser la **méthode des moindres carrés**. Lorsque le choix d'une loi est fait pour modéliser loi conditionnelle $Y|X = x$, la **méthode du maximum de vraisemblance** est la méthode d'estimation naturelle.

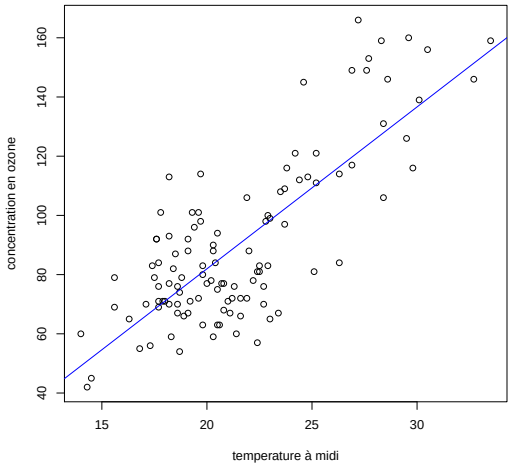
6.1.3 Modélisation de l'espérance

Elle est guidée par la théorie ou par l'observation de l'allure du phénomène, et peut prendre des formes très diverses. La fonction de régression $m = m(x_1, \ldots, x_p)$ peut être :

— linéaire en les paramètres

$$m(x) = \theta_0 + \theta_1 x_1$$
$$m(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$$
$$m(x) = \theta_0 + \sum_{j=1}^{p} \theta_j x_j$$

FIGURE 6.1 – Régression linéaire : Concentration maximale journalière d'ozone (en μg/m^3) en fonction de la température à midi dans la région de Rennes Cornillon et Matzner-Løber (2007)



— linéaire (après changement de variables). Le modèle de Cobb et Douglas (1928) par exemple explique la production en fonction du capital (valeur des usines) et du travail fourni (basé sur le calcul du nombre total de travailleurs)

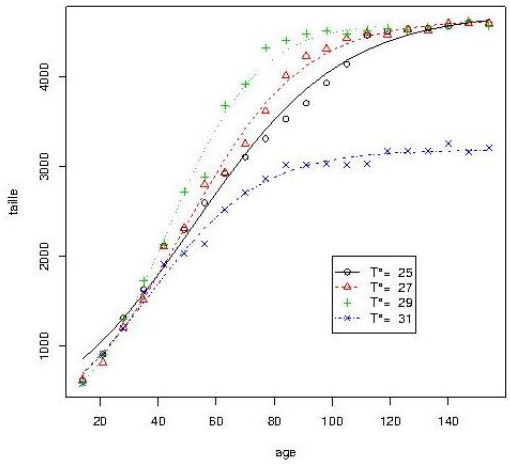
$$m(K, T) = \alpha_1 K^{\alpha_2} T^{\alpha_3}.$$

Les régresseurs sont ici $x_i = (1, \log(K_i), \log(T_i))$ et la réponse est $y_i = \log(P_i)$. Mais attention à la modélisation de l'erreur ! Elle est multiplicative dans la première formulation, et additive dans l'autre. Si on considère le modèle additif en $\log(Y)$, avec erreur gaussienne de variance s^2, alors l'erreur est multiplicative dans le modèle en Y, de loi log-normale de variance $e^{s^2}(e^{s^2} - 1)$.

— non linéaire, par exemple pour les courbes de croissance (Figure 6.2). Il est illusoire de vouloir proposer un modèle théorique fin de la croissance d'un organisme vivant. Pourtant, on observe en général que cette croissance présente une grande régularité, dépendant fortement de peu de facteurs. Les points expérimentaux, une fois portés sur le graphique, se répartissent suivant des formes simples, par exemple une fonction monotone croissante, avec deux asymptotes et un point d'inflexion comme sur la Figure 6.2 :

$$m(x) = \frac{k}{1 + \frac{k-n_0}{n_0} e^{-rx}},$$

FIGURE 6.2 – Régression non linéaire : Croissance des alevins en fonction de la température de l'eau



— linéaire dans une nouvelle base de fonctions

$$m(x) = \sum_{k=1}^{K} \theta_k \phi_k(x)$$

où les ϕ_k forment une suite de fonctions convenables (polynomiales, Fourrier, ondelettes) de $\mathbb{R}^p$ dans $\mathbb{R}$. Le modèle est linéaire dans cette nouvelle base mais plus en X. Le nombre K de fonctions de base indexe la dimension du modèle indépendamment de la taille n de l'échantillon. Ce paramètre de lissage devra être choisi au vu des données.

— additifs : ces modèles font l'hypothèse

$$m(X) = \sum_{j=1}^{p} f_j(X_j)$$

où les f_j sont des fonctions vérifiant un certain degré de régularité et sont estimées par exemple par des méthodes de lissage et ou décomposition sur des bases de fonctions. L'hypothèse d'additivité permet de réduire la dimension du problème : au lieu d'estimer une fonction p-dimensionnelle, on estime des fonctions à variables réelles en imposant une contrainte de structure.

La paramétrisation répond à des objectifs mathématiques (facilité de calcul,...) ou de modélisation (les paramètres sont des grandeurs interprétables). Il n'est pas conseillé, sous prétexte de décrire finement l'allure du phénomène observé, de choisir une fonction de régression avec un trop grand nombre de paramètres : il faut se limiter à des paramètres correspondant à des aspects essentiels du phénomène de l'étude en cours. Si le nombre de paramètres est trop important (*sur-paramétrisation* du modèle), la courbe ajustée sera trop proche de l'ensemble des points expérimentaux, et les paramètres seront estimés avec une mauvaise précision.

6.1.4 Choix de la loi $Y|X = x$

Régression gaussienne

C'est un cas d'un modèle à bruit additif pour lequel la loi du bruit est gaussienne, soit :

$$Y_{|x} \sim \mathcal{N}(m(x), \sigma^2).$$

La régression peut être linéaire ou non linéaire en fonction du choix de la fonction de régression $m(x)$. En général, la variance σ^2 est constante (modèle homoscédastique) et les observations indépendantes. La régression linéaire gaussienne est parfois appelée **modèle linéaire gaussien**, quoi que cette dernière dénomination fasse parfois référence au cas plus large d'hétéroscédasticité des données (variance non constante), voire de corrélation du bruit.

Autres lois

Le choix d'une loi dépend du phénomène observé. Par exemple, si l'observation est un comptage (nombre de débits mensuels ayant dépassé un certain seuil par exemple, nombre d'appels arrivant à un standard téléphonique, ...), la loi naturelle est la loi de Poisson.

Si l'observation est binaire, la loi naturelle à utiliser est la loi de Bernoulli. On parlera de **classification** (supervisée) lorsqu'il s'agit non seulement de déterminer les paramètres qui expliquent le modèle, mais aussi de définir une règle d'attribution de chaque individu à l'une des deux classes. Dans le cas du scoring de crédit, la banque étudie la façon dont elle attribue un crédit à des clients (réponse 0 ou 1). Une modélisation possible est : $Y_i \sim \mathcal{B}(p_i)$, avec $p_i = g(\sum_{j=0}^{p} \theta_j x_{ji})$. Les covariables sont par exemple l'âge, la catégorie socio-professionnelle, la date d'achat...

Quand la loi choisie fait partie d'une famille exponentielle (binomiale, Poisson, multinomiale, exponentielle), la variance n'est nativement pas constante et dépend de l'espérance, donc des covariables.

Méthodes non paramétriques

Notons qu'il existe des méthodes supervisées en régression Hastie et al. (2001), que nous n'aborderons pas dans ce cours :

— la méthode des K plus proches voisins définit $\widehat{m}_K(x) = \text{moyenne}\{y_i | x_i \in N_K(x)\}$, où $N_K(x)$ est le voisinage de l'ensemble d'apprentissage contenant les K plus proches voisins de x. On obtient une approximation localement constante. Si la dimension de l'espace des covariables est grand, les plus proches voisins peuvent être très dispersés, ce qui impliquera une grande erreur de prédiction. Imposer une structure, si elle existe, permet de réduire le biais et la variance des estimateurs.

— méthodes à noyaux

$$\widehat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

où K_λ est le noyau : on assigne des poids qui s'amenuisent avec la distance au point cible
— arbres de décision : Classification And Regression Tree, basé sur un emboîtement de règles binaires.

6.2 Qu'est-ce qu'un bon modèle ?

Après avoir défini la forme générale du modèle, il faut l'estimer (en estimer les paramètres dans le cas paramétrique), avant de pouvoir l'utiliser. A première vue, un bon modèle est un modèle qui s'**ajuste** bien aux données.

Cependant, en augmentant le nombre de paramètres du modèle, on peut rendre l'ajustement parfait (si nombre de paramètres = nombre d'observations) : dans ce cas, le modèle proposé n'a pas permis de réduire la dimension du modèle. Il faut donc rechercher un **modèle parcimonieux**, qui permette de synthétiser l'information et qui offre de **meilleures prédictions** qu'un modèle contenant de nombreux paramètres non significatifs.

Une importante propriété d'un modèle, c'est sa **robustesse** vis à vis de l'échantillonnage des données : on peut souhaiter que les prédictions issues d'un modèle ne soient pas trop dépendantes de l'échantillon d'apprentissage. Or, l'erreur de prédiction d'un nouveau Y_{x_0} en x_0 s'écrit :

$$\begin{aligned}\mathbb{E}[(Y_{x_0} - \widehat{m}(x_0))^2 | X = x_0] &= \mathbb{E}[(Y_{x_0} - \mu_0 + \mu_0 - \widehat{m}(x_0))^2 | X = x_0] \\ &= \mathbb{E}[(Y_{x_0} - \mu_0)^2 | X = x_0] + \mathbb{E}(\mu_0 - \widehat{m}(x_0))^2 \\ &= \sigma^2 + (\text{biais } \widehat{m}(x_0))^2 + \text{var}(\widehat{m}(x_0)),\end{aligned}$$

où le terme $\mathbb{E}(Y_{x_0} - \mu_0)(\mu_0 - \widehat{m}(x_0)) = 0$ car Y_{x_0} et $\widehat{m}(x_0)$ sont indépendants. Lorsque le nombre de paramètres (complexité du modèle) augmente, le terme de biais diminue, mais celui de variance augmente : intuitivement, on comprend qu'il faut un compromis entre parcimonie et robustesse, entre biais et variance.

Nous verrons qu'une difficulté supplémentaire tient au fait que l'erreur moyenne de prédiction fait intervenir le paramètre inconnu, et n'est donc pas calculable directement sur les données. Nous verrons les stratégies qui peuvent être mises en œuvre pour pallier cette difficulté.

La modélisation reste, au moins en partie, un art, mais il existe des principes pour la guider. McCullagh et Nelder (1989) proposent trois principes généraux :

1. Si tous les modèles sont faux, certains sont plus utiles que d'autres (maxime attribuée à George Box).
2. Il n'y a pas un unique bon modèle en général, mais plusieurs.
3. Un modèle doit toujours être validé.

Ajoutons qu'un modèle a une utilité en fonction de l'objectif fixé (décrire, évaluer, prédire...), et un modèle n'est pas forcément bon pour tous les objectifs.

6.3 Les étapes de la démarche statistique

A partir d'un échantillon de n observations, il s'agit de déduire -ou d'inférer- certaines propriétés du modèle probabiliste inconnu qui les a générées, puis d'utiliser ces estimations à des fins de décision ou de prédiction. C'est la démarche générale de l'inférence statistique qui peut être résumée par les étapes suivantes :

1. Acquérir et préparer les données, prendre en compte leur nature, effectuer une analyse descriptive et exploratoire.

2. Définir un modèle adapté à la situation observée :
   — loi de probabilité de la variable réponse,
   — équation liant l'espérance de la réponse et les covariables.
3. Estimer les paramètres du modèle grâce aux observations.
4. Vérifier l'adéquation de l'estimation aux observations (diagnostics d'ajustement, analyse des résidus, tests de loi, indépendance).
5. Faire un choix entre plusieurs modélisations
6. Vérifier la capacité de généralisation du modèle choisi, c'est à dire le bon comportement du modèle sur des données non encore observées.
7. Utiliser le modèle à des fins de décision ou de prédiction.

Ces étapes ne s'enchaînent pas forcément de façon linéaire. En effet, le processus est itératif en fonction des résultats obtenus à chacune d'elle. Plusieurs modèles pourront être mis en compétition, que des procédures de choix de modèle permettront de départager.

Chapitre 7

Régression non linéaire

La régression linéaire a été étudiée dans le module STA201, cf annexe B pour des révisions : définition, estimateurs, et leur loi à distance finie dans le cas gaussien, i.e. pour un échantillon de taille finie. Mais nous n'avons pas répondu à la question de la consistance de l'estimateur, ni précisé sa loi dans un cadre non gaussien. L'étude de l'asymptotique, un des domaines de la statistique mathématique, va permettre de répondre à ces questions. Nous étudierons l'asymptotique des estimateurs en nous plaçant dans le cadre plus général de la régression paramétrique non linéaire et non (forcément) gaussienne. Nous aborderons les estimateurs des moindres carrés et du maximum de vraisemblance, deux exemples d'estimateurs du minimum de contraste.

Bibliographie : Da Cunha et Duflo (1983), Van der Vaart (1998), Azais et Bardet (2005), Huet et al. (2003), Myers et al. (2012), Gallant et Goebel (1976), Jennrich (1969).

Notation : nous noterons $||.||_n$ la norme euclidienne de $\mathbb{R}^n$ et $< ., . >_n$ son produit scalaire.

7.1 Cadre de l'étude

Le modèle de régression non linéaire étend celui de la régression linéaire.

Définition 12. *Soit* $(\mathrm{x}_i, Y_i)_{i=1,\dots,n}$ *un échantillon où* Y_i *est une variable aléatoire réponse et* x_i *le vecteur ligne de* p *variables explicatives. La régression non linéaire paramétrique postule un modèle à* ***bruit additif*** *tel que*

$$Y_i = f(\mathrm{x}_i, \theta) + \varepsilon_i, \tag{7.1}$$

avec les hypothèses suivantes :

(i) la fonction de régression f *est une fonction* ***continue*** *à valeurs réelles sur un* ***compact*** Θ *de* $\mathbb{R}^p$ *:* $f \in \{f(., \theta), \theta \in \Theta\}$*,*

(ii) les ε_i *sont des variables aléatoires* ***i.i.d.****,* ***centrées*** $\mathbb{E}(\varepsilon_i|\mathrm{x}_i) = 0$*, de variance finie* σ^2 *ne dépendant pas de* x_i*.*

La forme de la fonction f est donnée, mais la valeur du paramètre avec laquelle les observations sont supposées être générées est inconnue : il s'agit de l'estimer à partir de l'échantillon d'observations indépendantes (x_i, Y_i), appelé **échantillon d'apprentissage**.

En notant $\mathrm{F}(X, \theta)$ vecteur des espérances de chaque observation

$$\mathrm{F}(X, \theta) = \begin{pmatrix} f(\mathrm{x}_1, \theta) \\ \vdots \\ f(\mathrm{x}_n, \theta) \end{pmatrix} \tag{7.2}$$

le modèle s'écrit matriciellement

$$Y = \mathrm{F}(X, \theta) + \varepsilon; \;\; \mathbb{E}(\varepsilon) = 0; \;\; \mathrm{var}(\varepsilon) = \sigma^2 I_n. \tag{7.3}$$

Comme dans le cas linéaire, toutes les espérances sont conditionnelles aux variables explicatives, mais nous l'omettons dans la notation.

7.1.1 Modélisation de l'espérance

La fonction f est donc une fonction non linéaire de θ, ce qui arrive dans de nombreuses applications. Elle est déterminée soit a priori par la connaissance du phénomène physique qu'elle veut représenter, soit par l'observation des représentations graphiques générées à partir de l'échantillon. Citons par exemple :

— *le modèle de Michaelis-Menten.* Dans le cadre d'une étude enzymatique, ce modèle définit la relation entre Y, la vitesse initiale de formation du produit et x, la concentration initiale du substrat :

$$f(x, \theta) = \frac{\theta_1 x}{\theta_2 + x}. \tag{7.4}$$

Ce modèle est fondé sur des approximations, mais il a montré son caractère opératoire. Le modèle est de dimension 2, et pourra être estimé dès que l'échantillon contient au moins deux observations de concentrations initiales différentes.

— *les courbes sigmoïdales.* La première étape d'un dosage radio-immunologique consiste à établir une courbe étalon : à partir d'une gamme de dilutions connues d'une quantité déterminée d'hormone, on mesure la radioactivité correspondante. L'allure du graphique généré suggère de prendre une courbe de type logistique (où x est le logarithme de la dose d'hormone et $f(x, \theta)$ est la radioactivité mesurée)

$$f(x, \theta) = b + \frac{a - b}{1 + \exp(c(x - d))} \tag{7.5}$$

Nous avons déjà évoqué ce modèle pour une autre application, la croissance des alevins, cf Figure6.2.

— *les fonctions périodiques de période inconnue.* On peut alors poser la fonction de régression suivante

$$f(t, \theta) = \theta_1 + \theta_2 \cos(\theta_4 t) + \theta_3 \sin(\theta_4 t).$$

Notons que ce modèle serait linéaire si θ_4 était connu !

Dans ces exemples, il y a une unique covariable, $\mathrm{x} = (x)$, ou $\mathrm{x} = (t)$, mais la dimension de θ est variée. Gallant et Goebel (1976) propose une application avec plusieurs variables explicatives :

— *une des covariables affecte la réponse exponentiellement*

$$f(\mathrm{x}, \theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 \exp(\theta_3 x_3).$$

Enfin, la régression non linéaire pourrait considérer le cas particulier d'une fonction f linéaire en θ. Notons cependant que dans un modèle de régression non linéaire (au sens strict), au moins l'une des dérivées partielles de l'espérance $\mathbb{E}(Y)$ par rapport aux composantes du paramètre dépend d'au moins un paramètre ; dans un modèle de régression linéaire, ces dérivées ne dépendent pas des paramètres inconnus.

7.1.2 Modélisation de la loi de l'erreur

Le modèle linéaire gaussien est souvent utilisé car il est facile à estimer, mais il est très restrictif et l'hypothèse de normalité peut être irréaliste : elle implique une densité à support infini et symétrique. C'est pourquoi l'extension à un bruit de densité continue peut permettre une plus grande robustesse.

Par ailleurs, l'hypothèse d'homoscédasticité du bruit pourrait être critiquée. En effet, sa variance pourrait changer au cours de l'expérience, par exemple en dépendant de l'espérance : $\operatorname{var} Y_i = f(\mathrm{x}_i)^2\sigma_0^2$, les grandes valeurs sont plus dispersées que les petites. La variance pourrait faire intervenir un paramètre β (éventuellement multi-dimensionnel) spécifique :

$$\operatorname{var} Y_i = (f(\mathrm{x}_i))^{\beta_1}\sigma_0^2 \text{ ou } \operatorname{var} Y_i = (1+\beta_1 f(\mathrm{x}_i))\sigma_0^2.$$

Il existe des techniques de transformations de variables (Box-Cox par exemple) pour linéariser ou pour stabiliser la variance : elles nécessitent une utilisation experte car elles modifient la structure du modèle, mais peuvent apporter des réponses en cas d'hétéroscédasticité. Les moyens de calculs actuels permettent maintenant de traiter les modèles non-linéaires hétéroscédastiques directement, mais ce point dépasse le cadre de ce cours.

7.1.3 Transformation en modèle linéaire

Certains modèles non-linéaires peuvent être transformés en modèles linéaires. Considérons par exemple le modèle non-linéaire

$$Y_i = \theta_1 e^{\theta_2 x_i} + \varepsilon_i. \tag{7.6}$$

Son espérance peut être linéarisée par passage au logarithme

$$\log E(Y_i) = \log(\theta_1) + \theta_2 x_i$$

d'où la proposition du modèle linéaire suivant :

$$\log(Y_i) = Z_i = \tilde{\theta}_1 + \tilde{\theta}_2 x_i + \varepsilon_i. \tag{7.7}$$

Il faut prendre des précautions dans cette approche. En effet, l'estimation par moindres carrés du modèle linéaire (7.7) ne donne pas en général le même résultat que celle du modèle original non-linéaire (7.6) : dans le cas linéarisé, l'EMC minimise la somme des carrés résiduels de $\log Y$, tandis que dans le cas non-linéaire, c'est celle de Y qui est minimisée : le passage en logarithme peut "tasser" les estimations vers les plus faibles valeurs en donnant une plus forte influence à ces dernières : voir une illustration sur un modèle de Michaelis-Menten dans Myers et al. (2012).

De plus, le passage au logarithme dans un modèle non-linéaire additif (7.6) transforme la structure du bruit comme dans les méthodes de stabilisation de la variance : le bruit du modèle transformé ne peut être additif comme proposé dans (7.7). Cependant, si les erreurs sont multiplicatives

$$Y_i = (\theta_1 e^{\theta_2 x_i})(1+\varepsilon_i), \tag{7.8}$$

la transformation en logarithme donne

$$\log(Y_i) = Z_i = \tilde{\theta}_1 + \tilde{\theta}_2 x_i + \tilde{\varepsilon}_i, \tag{7.9}$$

où $\tilde{\varepsilon}$ est un bruit de variance constante. Si la loi de $\tilde{\varepsilon}$ est supposée normale, celle de $1+\varepsilon$ est log-normale. Un modèle de régression non-linéaire pouvant être transformé en modèle linéaire équivalent comme dans (7.8) et (7.9) est appelé **intrinsèquement linéaire**.

7.1.4 Conditions suffisantes d'identifiabilité

Il ne suffit pas que $f(.,\theta)$ soit identifiable en θ (c'est à dire que deux paramètres θ et θ' différents déterminent deux fonctions de régression $f(.,\theta)$ et $f(.,\theta')$ différentes) pour pouvoir correctement en estimer le paramètre θ : encore faut-il que l'échantillon contienne suffisamment d'information au "bon endroit", i.e. que l'application $\theta \to \mathrm{F}(X,\theta)$ de $\mathbb{R}^p$ dans $\mathbb{R}^n$ soit également injective.

Condition de submersion

Si F est dérivable par rapport à θ, de dérivée la matrice $\dot{\mathrm{F}}(\theta)$ de dimension $n \times p$,

$$\dot{\mathrm{F}}(\theta) = \left(\dot{\mathrm{F}}_k\right)_{k=,1\ldots,p}; \;\; \dot{\mathrm{F}}_k = \left(\frac{\partial f}{\partial \theta_k}(\mathrm{x}_i,\theta)\right),$$

où $\dot{\mathrm{F}}_k$ est une colonne de $\dot{\mathrm{F}}(\theta)$, la condition de **submersion** assure l'identifiabilité à distance finie :

$$\forall \theta \in \Theta, \dot{\mathrm{F}}(\theta) \text{ est de rang } p.$$

En effet, le théorème des accroissements finis donne, pour $\tilde{\theta} \in [\theta_1, \theta_2]$

$$\mathrm{F}(\theta_1) - \mathrm{F}(\theta_2) = \dot{\mathrm{F}}(\tilde{\theta})(\theta_1 - \theta_2) \neq 0 \text{ si } \theta_1 \neq \theta_2.$$

Si le modèle est linéaire, cette condition est aussi nécessaire puisqu'alors $\dot{\mathrm{F}}(\theta) = X$ doit être de rang p. Dans le cas non linéaire, on peut avoir l'identifiabilité alors que $\dot{\mathrm{F}}$ n'est pas dérivable.

De manière générale, cette condition est vérifiée pour un choix correct des x_i, i.e. du plan d'expérience. Dans la pratique, étant donnée la manière complexe dont les valeurs x_i interviennent dans l'expression, il est impossible de faire un choix a priori de ce plan d'expérience fondé sur des méthodes analytiques. On supposera donc cette condition vérifiée a priori.

Identifiabilité asymptotique

L'**identifiabilité asymptotique** est une condition suffisante garantissant que l'identifiabilité se maintient ou s'acquiert quand $n \to \infty$. S'il existe une fonction $c(\theta,\theta^\star)$ définie par la limite suivante

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} (f(\mathrm{x}_i,\theta) - f(\mathrm{x}_i,\theta^\star))^2 = c(\theta,\theta^\star),$$

d'**unique** minimum en la vraie valeur $\theta^\star$ du paramètre inconnu, alors $\theta^\star$ est identifiable : $c(\theta,\theta^\star)$ sépare les paramètres du modèle. Nous retrouverons cette condition dans l'hypothèse (iv) du Théorème 5 de consistance de la section suivante.

Si le modèle est linéaire (non forcément gaussien), une condition suffisante est de supposer le plan d'expérience X_n généré avec des x_i réalisations i.i.d. d'une loi Z sur $\mathbb{R}^p$ de carré intégrable

$$\frac{1}{n} X_n' X_n \to_{n\to\infty} J = E(ZZ'), \tag{7.10}$$

alors l'identifiabilité asymptotique est vérifiée :

$$\frac{1}{n}(\theta - \theta^\star)'(X_n' X_n)(\theta - \theta^\star) \to c(\theta,\theta^\star) = (\theta - \theta^\star)' J (\theta - \theta^\star).$$

Cette condition n'est pas nécessaire, cf Guyon (2001) : prendre par exemple $x_i = 1/\sqrt{i}$ pour estimer une régression linéaire simple. Alors, $(X'X)^{-1}$ tend vers 0, donc l'estimateur est consistant,

mais $n(X'X)^{-1}$ diverge. La vitesse de l'estimateur dans ce cas est plus lente que dans le cas x_i i.i.d. du fait de la plus faible dispersion des données.

Renforçons le fait qu'un choix incorrect des x_i peut amener à une non identifiabilité asymptotique, alors que le modèle est identifiable pour tout n. Considérons l'ANOVA1 paramétré en $E(Y_{ik}) = a_i$ où $i = 1, \ldots, I$ représente les niveaux du facteur, et $k = 1, \ldots, n_i$ le nombre de répétitions observées sur chaque niveau, $n = \sum_i n_i$. Alors, en supposant que $n_1 = \log(\tilde{n})$, $n_2 = \ldots n_I = \tilde{n}$

$$\lim_{n\to\infty} \frac{1}{n} ||\mathrm{F}(X,\theta) - \mathrm{F}(X,\theta)||_n^2 = \lim_{n\to\infty} \sum_{i=1}^{I} \frac{n_i}{n}(a_i - a_i^*)^2 = \sum_{i=2}^{I}(a_i - a_i^*)^2,$$

et a_1 n'est plus identifiable. La condition indique donc qu'il faut non seulement des observations qui alimentent l'estimation de chaque paramètre, mais aussi qu'il y en ait suffisamment.

7.2 Estimateur des moindres carrés

Quand la loi de l'erreur ε n'est pas précisée, l'estimation par moindres carrés est une des méthodes les plus populaires pour estimer une fonction de régression à partir d'un échantillon d'apprentissage $(\mathrm{x}_i, Y_i), i = 1, \ldots, n$, et nous l'avons déjà rencontrée en régression linéaire. Elle fait intervenir la somme des carrés des erreurs entre les observations et leurs estimations, appelée **somme des carrés résiduels** $SCR(\theta)$:

$$SCR(\theta) = \sum_{i=1}^{n}(Y_i - f(\mathrm{x}_i,\theta))^2 = ||Y - \mathrm{F}(X,\theta)||_n^2,$$

qu'il semble naturel de minimiser.

Définition 13. *L'estimateur $\widehat{\theta}_n$ des **moindres carrés** ordinaires (**EMC**) minimise la somme des carrés résiduels entre la réponse et la fonction de régression*

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} SCR(\theta).$$

7.2.1 Calcul de l'estimateur

L'estimateur est cherché parmi les solutions des **équations normales**, qui annulent la statistique du **score**, dérivée $\nabla C_n(\theta)$ du contraste C_n par rapport au paramètre :

$$\nabla C_n(\theta) = \Big(\frac{\partial C_n(\theta)}{\partial \theta_k}\Big)_{k=1,\ldots,p} = 0$$

et telles que la matrice des dérivées secondes du contraste (**hessien**)

$$\nabla^2 C_n(\theta) = \Big(\frac{\partial^2 C_n(\theta)}{\partial \theta_k \partial \theta_\ell}\Big)_{k=1,\ldots,p;\ell=1,\ldots,p}$$

soit définie positive.

Il existe une expression explicite de l'EMC dans le cas linéaire : $\widehat{\theta}_n = (X'X)^{-1}X'Y$. En revanche, des procédures numériques itératives de type Newton sont nécessaires pour calculer l'EMC dans un cadre général non-linéaire, car les équations normales ne sont pas linéaires en θ et n'admettent pas en général de solution explicite.

Schéma itératif de résolution des équations normales

L'algorithme de Newton est un algorithme itératif pour le calcul du zéro d'une fonction. Ici $\widehat{\theta}_n$, minimum de $C_n(\theta)$, annule $\nabla C_n(\theta)$, le gradient de $C_n(\theta)$ en θ : $\nabla C_n(\widehat{\theta}_n) = 0$. Au voisinage d'une approximation numérique $\theta^{(m)}$ de $\widehat{\theta}_n$, on choisit $\theta^{(m+1)}$ qui annule l'approximation linéaire de $\nabla C_n(.)$, donnée par son développement de Taylor au premier ordre :

$$\nabla C_n(\theta^{(m+1)}) = 0 = \nabla C_n(\theta^{(m)}) + \nabla^2 C_n(\theta^{(m)})(\theta^{(m+1)} - \theta^{(m)}),$$

Si $\nabla^2 C_n(\theta^{(m)})$ est inversible, la valeur de

$$\theta^{(m+1)} = \theta^{(m)} - [\nabla^2 C_n(\theta^{(m)})]^{-1} \nabla C_n(\theta^{(m)})$$

peut être calculée et l'algorithme réitéré avec cette nouvelle valeur. Ce schéma itératif est convergent vers un zéro de $\nabla C_n(\theta)$ pour une fonction localement convexe Ciarlet (1990).

Conditions initiales

Θ étant compact, l'EMC existe dès que $\theta \mapsto f(\theta)$ est continue. Mais rien ne garantit la convexité de la fonction $\theta \mapsto C_n(\theta)$, ni l'unicité de son minimum. Il est donc important de bien calibrer les valeurs initiales $\theta^{(0)}$ de l'algorithme et les choisir proches de la solution pour diminuer les problèmes de convergence : un mauvais choix peut conduire à un défaut de convergence, ou à un minimum local.

Plusieurs approches peuvent être utilisées :

— Interpréter les paramètres sous forme de caractéristiques de la fonction $\mathrm{x} \to f(\mathrm{x}, \theta)$ (asymptote, points d'inflection, dérivée à l'origine) ou de caractéristiques physiques du phénomène observé, pour permettre une première estimation visuelle. Dans le cas d'une courbe sigmoïde (7.5) par exemple, les droites $y = a$ et $y = b$ sont deux asymptotes horizontales.

— Calculer des paramètres dans un espace linéarisé, puis effectuer la transformation inverse. Le modèle de Michaelis-Menten (7.4) peut se linéariser en

$$\frac{1}{f(x,\theta)} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{x} = \alpha_1 + \alpha_2 \frac{1}{x}.$$

Le modèle linéaire en α est estimé, pour donner une initialisation de θ.

7.2.2 Minimum de contraste

L'EMC choisit dans la famille de courbes $\{f(x,\theta), \theta \in \Theta\}$ celle qui passe le plus près de l'ensemble des valeurs observées, la proximité étant mesurée par $SCR(\theta)$. L'erreur moyenne commise entre en choisissant θ à la place de θ^* est $\mathbb{E}(SCR(\theta) - SCR(\theta^*))/n$. On est ainsi amené à définir $C_n(\theta)$ la moyenne des carrés résiduels par

$$\theta \mapsto C_n(\theta) = \frac{1}{n} SCR(\theta) = \frac{1}{n} ||Y - \mathrm{F}(X,\theta)||_n^2, \tag{7.11}$$

et

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} SCR(\theta) = \arg\min_{\theta \in \Theta} C_n(\theta).$$

L'heuristique est la suivante : soit $\theta^\star$ la valeur du paramètre avec lequel le modèle (7.1) a été généré, i.e. de fonction de régression $\mathbb{E}(Y_i) = f(\mathrm{x}_i, \theta^\star)$. L'excès de risque

$$\begin{aligned}\ell(\theta, \theta^\star) &= \mathbb{E}[C_n(\theta)] - \mathbb{E}[C_n(\theta^\star)] \\ &= \frac{1}{n}\mathbb{E}\left[||Y - \mathrm{F}(X,\theta)||_n^2 - ||Y - \mathrm{F}(X,\theta^\star)||^2\right] \\ &= \frac{1}{n}||\mathrm{F}(X,\theta) - \mathrm{F}(X,\theta^\star)||^2 \geq 0,\end{aligned}$$

est minimal en $\theta^\star$. Bien sûr, il est impossible de minimiser $\ell(\theta, \theta^\star)$ (ou $\mathbb{E}[C_n(\theta)]$, qui est aussi minimale en $\theta^\star$), puisque θ^* est inconnu, mais il est tout à fait possible de minimiser sa version empirique $C_n(\theta)$.

Si les observations sont identiquement distribuées selon la loi jointe $P(y, \mathrm{x}; \theta^*)$, et si $\ell(\theta, \theta^\star)$ tend asymptotiquement ($n \to \infty$) vers une fonction $c(\theta, \theta^\star)$ d'**unique** minimum $\theta^\star$ (cf. identifiabilité asymptotique), on peut intuitivement penser que $C_n(\theta)$ tend vers $\mathbb{E}(C_n(\theta))$, et que $\widehat{\theta}_n$, le lieu du minimum du contraste empirique $C_n(.)$ tend vers $\theta^\star$, le lieu du minimum de $c(., \theta^\star)$. A n fixé, l'excès de risque associé à $\widehat{\theta}_n$ est $\mathbb{E}[\ell(\widehat{\theta}_n, \theta^\star)]$.

Pour que ce schéma fonctionne, nous verrons que la convergence uniforme de $C_n(.)$ est nécessaire, pour conserver la forme de la fonction à la limite. La fonction $C_n(.)$ est appelée contraste empirique :

Définition 14. *On appelle* ***contraste empirique*** *en θ^*, une fonction $\theta \mapsto C_n(\theta)$ telle que*

$$\lim_{n\to\infty}\Big(\mathbb{E}[C_n(\theta)] - \mathbb{E}[C_n(\theta^\star)]\Big) = c(\theta, \theta^\star)$$

où $\theta \mapsto c(\theta, \theta^\star)$ a pour ***unique*** *minimum $\theta^\star$. $\theta \mapsto c(., \theta^\star)$ est appelée* ***fonction de contraste****.*

Le schéma présenté est assez général et s'applique à d'autres fonctions de contraste, comme par exemple la vraisemblance. L'EMC est un exemple d'**estimateur du minimum de contraste**.

7.2.3 Consistance de l'EMC

La consistance de l'EMC est démontrée par le théorème suivant :

Théorème 5 (Théorème 6 de Jennrich (1969))**.** *Soit $(\widehat{\theta}_n)$ une suite d'estimateurs des moindres carrés du modèle de régression non-linéaire paramétrique. Sous les hypothèses de la Définition 12, et en supposant de plus que :*

(iii) pour tout $\theta \in \Theta$, la limite suivante existe et est finie

$$\lim_{n\to\infty}\frac{1}{n}||\mathrm{F}(X,\theta)||_n^2, \tag{7.12}$$

(iv) et la fonction c définie par

$$\lim_{n\to\infty}\frac{1}{n}||\mathrm{F}(X,\theta) - \mathrm{F}(X,\theta^\star)||_n^2 = c(\theta, \theta^\star) \tag{7.13}$$

a un unique minimum en $\theta = \theta^\star$,

alors $\widehat{\theta}_n$ et $C_n(\widehat{\theta}_n)$ sont des estimateurs fortement consistant de $\theta^\star$ et σ^2.

Schéma général de la preuve. Les hypothèses de continuité de f en θ et de compacité de Θ suffisent à garantir l'existence d'estimateurs (mesurables) des moindres carrés dans le cas non-linéaire.

Le point clé de la preuve est la convergence uniforme de $C_n(\theta)$ quand $n \to \infty$. Commençons par étudier la convergence simple :

$$\begin{aligned} C_n(\theta) &= \frac{1}{n}||Y - \mathrm{F}(X,\theta)||_n^2 \\ &= \frac{1}{n}||Y - \mathrm{F}(X,\theta^\star) + \mathrm{F}(X,\theta^\star) - \mathrm{F}(X,\theta)||_n^2 \\ &= \frac{1}{n}||\varepsilon||_n^2 + \frac{1}{n}||\mathrm{F}(X,\theta) - \mathrm{F}(X,\theta^\star)||_n^2 + \frac{2}{n} < \varepsilon, \mathrm{F}(X,\theta) - \mathrm{F}(X,\theta^\star) >_n \\ &\xrightarrow{p.s.} \sigma^2 + c(\theta,\theta^\star) = c(\theta) \end{aligned}$$

Le premier terme tend vers σ^2 (loi des grands nombres), le deuxième vers $c(\theta,\theta^\star)$ par l'hypothèse (iv). On utilise le théorème de Chow, généralisation de la loi des grands nombres à des variables indépendantes centrées et hétéroscédastiques, pour traiter le troisième terme :

Théorème 6 (LGN de Chow (1967))**.** *Soit $S_n = \frac{1}{n}\sum_i Z_i$ la moyenne de n variables aléatoires Z_i centrées indépendantes (non forcément identiquement distribuées). Si*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{i^2} \operatorname{var}(Z_i) < +\infty$$

alors $\lim_n S_n = 0$ *presque sûrement.*

On applique le théorème avec $Z_i = (f(x_i,\theta) - f(x_i,\theta^\star))\varepsilon_i$, la condition est remplie grâce à (iii) et le troisième terme tend également vers 0. $C_n(\theta)$ tend donc vers $c(\theta) = \sigma^2 + c(\theta,\theta^\star)$. De plus la convergence précédente est uniforme grâce à la compacité de Θ :

$$\sup_\theta |C_n(\theta) - c(\theta)| \xrightarrow{p.s.} 0,$$

et la limite $c(\theta)$ est continue en θ. La convergence uniforme indique ainsi que la **forme** de $c(\theta)$ est respectée, ce qui est important car la définition de $\widehat{\theta}$ dépend de toute la fonction C_n.

Par l'hypothèse (iv), $\theta^\star$ est l'unique minimum de $c(\theta)$ d'où

$$\begin{aligned} 0 \le c(\widehat{\theta}_n) - c(\theta^\star) &\le c(\widehat{\theta}_n) - C_n(\widehat{\theta}_n) + C_n(\widehat{\theta}_n) - C_n(\theta^\star) + C_n(\theta^\star) - c(\theta^\star) \\ &\le c(\widehat{\theta}_n) - C_n(\widehat{\theta}_n) + C_n(\theta^\star) - c(\theta^\star) \qquad (7.14) \\ &\le \sup_\theta |C_n(\theta) - c(\theta)| + C_n(\theta^\star) - c(\theta^\star). \end{aligned}$$

La simplification dans (7.14) est la prise en compte du fait que $\widehat{\theta}_n$ minimise le contraste empirique. Grâce à la convergence uniforme de $C_n(.)$, on en déduit que $c(\widehat{\theta}_n) - c(\theta^\star) \xrightarrow{p.s.} 0$. La continuité de c et l'unicité du minimum appliquées à la dernière inégalité impliquent alors la consistance de $\theta^\star$: $\widehat{\theta}_n \xrightarrow{p.s.} \theta^\star$. ◊

7.2.4 Loi asymptotique de l'EMC

L'EMC étant consistant, il est possible de faire un développement de Taylor de ∇C_n autour de $\theta^\star$. L'étude asymptotique des différents termes de ce développement permettent de démontrer un comportement asymptotiquement gaussien de $\widehat{\theta}$ sous certaines hypothèses de régularité du modèle concernant les dérivées premières et secondes de F, que nous notons $\dot{\mathrm{F}}_k$ et $\ddot{\mathrm{F}}_{k\ell}$ (sans rappeler qu'elle sont des fonctions de la variable θ) :

$$\dot{\mathrm{F}}_k = \left(\frac{\partial f}{\partial \theta_k}(\mathrm{x}_i,\theta)\right)_{i=1,\dots n} \text{ et } \ddot{\mathrm{F}}_{k\ell} = \left(\frac{\partial^2 f}{\partial \theta_k \partial \theta_\ell}(\mathrm{x}_i,\theta)\right)_{i=1,\dots n}$$

Théorème 7 (Théorème 7 de Jennrich (1969))**.** *Soit $(\widehat{\theta}_n)$ une suite d'estimateurs des moindres carrés de $\theta^\star$ du modèle (7.1). Sous les hypothèses (i) à (iv) et en supposant de plus :*

*(v) f est **deux fois continûment différentiable** sur Θ, et toutes les limites suivantes existent et sont des nombres réels pour tout g, h à prendre parmi les fonctions $\mathrm{F}, \dot{\mathrm{F}}_k, \ddot{\mathrm{F}}_{k\ell}$:*

$$\lim_{n\to\infty} \frac{1}{n} < g, h >_n . \tag{7.15}$$

En particulier,

$$\lim_{n\to\infty} \frac{1}{n} < \dot{\mathrm{F}}_k, \dot{\mathrm{F}}_\ell >_n = \mathbf{J}_{\theta k\ell}. \tag{7.16}$$

*(vi) $\theta^\star$ est un point **intérieur** de Θ et $J_{\theta^\star}$ est **inversible**.*

Alors,

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \sigma^2 J_{\theta^\star}^{-1}). \tag{7.17}$$

De plus, $\frac{1}{n} < \dot{\mathrm{F}}_k(\widehat{\theta}_n), \dot{\mathrm{F}}_\ell(\widehat{\theta}_n) >_n$ est un estimateur fortement consistant de $J_{\theta^\star}$.

Avant d'aborder des éléments de preuve, commençons par deux propriétés asymptotiques des dérivées du contraste en $\theta^\star$.

Proposition 1. *Sous les hypothèses (i) et (v),*

1. *la **statistique de score** $\nabla C_n(\theta^\star)$, gradient du contraste en $\theta^\star$, tend asymptotiquement vers 0 avec le comportement suivant*

$$\sqrt{n}\nabla C_n(\theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma(\theta^\star)),$$

où $\Gamma(\theta^\star)$ est la variance limite

$$\mathrm{var}(\sqrt{n}\nabla C_n(\theta^\star)) \xrightarrow{\mathcal{P}} 4\sigma^2 J_{\theta^\star} = \Gamma(\theta^\star).$$

2. *le **hessien** $\nabla^2 C_n(\theta^\star)$ du contraste en $\theta^\star$ tend asymptotiquement vers $I(\theta^\star) = 2J_{\theta^\star}$.*

Preuve. La k-ième composante du gradient du contraste s'écrit :

$$\begin{aligned}
\nabla_k C_n(\theta^\star) &= \Big(\frac{\partial C_n(\theta)}{\partial \theta_k}\Big)_{|\theta=\theta^\star} \\
&= -\frac{2}{n} < \dot{\mathrm{F}}_k, Y - \mathrm{F}(X,\theta) >_{n|\theta=\theta^\star} \\
&= -\frac{2}{n} < \dot{\mathrm{F}}_k, \varepsilon >_{n|\theta=\theta^\star} .
\end{aligned}$$

Elle est d'espérance nulle car $\mathbb{E}(Y) = \mathrm{F}(X, \theta^\star)$. De plus, elle tend vers 0 par la loi des grands nombres généralisée que nous avons déjà rencontrée, valide grâce à (v). Cette même hypothèse permet d'obtenir le comportement asymptotiquement gaussien de $\sqrt{n}\nabla C_n$ (extension du théorème central limite pour des suites centrées indépendantes hétéroscédastiques) et la variance limite se calcule aisément en utilisant le fait que $\mathrm{var}(\varepsilon) = \sigma^2 I_n$. En effet, pour toute coordonnée $\nabla_k C_n(\theta^\star) = \partial C_n(\theta^\star)/\partial\theta_k$ et $\nabla_\ell C_n(\theta^\star) = \partial C_n(\theta^\star)/\partial\theta_\ell$ de $\nabla C_n(\theta^\star)$ on a :

$$\begin{aligned}
\mathrm{cov}(\sqrt{n}\nabla_k C_n(\theta^\star), \sqrt{n}\nabla_\ell C_n(\theta^\star)) &= \mathbb{E}\left(\sqrt{n}\nabla_k C_n(\theta^\star)\sqrt{n}\nabla_\ell C_n(\theta^\star)\right) \\
&= \frac{4}{n}\mathbb{E}\left(\sum_{i,i'} \varepsilon_i\varepsilon_{i'}\dot{\mathrm{F}}^\star_k(i)\dot{\mathrm{F}}^\star_\ell(i')\right) \\
&= \frac{4\sigma^2}{n} < \dot{\mathrm{F}}^\star_k, \dot{\mathrm{F}}^\star_\ell >_n \qquad \to 4\sigma^2[J_{\theta^\star}]_{k\ell}
\end{aligned}$$

d'où la première propriété. De plus,

$$\begin{aligned}
\frac{\partial^2 C_n(\theta^\star)}{\partial\theta_k\partial\theta_\ell} &= -\frac{2}{n}\sum_i (Y_i - f(\mathrm{x}_i, \theta^\star))\ddot{\mathrm{F}}_{k\ell}(i) + \frac{2}{n}\sum_i \dot{\mathrm{F}}_k(i)\dot{\mathrm{F}}_\ell(i) \\
&= -\frac{2}{n} < \varepsilon, \ddot{\mathrm{F}}_{k\ell} >_n + \frac{2}{n} < \dot{\mathrm{F}}_k, \dot{\mathrm{F}}_\ell >_n \qquad \to 2[J_{\theta^\star}]_{k\ell}.
\end{aligned}$$

$\Diamond$

Elements de preuve du Théorème 7. L'EMC $\widehat{\theta}_n$ est solution des équations normales $\nabla C_n(\widehat{\theta}_n) = 0$. Comme il est fortement consistant en $\theta^\star$, point intérieur de Θ par l'hypothèse (vi), il prend presque sûrement ses valeurs dans un voisinage **compact convexe** de $\theta^\star$ pour n suffisamment grand. Pour chaque coordonnée $\nabla_k C_n(.)$ du score, le théorème des accroissements finis donne l'existence de $\overline{\theta}_n$ aléatoire tel que $|\overline{\theta}_n - \theta^\star| \leq |\widehat{\theta}_n - \theta^\star|$ et :

$$\nabla_k C_n(\theta^\star) = \nabla_k C_n(\widehat{\theta}_n) + [\nabla\nabla_k C_n(\overline{\theta}_n)]'(\theta^\star - \widehat{\theta}_n),$$

où $\nabla\nabla_k C_n$ est le vecteur des dérivées de la k-ème coordonnée du score. En multipliant par $\sqrt{n}$,

$$[\nabla\nabla_k C_n(\overline{\theta}_n)]'\sqrt{n}(\widehat{\theta}_n - \theta^\star) = \sqrt{n}\nabla_k C_n(\widehat{\theta}_n) - \sqrt{n}\nabla_k C_n(\theta^\star).$$

Or,

- $\nabla_k C_n(\widehat{\theta}_n) = 0$ dès que $\widehat{\theta}_n$ est intérieur à Θ,
- d'après la forte consistance de $\widehat{\theta}_n$ et la propriété 2 de la proposition 1, $[\nabla\nabla_k C_n(\overline{\theta}_n)]$ tend presque sûrement vers la k-ème colonne de $I(\theta^\star) = 2J_{\theta^\star}$,
- la propriété 1 de la proposition 1 précise le comportement asymptotique normal $\mathcal{N}(0, I(\theta^\star))$ de $\nabla C_n(\theta^\star)$.

D'où sous la condition d'inversibilité (vi) :

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I(\theta^\star)^{-1}\Gamma(\theta^\star)I(\theta^\star)^{-1}\right),$$

avec

$$I^{-1}(\theta^\star)\Gamma(\theta^\star)I^{-1}(\theta^\star) = \frac{1}{2}J^{-1}_{\theta^\star}4\sigma^2 J_{\theta^\star}\frac{1}{2}J^{-1}_{\theta^\star} = \sigma^2 J^{-1}_{\theta^\star}$$

$\Diamond$

Nous verrons à la section 7.4 comment définir des lois à distance finie. Celles-ci permettront en particulier de construire des intervalles ou des régions de confiance de fonctions du paramètre.

7.2.5 Asymptotique du constraste empirique C_n

On en déduit l'asymptotique du contraste $C_n(\widehat{\theta}_n)$ par un développement de Taylor dont on peut montrer que le reste tend vers 0 :

$$\begin{aligned} n(C_n(\widehat{\theta}_n) - C_n(\theta^\star)) &\simeq n[\nabla C_n(\theta^\star)]'(\widehat{\theta}_n - \theta^\star) + \frac{n}{2}(\widehat{\theta}_n - \theta^\star)'\nabla^2 C_n(\theta^\star)(\widehat{\theta}_n - \theta^\star) \\ &\simeq -n(\widehat{\theta}_n - \theta^\star)'\nabla^2 C_n(\theta^\star)(\widehat{\theta}_n - \theta^\star) + \frac{n}{2}(\widehat{\theta}_n - \theta^\star)'\nabla^2 C_n(\theta^\star)(\widehat{\theta}_n - \theta^\star) \\ &\simeq -\frac{n}{2}(\widehat{\theta}_n - \theta^\star)'\nabla^2 C_n(\theta^\star)(\widehat{\theta}_n - \theta^\star) \\ &\simeq -\frac{n}{2}(\widehat{\theta}_n - \theta^\star)' I(\theta^\star)(\widehat{\theta}_n - \theta^\star) \end{aligned}$$

Or, $I(\theta^\star)/2 = J_{\theta^\star}$, d'où

$$n\frac{(C_n(\theta^\star) - C_n(\widehat{\theta}_n))}{\sigma^2} \simeq \sqrt{n}(\theta^\star - \widehat{\theta}_n)\frac{J_{\theta^\star}}{\sigma^2}\sqrt{n}(\theta^\star - \widehat{\theta}_n),$$

soit :

$$T_{RV} = n\frac{C_n(\theta^\star) - C_n(\widehat{\theta})}{\sigma^2} \xrightarrow{\mathcal{L}} \chi^2_p$$

où p est la dimension du modèle. En particulier, $C_n(\widehat{\theta}_n)$ est un estimateur consistant de σ^2.

Afin de bâtir des tests de sous-modèles, il est intéressant d'étudier l'extension du résultat précédent aux observations issues d'un modèle ω de dimension $p-q$ emboîté dans Ω de dimension p. Dans ce cas et si σ est connu, le comportement asymptotique de la statistique T_{RV} est défini par

$$T_{RV} = n\frac{C_n(\widehat{\theta}_\omega) - C_n(\widehat{\theta}_\Omega)}{\sigma^2} \xrightarrow{\mathcal{L}} \chi^2_r,$$

où r est la différence de dimension entre les deux modèles. On reconnaît au numérateur $SCR(\widehat{\theta}_\omega) - SCR(\widehat{\theta}_\Omega)$. Si la variance σ^2 est inconnue, elle est estimée par $\widehat{\sigma}^2 = SCR(\widehat{\theta}_\Omega)/(n-p)$ dans Ω soit

$$T_{RV} = \frac{SCR(\widehat{\theta}_\omega) - SCR(\widehat{\theta}_\Omega)}{SCR(\widehat{\theta}_\Omega)/(n-p)} = rF \xrightarrow{\mathcal{L}} \chi^2_r \tag{7.18}$$

On retrouve une expression voisine de celle la statistique de Fisher $F = T_{RV}/r$ déjà rencontrée dans le test de sous-modèles du modèle linéaire. Mais dans le cas non-linéaire, sa loi n'est connue qu'asymptotiquement.

7.2.6 Asymptotique de la régression linéaire

On en déduit l'asymptotique de la régression linéaire, l'hypothèse gaussienne n'étant pas nécessaire :

Théorème 8. *Sous les hypothèses suivantes*
- $\frac{1}{n}(X'X) \to J$ *où* J *est définie positive,*
- $\sum_{i=1}^n ||x_i||^3 = O(n^{3/2})$
- $\mathbb{E}(|\varepsilon|^3) < \infty$

On a :
- $\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, \sigma^2 J^{-1})$
- $\widehat{\sigma}^2_n$ *tend en probabilité vers* σ^2
- *si de plus* $0 < \mathrm{var}(\varepsilon^2) < \infty$ *et* ε *gaussien, alors* $\sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 2\sigma^4)$.

7.3 Estimation par Maximum de Vraisemblance

Dans le cas où la loi de ε est spécifiée et connue (en général gaussienne), et la famille de loi dominée par une mesure commune, il est possible d'écrire la vraisemblance des observations, version de la densité de l'échantillon vue comme une fonction du paramètre θ. Soit $\phi(y; \mathrm{x}, \theta)$ la densité d'une observation de régresseur x. La vraisemblance de l'échantillon est le produit des densités de chacune des observations puisque celles-ci sont indépendantes, soit $L(\theta; Y) = \prod_i \phi(Y_i; \mathrm{x}, \theta)$. La méthode du maximum de vraisemblance choisit une valeur des paramètres qui maximise la vraisemblance de l'échantillon. Il est identique (et plus pratique) de maximiser le logarithme de la vraisemblance $\log L(\theta; Y) = \sum_i \log_i \phi(Y_i; \theta)$. Dans le cas gaussien,

$$\log L(\theta; Y) = -\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} SCR(\theta).$$

Ainsi, comme dans le cas linéaire, l'estimateur du maximum de vraisemblance (EMV) de θ a la même expression que celui de l'EMC de θ :

$$\arg\max_\theta \log L(\theta) = -\arg\min_\theta SCR(\theta).$$

mais ce dernier ne nécessite pas de spécifier la loi de Y. Remarquons que maximiser la log-vraisemblance revient à minimiser

$$\tilde{C_n}(\theta) = -\frac{1}{n} \log L(\theta; Y) = -\frac{1}{n} \sum_i \log \phi(Y_i; \theta).$$

et que

$$\tilde{C_n}(\theta) = \frac{1}{2} C_n(\theta)$$

dans le cas gaussien. L'EMV est également un estimateur du minimum du contraste, $\tilde{C_n}(\theta)$, et le schéma général que nous avons développé pour la consistance et la normalité asymptotique de l'EMC peut être repris pour l'EMV. Sous quelques hypothèses de régularité de $L(\theta; Y)$ et pour un modèle régulier, i.e. tel que

$$\mathbb{E}\left[(\nabla \log L(Y; \theta^\star))' \nabla \log L(Y; \theta^\star)\right] = -E\left[\nabla^2 \log L(Y; \theta^\star)\right],$$

et

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[(\nabla \log L(Y; \theta^\star))' \nabla \log L(Y; \theta^\star)\right] = I(\theta^\star),$$

alors l'EMV est consistant et asymptotiquement gaussien :

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I(\theta^\star)^{-1}\right). \tag{7.19}$$

7.3.1 Efficacité

L'estimateur du maximum de vraisemblance est asymptotiquement efficace. Mais on ne peut pas parler d'efficacité au sens de la borne de Cramer-Rao pour l'EMC, puisque, faute de connaître la vraisemblance, il est impossible d'en calculer l'information de Fisher. Cependant, dans le cadre d'erreurs gaussiennes i.i.d, l'EMC est (asymptotiquement) efficace, étant équivalent à l'EMV. En général, les EMC ne sont pas efficaces si le modèle n'est pas gaussien (Jennrich, 1969; Guyon, 2001).

7.3.2 Statistique du rapport de vraisemblance

On en déduit le comportement de la statistique du rapport de vraisemblances maximales quand σ^2 est connu :

$$\tilde{T}_{RV} = 2\log\frac{L(\widehat{\theta}_\Omega, Y)}{L(\widehat{\theta}_\omega, Y)} = 2n(\tilde{C}_n(\widehat{\theta}_\omega) - 2\tilde{C}_n(\widehat{\theta}_\Omega)) \xrightarrow{\mathcal{L}} \chi^2_r.$$

Si le modèle est gaussien,

$$\tilde{T}_{RV} = \frac{SCR(\widehat{\theta}_\omega) - SCR(\widehat{\theta}_\Omega)}{\sigma^2} \xrightarrow{\mathcal{L}} \chi^2_r.$$

Si de plus la variance σ^2 est inconnue, elle est estimée par $SCR(\widehat{\theta}_\Omega)/n$ dans Ω et $SCR(\widehat{\theta}_\omega)/n$ dans ω, soit

$$\begin{aligned} 2\log L(\widehat{\theta}_\Omega, Y) - \log L(\widehat{\theta}_\omega, Y) &= n\log\frac{SCR(\widehat{\theta}_\omega)}{SCR(\widehat{\theta}_\Omega)} \\ &= n\log(\frac{SCR(\widehat{\theta}_\omega) - SCR(\widehat{\theta}_\Omega)}{SCR(\widehat{\theta}_\Omega)} + 1) \\ &\simeq n\frac{SCR(\widehat{\theta}_\omega) - SCR(\widehat{\theta}_\Omega)}{SCR(\widehat{\theta}_\Omega)} \simeq rF \end{aligned}$$

quand la valeur de rF/n est faible. Ici, F est l'expression de la statistique de Fisher (B.7) telle que nous l'avons rencontrée au Chapitre 2. Ainsi, par référence au cas gaussien, la statistique rF en régression non-linéaire, de loi approchée $\chi^2(r)$, est parfois appelée statistique du "rapport de vraisemblance" dans le cas des moindres carrés, ce que nous avons fait en (7.18). Quand $n \to +\infty$, la loi de Fisher tend vers un χ^2/r , ce qui rend cohérent asymptotiquement la régression linéaire gaussienne et la régression non linéaire.

7.4 Lois à distance finie

L'EMC et l'EMV ont donc un comportement asymptotiquement normal

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, V_0(\theta^\star)),$$

où $V_0(\theta^\star)$ est la variance de la loi limite et p est la dimension du modèle, avec :

— $V_0(\theta^\star) = \sigma^2 J_{\theta^\star}^{-1}$ pour l'EMC,
— $V_0(\theta^\star) = I(\theta^\star)^{-1}$ pour l'EMV.

Soit $\widehat{V}_n$ un estimateur de la variance de $\widehat{\theta}_n$ tel que $n\widehat{V}_n \to V_0(\theta^\star)$. Par exemple

— cas EMV

$$\widehat{V}_n = -\left(\sum_i \nabla^2 \log(\phi(y_i; \mathrm{x}_i, \widehat{\theta}_n)\right)^{-1},$$

$$\widehat{V}_n = \frac{\widehat{\sigma}^2}{2}\left(n\nabla^2 C_n(\widehat{\theta}_n)\right)^{-1}.$$

Alors, en utilisant le théorème de Slutsky, on obtient les comportements asymptotiques suivants :

$$\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, Id_p) \tag{7.20}$$

$$(\widehat{\theta}_n - \theta^\star)'\widehat{V}_n^{-1}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \chi^2(p) \tag{7.21}$$

À distance finie, toutes les lois des statistiques sont approchées par celle de leur comportement limite.

Proposition 2. *Pour n suffisamment grand, la loi des estimateurs (EMV ou EMC) peut être approchée par celle de leur comportement limite. Ainsi,*

- *la loi de $\widehat{\theta}$ peut être approchée par une loi $\mathcal{N}(\theta^\star, \widehat{V}_n)$,*
- *la loi de la statistique de Wald $W = (A\widehat{\theta}_n - A\theta^\star)'(A\widehat{V}_nA')^{-1}(A\widehat{\theta}_n - A\theta^\star)$ est approchée par une loi $\chi^2(r)$,*
- *la loi de la statistique de rapport de vraisemblance T_{RV} ou $\tilde{T}_{RV}$ est approchée par une loi $\chi^2(r)$.*

De façon pratique, les logiciels calculent la variance estimée $\widehat{V}_n$ de $\widehat{\theta}$ et non la variance limite $V_0(\widehat{\theta}_n)$ de $\sqrt{n}\widehat{\theta}_n$: attention, il y a un rapport n entre les deux ! Par ailleurs, on peut écrire dans le cas de l'EMC :

$$\frac{\sqrt{n}J_{\widehat{\theta}_n}^{-1/2}(\widehat{\theta}_n - \theta^\star)}{\widehat{\sigma}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p).$$

Si l'estimateur $\widehat{\sigma}^2$ est $SCR(\widehat{\theta})/(n-p)$, il est possible de considérer l'approximation suivante :

Proposition 3. *Pour n suffisamment grand, et si la variance du bruit est estimée par $SCR(\widehat{\theta}_n)/(n-p)$,*

- *la loi de l'EMC centré réduit peut être est approchée par une loi de Student à $n-p$ degré de liberté $\mathcal{T}(n-p)$.*
- *la loi de W/r peut être approchée par une loi de Fisher $\mathcal{F}(r, n-p)$,*
- *la loi de T_{RV}/r peut être approchée par une loi de Fisher $\mathcal{F}(r, n-p)$.*

C'est le parti que prennent certains logiciels de statistiques, comme SAS, par analogie au cas gaussien linéaire pour lequel ces lois sont exactes.

7.4.1 Delta-méthode

Le comportement de $h(\widehat{\theta}_n)$, où h est une fonction (éventuellement) non linéaire de θ à valeurs dans $\mathbb{R}^r$ est déduit de celui de $\widehat{\theta}_n$ en effectuant un développement limité de h, et en négligeant les termes d'ordre 2 en $\widehat{\theta} - \theta^\star$. Soit

$$H = \begin{pmatrix} \frac{\partial h_1}{\partial \theta_1} & \cdots & \frac{\partial h_1}{\partial \theta_p} \\ \vdots & \frac{\partial h_j}{\partial \theta_k} & \vdots \\ \frac{\partial h_r}{\partial \theta_1} & \cdots & \frac{\partial h_r}{\partial \theta_p} \end{pmatrix}$$

Proposition 4. *Soit $h : \mathbb{R}^p \to \mathbb{R}^r$ différentiable en $\theta^\star$, tel que H soit de rang $q < p$. Alors,*

$$(H(\widehat{\theta}_n)\widehat{V}_nH(\widehat{\theta}_n)')^{-1/2}(h(\widehat{\theta}_n) - h(\theta^\star)) \xrightarrow{\mathcal{L}} \mathcal{N}_q(0, Id_q), \tag{7.22}$$

$$(h(\widehat{\theta}_n) - h(\theta^\star))'(H(\widehat{\theta}_n)\widehat{V}_nH(\widehat{\theta}_n)')^{-1}(h(\widehat{\theta}_n) - h(\theta^\star)) \xrightarrow{\mathcal{L}} \chi^2(q). \tag{7.23}$$

On en déduit aisément les lois approchées à distance finie. Si $h(\theta) = A\theta$ est linéaire, on retrouve la statistique de Wald

$$W = (A\widehat{\theta}_n - A\theta^\star)'(A\widehat{V}_n A')^{-1}(A\widehat{\theta}_n - A\theta^\star). \tag{7.24}$$

L'utilisation de la delta-méthode n'est pas toujours judicieuse, en particulier si h est un fonction monotone du paramètre : dans ce cas, si $[\widehat{IC}_{min}, \widehat{IC}_{max}]$ est un intervalle de confiance de θ, il est plus efficace de définir l'intervalle de confiance de $h(\theta)$ par $[h(\widehat{IC}_{min}), h(\widehat{IC}_{max})]$ si h est croissante et $[h(\widehat{IC}_{max}), h(\widehat{IC}_{min})]$ si h est décroissante.

7.4.2 Régression linéaire, non linéaire : analogie et différences

L'EMC d'une régression linéaire est explicite, linéaire en Y et sans biais ; sa variance est connue et vaut $\sigma^2(X'X)^{-1}$. De plus, si le modèle est gaussien, l'estimateur est gaussien pour tout n, et on dispose de lois exactes.

Mais ces propriétés ne se maintiennent pas pour une régression non-linéaire à n fini :

— l'estimateur $\widehat{\theta}_n$ est biaisé„
— il n'y a pas de formule explicite de $\widehat{\theta}_n$, ni de sa variance,
— on ne connaît pas la loi exacte de $\widehat{\theta}_n$ à distance finie, même dans le cas gaussien !

En revanche, sous certaines hypothèses, $\widehat{\theta}_n$ retrouve asymptotiquement les bonnes propriétés du modèle linéaire gaussien :

— $\widehat{\theta}_n$ converge vers $\theta^\star$, sa limite est gaussienne de variance identifiable,
— on dispose de tests asymptotiques pour une sous-hypothèse, que cette sous-hypothèse soit linéaire ou non.

Ainsi, asymptotiquement, la non-linéarité et/ou la non-gaussianité du modèle ne détériorent pas les propriétés statistiques du modèle. En non linéaire-gaussien, SCR suit approximativement un χ^2, Fisher est une approximation, les résidus sont approximativement gaussiens.

7.4.3 Intervalles de confiance et tests

Tous les principes de construction des tests et intervalles de confiance vus en régression linéaire restent valables en non-linéaire. Nous ne les rappelons donc pas ici. Il faut cependant rester attentif au fait que les lois utilisées sont maintenant approchées, et les niveaux observés ne peuvent être qu'approximativement le niveau désiré α. Par ailleurs, il faut aussi prendre soin de bien choisir le type de quantile en fonction de l'utilisation de la Propriété 2 ou de la Propriété 3 : i.e., vérifier l'option choisie par le logiciel pour calculer $\widehat{V}_n$, la variance estimée de $\widehat{\theta}_n$.

7.4.4 Comparaison TRV et test de Wald

Les tests de Wald et de rapport de vraisemblance comparent deux modèles emboîtés, ω restreignant Ω par $h(\theta) = 0$.

— Les deux tests de Wald et de RV sont conçus pour que le niveau asymptotique soit égal à α fixé. On peut vérifier que leur puissance tend vers 1. Pour n petit, le seuil $q_{\chi^2_q}(1-\alpha)$ est une approximation du seuil réel qui garantit un niveau donné.
— Le test de Wald est plus simple numériquement, car il ne nécessite qu'une seule optimisation du contraste ; mais ses qualités dépendent de l'estimation de la matrice de covariance, toujours délicate en pratique.

— Le test de RV est plus compliqué numériquement : il nécessite deux optimisations du contraste, dont une sous contrainte. D'un point de vue théorique, il est réputé meilleur, au sens où la différence entre le niveau réel du test et le niveau attendu α est minimale. Ceci se traduit par de meilleures performances 'à distance finie', constatées par des simulations dans de nombreuses situations.
— Dans le modèle linéaire gaussien, les deux tests sont équivalents, et équivalents au test de Fisher. Dans les autres modèles, on dispose de deux procédures fondées sur deux approximations différentes du contraste.

En présence de décisions contradictoires sur ces deux tests, on pourra préférer le test du rapport de vraisemblance.

Chapitre 8

Régression logistique

Malgré leurs nombreux avantages théoriques et pratiques, les régressions linéaires et non linéaires ne sont pas toujours les mieux adaptées, par exemple quand la réponse n'est pas continue, qu'elle est contrainte à être positive, ou qu'elle est hétéroscédastique.

Une nouvelle classe de modèles, les **modèles linéaires généralisés**, permettent de modéliser les lois de réponse qui appartiennent à une famille exponentielle de loi (loi de Bernoulli, exponentielle, de Poisson par exemple), réglant certains des inconvénients précédents. Ces modèles bénéficient de la simplicité d'un régresseur linéaire alliée à la flexibilité d'un modèle non linéaire. Utilisant des propriétés caractéristiques des familles exponentielles de loi, l'estimation y est aisée.

Dans le cas où la loi de la réponse est binaire, celle-ci est naturellement modélisée par une loi de Bernoulli ou une loi binomiale : c'est la régression **logistique**, qui fait l'objet de ce chapitre. C'est un exemple de modèle linéaire généralisé.

Bibliographie : Dobson (2002), Fahrmeir et al. (2013), Fahrmeir et Kaufmann (1985), McCullagh et Nelder (1989), Myers et al. (2012), Nelder et Wedderburn (1972), Cornillon et autres (2008)

8.1 Deux exemples

La réponse d'une expérience est binaire lorsqu'elle ne peut prendre que deux niveaux possibles de façon exclusive : par exemple, succès ou échec, mort ou vie, présence ou absence : il peut s'agir de classer un courrier électronique en indésirable ou normal, d'étudier la réussite ou l'échec à un examen, de définir si un individu est malade ou sain, ou de décider l'octroi d'un crédit par exemple. Commençons par présenter deux situations dans laquelle la réponse est binaire.

8.1.1 Fragilité d'un alliage

Un métallurgiste étudie le comportement en rupture d'éprouvettes d'acier, soumises à des cycles de fatigue. Quinze variables explicatives sont enregistrées pour chaque éprouvette : composition de l'alliage, temps de cuisson et température du four. Après les essais de fatigue, les éprouvettes sont classées en deux groupes : celui des éprouvettes robustes, et celui des éprouvettes fragiles qui ont cassé. Le métallurgiste se demande si les conditions de fabrication de l'acier (composition de l'alliage, variables de cuisson) influent sur la fragilité des éprouvettes [1].

1. Données gracieusement transmises par Patrick Pamphile, Université Paris Sud

La variable à expliquer est la classe de l'éprouvette : niveau 1, l'alliage de l'éprouvette est considéré comme fragile, niveau 0, il est robuste. C'est une variable binaire.

Les valeurs des covariables $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ étant différentes pour chaque éprouvette i, l'expérience est modélisée comme la réalisation de n variables aléatoires indépendantes Z_i de loi de Bernoulli d'espérance $\pi(\mathrm{x}_i) = P(Z_i = 1; \mathrm{x}_i)$:

$$Z_i \sim \mathcal{B}(1, \pi(\mathrm{x}_i)); i = 1, \ldots, n.$$

La modélisation linéaire

$$\pi(\mathrm{x}_i) = \mathrm{x}_i\theta$$

n'est pas adaptée, car elle ne contraint pas π à appartenir à $[0; 1]$. En revanche, la fonction `logit` par exemple, bijective de $]0, 1[$ sur $]-\infty, +\infty[$, permet de prendre en compte cette contrainte :

$$\mathrm{logit}(\pi(\mathrm{x}_i)) = \log\left(\frac{\pi(\mathrm{x}_i)}{1 - \pi(\mathrm{x}_i)}\right).$$

La régression logistique postule alors que dans l'échelle de la fonction `logit`, la variation de l'espérance est linéaire en un paramètre θ

$$\mathrm{logit}(\pi(\mathrm{x}_i)) = \mathrm{x}_i\theta.$$

Cette hypothèse permet donc de définir une variation non linéaire de l'espérance $\pi(\mathrm{x}_i)$ en fonction d'un régresseur linéaire $\mathrm{x}_i\theta$ des covariables. La fonction `logit` est appelée **fonction de lien**. Il existe d'autres fonctions de lien (voir Section 8.2.1), mais nous verrons que la fonction `logit` se distingue des autres pour la régression logistique et est appelée fonction de lien canonique.

Remarque Comme en régression, nous travaillons conditionnellement aux covariables, et nous y intégrons l'intercept. Ainsi, la première colonne de la matrice du plan d'expérience X (dont chaque ligne i est le vecteur ligne x_i des covariables pour l'observation i) est une colonne de 1. Toutes les variables étant ici quantitatives, la dimension du modèle est donc égale au nombre de variables plus un : $p = 15 + 1 = 16$.

8.1.2 Cancer de l'œsophage

Les données de l'exemple précédent sont individuelles : les conditions d'expérience sont différentes pour chaque éprouvette. Dans le cas présenté maintenant, plusieurs individus sont observés sous les mêmes conditions, et on ne retient que le nombre global de succès (et d'échec) : c'est un cas de données groupées ou répétées. Breslow et Day (1980) ont procédé à une étude cas-contrôle pour déterminer l'impact combiné de l'âge, de la consommation d'alcool et du tabac sur l'apparition du cancer de l'œsophage [2]. L'étude est menée sur 200 hommes ayant reçu un diagnostic de cancer de l'œsophage dans l'un des hôpitaux régionaux d'Ille-et-Villaine entre janvier 1972 et avril 1974. Un échantillon de contrôle est constitué de 775 hommes adultes tirés aléatoirement à partir des listes électorales des communes. Les variables explicatives sont l'âge `agegp` codé sur six niveaux, la consommation moyenne d'alcool `alc` en gr/l, et la consommation de tabac `tobgp`, codée sur quatre niveaux. Le nombre `ncases` de personnes de l'échantillon des cas diagnostiqués et le nombre de personnes `ncontrols` de l'échantillon de contrôle sont enregistrés pour chaque combinaison des variables explicatives.

Le fichier complet comporte autant de lignes qu'il y a de groupes d'individus. La variable à expliquer Z est bien binaire (*cancer* ou *indemne*), mais nous n'avons plus accès aux information individuelles, qui ont été globalisées au niveau du groupe. Ainsi, nous choisissons de modéliser

2. http ://www.iarc.fr/en/publications/pdfs-online/stat/sp32/

cette expérience par K variables aléatoires indépendantes Y_k de loi binomiale d'espérance π_k, agissant sur un effectif n_k, somme du nombre de cas *indemne* et du nombre de cas de *cancer* pour un groupe donné :

$$Y_k \sim \mathcal{B}(n_k, \pi_k); k = 1, \ldots, K$$

L'espérance de la loi binomiale vaut $\mu_k = n_k\pi_k$, et nous supposons qu'elle dépend des valeurs des variables explicatives du groupe k :

$$\pi_k = \pi(\mathrm{x}_k) = P(Z = cancer; \mathrm{x}_k).$$

Nous terminons la définition du modèle par le choix d'une fonction de lien entre l'espérance de la loi de Y_k et le régresseur linéaire $\mathrm{x}_k\theta$, par exemple la fonction

$$g(\mu_k) = \log\left(\frac{\mu_k}{n_k - \mu_k}\right) = \mathrm{x}_k\theta,$$

ce qui revient à utiliser la fonction `logit` sur la variable individuelle Z :

$$g(\mu_k) = \mathrm{logit}(\pi_k) = \log\left(\frac{\pi_k}{1 - \pi_k}\right) = \mathrm{x}_k\theta. \tag{8.1}$$

Le modèle ainsi défini comporte deux variables explicatives qualitatives (niveau d'âge et niveau de consommation d'alcool), et une variable quantitative (consommation moyenne). Sa dimension est donc $1 + (6 - 1) + (4 - 1) + 1 = 10$ sans tenir compte des interactions.

Remarque : La taille de l'échantillon de contrôle a été fixée arbitrairement et a posteriori. Ce type d'étude est appelée étude rétrospective (*case-control study* en anglais), et est très souvent utilisée dans les applications médicales : elle permet en particulier d'assurer un échantillon suffisamment large de sujets malades à étudier, et proportionnellement plus important que celui auquel permettrait d'accéder un échantillonnage aléatoire dans toute la population. Ainsi, la probabilité de déclarer un cancer de l'œsophage dans la population d'Ille-et-Villaine ayant les caractéristiques du groupe k est-elle bien inférieure à la probabilité π_k ici modélisée.

8.2 Définition

La loi de Bernoulli étant un cas particulier de la loi binomiale ($n_k = 1$), les deux exemples précédents peuvent être englobés dans un même modèle de réponse binaire appelé régression logistique.

Définition 15. *On appelle* ***régression logistique*** *un modèle de variables* ***indépendantes*** *de loi* ***binomiale*** *dont l'****espérance*** *est une fonction* ***non linéaire*** *d'un* ***régresseur linéaire****. Le régresseur linéaire est donc une fonction non linéaire de l'espérance, appelée* ***fonction de lien****.*

Rappelons la définition de la loi binomiale : si $Y \sim \mathcal{B}(n, \pi)$, alors

$$\mathbb{P}(Y = y) = \left(\begin{array}{c} n \\ y \end{array}\right) \pi^y (1 - \pi)^{n-y}$$

qui peut s'écrire

$$\begin{aligned} \mathbb{P}(Y = y) &= \exp\left(y \log\frac{\pi}{1 - \pi} + n \log(1 - \pi) + \log\left(\begin{array}{c} n \\ y \end{array}\right)\right) \\ &= \exp\left[y\, a(\mathbb{E}(Y)) + b(\mathbb{E}(Y)) + c(y)\right], \end{aligned} \tag{8.2}$$

où l'on reconnaît dans (8.2) la forme caractéristique d'une famille exponentielle de loi. Nous remarquons que la fonction de lien `logit` choisie dans (8.1) apparaît naturellement dans l'équation (8.2) où $a = \text{logit}$, et cette fonction de lien est appelée **canonique**.

La régression logistique remplit les conditions d'un m**odèle linéaire généralisé** : des observations **indépendantes**, de loi appartenant à une **famille exponentielle**, dont l'espérance dépend d'un **régresseur linéaire** au travers d'une **fonction de lien non linéaire**. La régression logistique se situe donc entre la régression linéaire dont elle reprend un régresseur linéaire en un paramètre θ et la régression non linéaire, dont elle reprend une modélisation non linéaire (et un peu particulière) de l'espérance en fonction de θ. De plus, le schéma de bruit n'est plus additif : la loi de l'observation a été modélisée de façon directe.

8.2.1 Les fonctions de lien

Toute fonction assurant la bijectivité entre le domaine de définition de l'espérance de la loi et $\mathbb{R}$, domaine de définition du régresseur, peut être utilisée comme fonction de lien. Ainsi, les fonctions de lien en régression logistique sont des bijections de $]0,1[$ sur $]-\infty,+\infty[$ de la forme :

$$F^{-1}(\pi(\mathrm{x})) = \mathrm{x}\theta,$$

où $F(x) = \int_{-\infty}^{x} f(s)ds$ est une fonction de répartition de densité f appelée distribution de tolérance.

Remarquons que la fonction `logit` est dans ce cas : c'est l'inverse de la fonction de répartition de la loi logistique, dont la densité s'écrit :

$$f(x) = \frac{e^x}{(1+e^x)^2}$$

Elle possède un centre de symétrie : $\text{logit}(\pi) = -\text{logit}(1-\pi)$. De plus, elle permet une interprétation simple de la notion de rapport de cotes (*odds ratio*) que nous verrons section 8.3.6.

La fonction `probit` est l'inverse de la fonction de répartition $F_{\mathcal{N}(0,1)}$ de la loi gaussienne centrée réduite, et possède également un centre de symétrie. Cette fonction est particulièrement adaptée à la situation où la variable binaire Z est déterminée par une variable latente (ou cachée) gaussienne T qui s'explique linéairement, et que l'on seuille :

$$T = \mathrm{x}\theta + \epsilon, \epsilon \sim \mathcal{N}(0,\sigma^2), \quad Z = \mathbb{1}_{T \leq s}$$

Alors, $\mathbb{P}(Z=1) = \mathbb{P}(T \leq s) = F_{\mathcal{N}(0,1)}((s - \mathrm{x}\theta)/\sigma)$.

Une autre fonction de lien souvent utilisée est la fonction `cloglog`, inverse de la loi de Weibull,

$$\text{cloglog}(\pi) = \log(-\log(1-\pi))$$

qui ne possède pas de centre de symétrie (Figure 8.1), et qui permet parfois de mieux représenter les données.

Les fonctions `logit` et `probit` sont quasiment linéaires sur l'intervalle $0.1 \leq \pi \leq 0.9$. Pour les petites valeurs de la probabilité π, la fonction `cloglog` est proche de la fonction `logit`, toutes les deux étant équivalentes à la fonction `log`. Quand π tend vers 1, la fonction `cloglog` tend vers l'infini beaucoup plus lentement que les fonctions `logit` ou `probit`.

8.2.2 Représentations graphiques

Lorsque le cas d'étude concerne des données groupées, il est possible d'estimer, pour chaque groupe, la probabilité $\pi = P(Z=1)$ par la fréquence observée de cet événement dans le groupe.

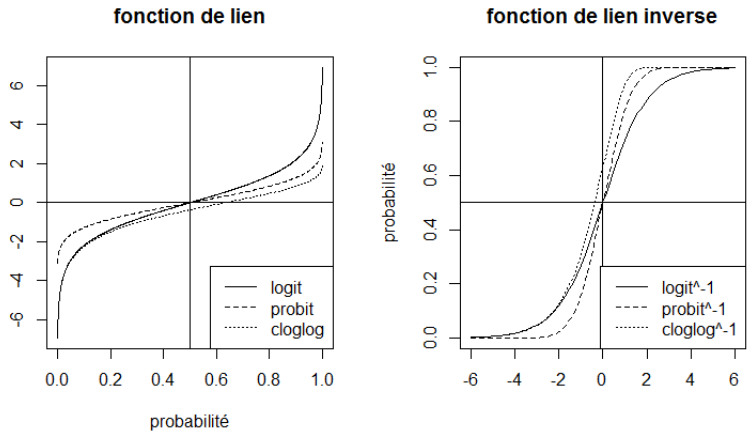


FIGURE 8.1 – Quelques exemples de fonctions de lien et leur inverse

Ainsi, dans l'exemple du cancer de l'œsophage, nous traçons les fréquences observées en fonction de la consommation d'alcool. Dans ce jeu de données, il y a plusieurs groupes observés par niveau de consommation (Figure 8.2), et la fréquence moyenne est également reportée.

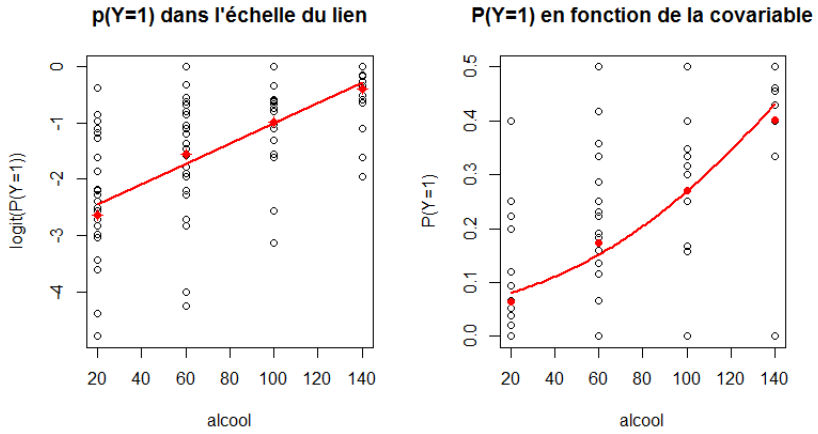


FIGURE 8.2 – Fréquence observée de l'apparition de cancer en fonction de la consommation d'alcool (à gauche), représentation faite dans l'échelle du lien (à droite)

La consommation d'alcool est une variable quantitative c, le lien entre la probabilité prédite et le régresseur linéaire s'écrit

$$\text{logit}(\pi(c)) = \begin{pmatrix} 1 & c \end{pmatrix} \theta = \theta_1 + c\theta_2$$

soit

$$\pi(c) = \frac{\exp(\theta_1 + c\theta_2)}{1 + \exp(\theta_1 + c\theta_2)}.$$

La probabilité prédite suit donc une courbe non linéaire. Il est également possible de représenter ces observations dans l'échelle du lien, c'est à dire représenter logit(π) en fonction de la covariable. La fonction logit empirique permet d'éviter les problèmes liés au calcul du logit pour des fréquences observées nulles :

$$\text{logit.emp}(ncases, ncontrols) = \log \frac{ncases + 0.5}{ncontrols + 0.5}.$$

La représentation graphique de la prédiction dans l'échelle du lien est maintenant linéaire : elle est en effet directement calculée par le prédicteur linéaire (figure 8.2).

Les fréquences observées ne sont pas accessibles dans le cas de données individuelles. Mais il est cependant possible de représenter la valeur de la réponse (0 ou 1) en fonction de la covariable. Dans l'exemple de fragilité de l'alliage, la réponse est tracée en fonction de la teneur en fer et chrome (`FeCr.MC`) du mélange initial. Aucun cas de fragilité (Y=1) n'est observé en dessous d'un seuil de 150. Cependant, le graphe n'est pas parlant pour les observations dont la teneur est supérieure à ce seuil. La figure 8.3 affiche de plus la probabilité prédite avec une fonction de lien `logit` ou une fonction de lien `probit`. Bien que les données de fragilité aient été obtenues par seuillage, il y a peu de différence entre ces deux modélisations.

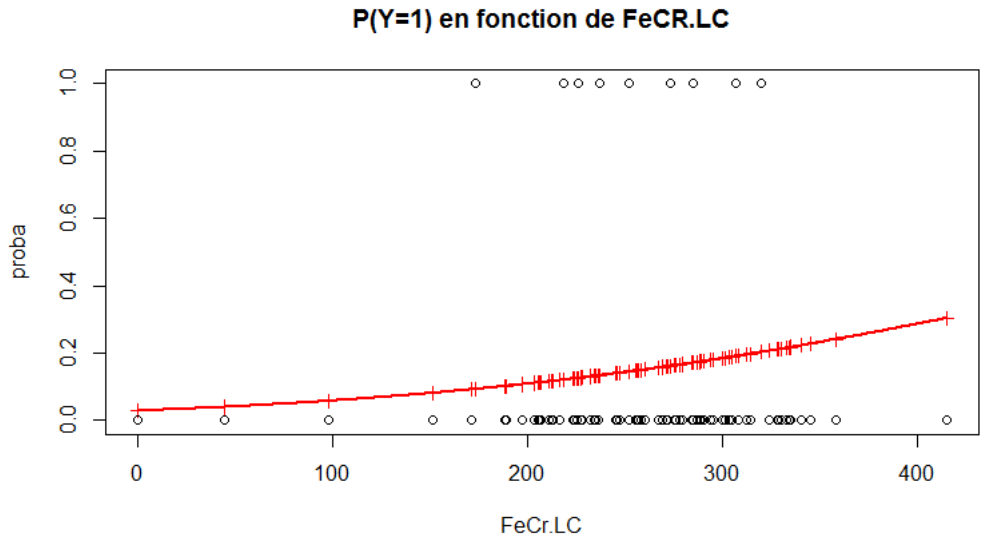


FIGURE 8.3 – Taux de pièces défectueuses en fonction de la concentration de FeCR.LC

8.3 Estimation

Ayant défini la loi des observations, il est naturel d'estimer les paramètres du modèle de régression logistique par maximum de vraisemblance, en suivant les étapes classiques : écriture de la vraisemblance, calcul des équations normales, résolution numérique, étude des propriétés asymptotiques.

8.3.1 Vraisemblance

La log-vraisemblance en données groupées $\mathbf{y} = (Y_1, \ldots, Y_k)$ s'écrit

$$l_n(\theta; \mathbf{y}) = \sum_{k=1}^{K} \log \mathbb{P}(Y_k)$$

en tenant compte de l'indépendance des observations, d'où

$$l_n(\theta; \mathbf{y}) = \sum_{k=1}^{K} y_k \log\left(\frac{\pi_k}{1-\pi_k}\right) + \sum_{k=1}^{K} n_k \log(1-\pi_k) + \sum_{k=1}^{K} \left(\begin{array}{c} n_k \\ y_k \end{array} \right). \tag{8.3}$$

Si la fonction de lien est la fonction `logit`, l'expression se simplifie

$$l_n(\theta; \mathbf{y}) = (\sum_{k=1}^{K} y_k \mathrm{x}_k)\theta - \sum_{k=1} n_k \log(1 + e^{\mathrm{x}_k\theta}) + c(\mathbf{y}).$$

Il apparaît :
- un terme **linéaire** en θ : $(\sum_k y_k \mathrm{x}_k)\theta$,
- un terme **non linéaire déterministe** : $b(\theta) = \sum_k n_k \log(1 + e^{\mathrm{x}_k\theta})$,
- un terme ne dépendant que des observations : $c(\mathbf{y})$.

La vraisemblance a été linéarisée, à la constante de normalisation $b(\theta)$ près. La partie linéaire apparaît grâce au choix de la fonction de lien canonique de la loi binomiale. Si une autre fonction de lien g est choisie, la log-vraisemblance ne se linéarise plus et s'écrit sous la forme :

$$l_n(\theta; \mathbf{y}) = \sum_{k=1}^{K} y_k \log\left(g^{-1}(\mathrm{x}_k\theta)\right) + \sum_{k=1}^{K} (n_k - y_k) \log(1 - g^{-1}(\mathrm{x}_k\theta)) + c(\mathbf{y})$$

La loi de Bernoulli étant un cas particulier de la loi binomiale ($n = 1$), les expressions trouvées s'appliquent au cas des données individuelles en remplaçant l'indice k par i, la valeur K par n et en imposant $n_i = 1$.

8.3.2 Estimateur du maximum de vraisemblance

Le principe de l'estimation par maximum de vraisemblance est la recherche l'estimateur $\widehat{\theta}$ qui rend la vraisemblance (et donc la log-vraisemblance) des observations la plus grande possible. C'est un problème d'optimisation, qui peut se résoudre en cherchant $\widehat{\theta}$ annulant la dérivée $U(\widehat{\theta})$ (appelée score) de la log-vraisemblance :

$$U(\widehat{\theta}) = \frac{\partial l_n(\theta; \mathbf{y})}{\partial \theta_j}_{|\theta=\widehat{\theta}} = 0.$$

Son existence est montrée dans le cadre général des modèles linéaires généralisés Fahrmeir et Kaufmann (1985). En particulier, si le modèle est identifiable (matrice du plan d'expérience X injective, comme dans le modèle linéaire), si la fonction de lien est log-concave (ce qui est le cas pour les trois fonctions présentées en section 8.2.1), et si pour chaque groupe k, $0 < y_k < n_k$, alors $\widehat{\theta}$ existe, est fini et la log-vraisemblance a un unique maximum $\widehat{\theta}$ consistant en $\theta^\star$ la vraie valeur du paramètre. L'argument de compacité de l'espace des paramètre utilisé en régression non-linéaire est remplacé ici par un argument de convexité du contraste dans cette famille de modèles.

Les composantes du vecteur du **score** (dérivée première de la log-vraisemblance par rapport à θ) et de la matrice du **hessien** (dérivée seconde) s'écrivent, dans le cas du lien `logit`,

$$U_j(\theta) = \frac{\partial l_n(\theta;\mathbf{y})}{\partial \theta_j} = \sum_{k=1}^{K} y_k x_{kj} - \sum_{k=1}^{K} n_k x_{kj} \frac{e^{\mathrm{x}_k\theta}}{1+e^{\mathrm{x}_k\theta}} = \sum_{k=1}^{K} x_{kj}(y_k - n_k\pi_k); \quad j=1,\ldots,p$$

$$H_{jl}(\theta) = \frac{\partial^2 l_n(\theta;\mathbf{y})}{\partial \theta_j \partial \theta_l} = -\sum_{k=1}^{K} n_k x_{kj} x_{kl} \frac{e^{x_k'\theta}}{(1+e^{x_k'\theta})^2} = -\sum_{k=1}^{K} x_{kj} x_{kl} n_k \pi_k (1-\pi_k); \quad j,l=1,\ldots,p$$

ou, sous forme matricielle

$$U(\theta) = X'WY^*, \quad H(\theta) = -X'WX \tag{8.4}$$

avec W, matrice diagonale de coefficient

$$\mathrm{var}(Y_k) = n_k\pi_k(1-\pi_k),$$

et Y^* le vecteur de composantes

$$Y_k^* = (Y_k - n_k\pi_k)/\,\mathrm{var}(Y_k).$$

Le hessien dans ce cas est déterministe et égal est l'opposé de l'**information de Fisher** de l'échantillon, définie comme la matrice de covariance du score :

$$H(\theta) = -I_n(\theta) = -\mathbb{E}\left[U(\theta)U(\theta)'\right]. \tag{8.5}$$

Cette propriété est spécifique au cas canonique. Dans le cas général, le hessien prend la forme $-X'\widetilde{W}X$ où $\widetilde{W}$ est une matrice aléatoire d'espérance W.

8.3.3 Résolution numérique

L'estimateur du maximum de vraisemblance peut être calculé par une méthode de Newton, soit en utilisant le hessien, soit en utilisant la matrice d'information de Fisher comme approximation du hessien (méthode du score de Fisher) et ces deux méthodes sont bien évidemment identiques dans le cas de la fonction de lien canonique `logit`.

Méthode de Newton

Au voisinage d'une approximation numérique $\theta^{(m)}$ de $\widehat{\theta}$, on choisit $\theta^{(m+1)}$ qui annule l'approximation linéaire de $U(\theta)$ donnée par son développement de Taylor au premier ordre :

$$U(\theta^{(m)}) + \frac{\partial U(\theta^{(m)})}{\partial \theta}(\theta^{(m+1)} - \theta^{(m)}) = 0,$$

soit

$$\theta^{(m+1)} = \theta^{(m)} - H^{-1}(\theta^{(m)})U(\theta^{(m)}).$$

L'algorithme est itéré à partir de cette nouvelle valeur, et ainsi de suite jusqu'à convergence.

Lorsque la vraisemblance est strictement concave, cette méthode converge vers $\widehat{\theta}$, quelle que soit l'initialisation. Le calcul du hessien permet de définir la direction de descente dans l'optimisation numérique permettant le calcul de l'EMV. Il est déterministe dans le cas de la fonction de lien canonique ($g = a$).

Algorithme de scoring de Fisher

Quand le hessien est aléatoire ($a \neq g$), il est possible de l'approximer par l'opposé de l'information de Fisher, qui est prise comme direction de descente. Il en résulte l'itération suivante de algorithme de scoring de Fisher :

$$\theta^{(m+1)} = \theta^{(m)} + I_n^{-1}(\theta^{(m)})U(\theta^{(m)}). \tag{8.6}$$

Cette méthode est bien sûr identique à celle de Newton dans le cas canonique, puisqu'alors $I_n = -H_n$. Les deux méthodes convergent et fournissent une estimation de la variance de $\widehat{\theta}$: $-H^{-1}(\widehat{\theta})$ ou $I_n^{-1}(\widehat{\theta})$.

Algorithme des moindres carrés pondérés itératifs

En utilisant (8.5), l'itération de scoring de Fisher (8.6) peut s'écrire :

$$\begin{aligned}
\theta^{(m+1)} &= \theta^{(m)} + (\mathrm{X}'W_m\mathrm{X})^{-1}\mathrm{X}'W_m\mathrm{Y}_m^* \\
&= (\mathrm{X}'W_m\mathrm{X})^{-1}(\mathrm{X}'W_m\mathrm{X}\theta^{(m)} + \mathrm{X}'W_m\mathrm{Y}_m^*) \\
&= (\mathrm{X}'W_m\mathrm{X})^{-1}\mathrm{X}'W_m(\mathrm{X}\theta^{(m)} + \mathrm{Y}_m^*) \\
&= (\mathrm{X}'W_m\mathrm{X})^{-1}\mathrm{X}'W_m\mathrm{Z}_m.
\end{aligned} \tag{8.7}$$

Ce qui permet de voir $\theta^{(m+1)}$ comme solution d'une procédure des **moindres carrés pondérés** dans le modèle linéaire suivant :

$$\mathrm{Z}_m = \mathrm{X}\theta + \varepsilon_m \tag{8.8}$$

pour lequel $\mathbb{E}(\varepsilon_m) = 0$ et $\mathrm{var}(\varepsilon_m) = W_m^{-1}$. Ainsi, $\widehat{\theta}$ solution de $U(\widehat{\theta}) = 0$ peut être obtenu comme la solution d'une procédure itérative des moindres carrés pondérés dans les modèles de régression linéaire successifs $Z_m = \mathrm{X}\theta + \varepsilon_m$, $m = 1, 2, \ldots$.

Cet algorithme peut être interprété comme une optimisation alternée de l'espérance et la variance : à $W^{(m)}$ fixé, c'est à dire si la variance est considérée comme connue, on cherche $\mu^{(m+1)}$. Le raffinement de $\mu^{(m+1)}$ permet d'adapter W, puis de mettre à jour μ jusqu'à convergence. C'est une procédure des **moindres carrés pondérés itératifs**.

Conditions initiales

Contrairement à la régression non-linéaire, le choix de la valeur initiale n'est pas critique en régression logistique, en particulier sous le lien canonique, puisque la vraisemblance est concave ; cependant, un bon choix de valeur initiale peut réduire le nombre de cycles de l'algorithme.

Qualité de l'estimation

Après obtention de la valeur de l'estimateur, l'utilisation de la fonction de lien inverse,

$$\pi_k = \frac{e^{\mathrm{x}_k\theta}}{1 + e^{\mathrm{x}_k\theta}} = \frac{e^{\theta_1 + c\theta_2}}{1 + e^{\theta_1 + c\theta_2}} \tag{8.9}$$

permet de tracer la courbe des probabilités estimées en fonction de c que nous avons déjà visualisées Figure 8.2.

Il est rare d'obtenir des défauts de convergence sauf si l'une des composantes de $\widehat{\theta}$ est infinie, ce qui peut arriver quand certaines probabilités sont estimées à zéro ou à un. Dans ce cas,

	estimation	écart-type
(Intercept)	-3,4637	73,9645
Power	0,7885	0,1986
Vac1	-1,1701	0,3808
FeSi.HP	-0,3829	0,2162
FeCr.LC	1,1264	0,4073
FeV	-0,9802	0,2659
Ni	0,3787	0,2442
CASI.WI	-3,9051	407,9807
Total.Slag	-0,5942	0,2433

TABLE 8.1 – Exemple de défaut d'estimation sur les données de fragilité

W n'est plus inversible, et $X'WX$ non plus. Il y a une non identifiabilité sur la position de séparation dans le processus de calcul par MCPI. Une convergence anormale signifie, soit que la log-vraisemblance est très plate, soit que la log-vraisemblance possède une asymptote. Dans ce cas, ni les paramètres, ni leur variabilité ne peuvent être estimés de façon fiable.

Des écarts-types anormalement élevés doivent alerter sur la configuration des données étudiées. Prenons l'exemple des données de fragilité, et effectuons une régression logistique sur les variables `Power`, `Vac1`, `FeSi.HP`, `FeCr.LC`, `FeV`, `Ni`, `CASI.WI` et `Total.Slag` par exemple. Les valeurs estimées des paramètres correspondants sont reportées dans la table 8.1. L'intercept et la variable `CASI.WI` présentent un écart-type très important proportionnellement à la valeur estimée, alors qu'aucun diagnostic n'a été émis sur un problème de convergence numérique. L'examen de la variable `CASI.WI` montre qu'elle vaut 25 sauf en de rares exceptions. Ainsi, cette variable est très proche de l'intercept, et même si le modèle est théoriquement identifiable, il est proche de la non identifiabilité : ceci se traduit par une très mauvaise estimation des composantes du paramètre associées à ces deux variables. Dans cette étude, il conviendra donc de ne plus inclure la variable `CASI.WI`.

8.3.4 Loi de l'estimateur et intervalle de confiance du paramètre

Comme en régression non-linéaire, $\widehat{\theta}$ est un estimateur biaisé de $\theta^\star$ pour un n donné. Mais la méthode du maximum de vraisemblance produit des estimateurs *asymptotiquement* non biaisés et de variance minimum.

En particulier, sous certaines conditions du plan d'expérience, l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ du modèle de régression logistique est consistant et possède un comportement asymptotiquement normal. Ces résultats ont été montrés par Fahrmeir et Kaufmann (1985), en suivant le paradigme de preuve d'un estimateur du moindre contraste (ici, la fonction de contraste est le logarithme de la vraisemblance) :

$$I_n^{1/2}(\theta^\star)(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p),$$

où $\theta^\star$ désigne la vraie valeur (inconnue) de θ. En utilisant le théorème de Slutsky, si $\widehat{V}_n$ est un estimateur consistant de $I_n(\theta^\star)$,

$$\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p).$$

Cette propriété permet de déduire une loi approchée de l'EMV à partir d'un estimateur $\widehat{V}_n$ du terme de normalisation $I_n(\theta^\star)$:

— quand l'information de Fisher est facile à calculer, $I_n(\theta^\star)$ est estimée par $I_n(\widehat{\theta})$ (méthode *plug-in*). C'est en particulier le cas lorsque pour une fonction de lien canonique, puisque le hessien est alors déterministe et $I_n(\theta) = -H(\theta)$
— quand ce n'est pas le cas, elle est remplacée par $-H(.)$, estimateur consistant, et $I_n(\theta^\star)$ est encore estimée par $-H(\widehat{\theta})$.

Ce sont ces propriétés asymptotiques qui justifient l'*optimalité* réputée de la méthode du maximum de vraisemblance ; elles garantissent une précision optimale des estimateurs, dès lors que n est suffisamment grand.

Proposition 5 (Loi approchée de l'EMV)**.** *Pour n suffisamment grand, la loi de $\widehat{\theta}$ est approchée par une loi gaussienne multivariée :*

$$\widehat{\theta} \overset{appr}{\sim} \mathcal{N}(\theta^\star, \widehat{V}_n),$$

où $\widehat{V}_n$ est un estimateur consistant de la matrice de variance-covariance de la loi de $\widehat{\theta}$, par exemple

$$\widehat{V}_n = \left(-H(\widehat{\theta})\right)^{-1}.$$

Rappelons que les algorithmes de calcul du maximum fournissent une estimation de $\widehat{V}_n$.

Ces informations permettent de construire des intervalles de confiance du paramètre θ. La table 8.2 donne les valeurs estimées du paramètre et l'écart-type de l'estimation dans le cas des données de cancer avec la variable explicative alcool :

Paramètre	Valeur estimée	écart type
Intercept	-2,8035	0,1713
alcool	0,0181	0,00200

TABLE 8.2 – Exemple des données cancer : estimations et écart-types

L'information apportée par la variabilité de l'estimation est importante. En effet, nous pourrions hâtivement penser que l'alcool n'est pas un facteur significatif, car la valeur du paramètre estimé est proche de zéro. Cependant, elle est estimée avec une grande précision. Ainsi un intervalle de confiance (approché) de θ_2 de niveau 95% est

$$\left[\widehat{\theta}_2 - q_{\mathcal{N}(0,1)}(0,95)\sqrt{(\widehat{V}_n)_{22}}; \widehat{\theta}_2 + q_{\mathcal{N}(0,1)}(0,95)\sqrt{(\widehat{V}_n)_{22}}\right]$$

ce qui donne $[0,014; 1,978]$ sur le jeu de données observées. 0 n'appartient pas à cet intervalle, l'alcool est significatif (au risque 5%) dans l'apparition du cancer de l'œsophage.

8.3.5 Prévision de la probabilité sous une condition donnée

En utilisant la fonction inverse du lien logit (8.9), il est possible d'estimer la probabilité d'apparition du cancer en fonction de la consommation c_0

$$\widehat{\pi}_k(c_0) = \frac{e^{\widehat{\theta}_1 + c_0\widehat{\theta}_2}}{1 + e^{\widehat{\theta}_1 + c_0\widehat{\theta}_2}}$$

et d'en calculer un intervalle de confiance de niveau approché $1-\alpha$. Celui-ci est l'inverse de la fonction logit (8.9), strictement monotone, appliquée aux bornes d'un intervalle de confiance de

$\mathrm{x}_0\theta = \theta_1 + c_0\theta_2$ de niveau approché $1-\alpha$:

$$\left[\mathrm{x}_0\widehat{\theta} - q_{\mathcal{N}(0,1)}(1-\frac{\alpha}{2})\sqrt{\mathrm{x}_0\widehat{V}_n\mathrm{x}_0'}; \mathrm{x}_0\widehat{\theta} + q_{\mathcal{N}(0,1)}(1-\frac{\alpha}{2})\sqrt{\mathrm{x}_0\widehat{V}_n\mathrm{x}_0'}.\right]$$

Les valeurs de intervalle de confiance de niveau 95% dans l'échelle du lien logit et dans l'échelle de la probabilité pour les consommations observées sont visualisées dans la table 8.3.

c0	min régresseur	max régresseur	min probabilité	max probabilité
20	-2,7117542	-2,17321021	0,06228332	0,1021821
40	-2,2925413	-1,87036312	0,09174257	0,1334997
60	-1,8903484	-1,55049603	0,13120475	0,1750146
80	-1,5189425	-1,19984187	0,17961729	0,2315033
100	-1,1835724	-0,81315200	0,23441048	0,3072192
120	-0,8727349	-0,40192949	0,29468555	0,4008488
140	-0,5742806	0,02167625	0,36024969	0,5054189

TABLE 8.3 – Exemple des données cancer : intervalles de confiance

De la même façon, il est possible de prévoir la probabilité d'apparition du cancer pour des consommations non observées dans le jeu de données. La figure 8.4 représente la forme des intervalles de confiance individuels pour différentes valeur de consommation d'alcool : dans l'échelle du lien, le prédicteur $\mathrm{x}\theta$ est linéaire en la consommation c, la variabilité de l'estimation dépend de la valeur de c. En effet

$$\mathrm{x}_0\widehat{V}_n\mathrm{x}_0' = (\widehat{V}_n)_{11} + 2c_0(\widehat{V}_n)_{12} + c_0^2(\widehat{V}_n)_{12}, \quad \text{avec } \mathrm{x}_0 = \begin{pmatrix} 1 & c \end{pmatrix}.$$

Les intervalles de confiance sont centrés autour de la valeur observée du prédicteur, mais non centrés autour de la fréquence observée dans l'échelle de la probabilité : c'est une conséquence de la non linéarité de la fonction de lien.

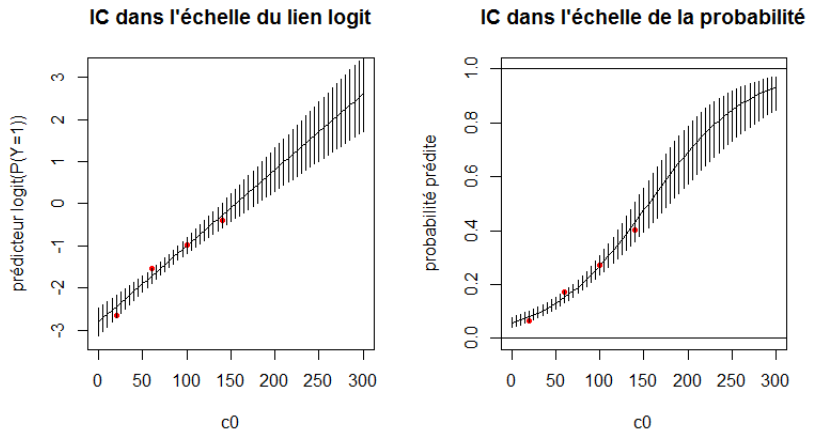


FIGURE 8.4 – Fréquence observée de l'apparition de cancer en fonction de la consommation d'alcool représentée dans l'échelle du lien, et dans l'échelle de la probabilité. Les triangles représentent l'incidence de cancer sur la population observée

Le calcul de l'intervalle de confiance de la probabilité comme image de la fonction de lien inverse permet d'imposer aux bornes de l'intervalle d'être contenues dans l'intervalle $[0;1]$. Le recours à la delta méthode (section 7.4.1) serait ici maladroit : en plus d'une approximation induite par la linéarisation de la log vraisemblance, la delta méthode calcule des intervalles de confiance centrés en $\widehat{\theta}$ dont les bornes pourraient sortir de l'intervalle $[0;1]$.

Remarque : La prévision de la probabilité π extrapolée à des valeurs de covariables en dehors du domaine d'observation est un exercice délicat et à déconseiller. En effet, celle-ci dépend fortement de la forme de la fonction de lien du modèle supposé. Il est commun de trouver deux modèles ayant des prévisions similaires sur le domaine d'observation, mais se comportant de façon bien différente lors de l'extrapolation. Ici de plus, des taux de consommation supérieurs à la dose léthale ne peuvent pas être observés, la prévision n'a alors plus de sens. Il convient donc de limiter les prévisions au domaine d'observation (voir chapitre 4 p. 123 de McCullagh et Nelder (1989) pour approfondir le traitement des cas d'extrapolation extrême).

Remarque : dans une étude de cas-contrôle, cette probabilité n'est pas directement utile, puisqu'elle ne correspond pas à la probabilité observée dans la population générale. Il est beaucoup plus intéressant d'utiliser la notion d'odds ratio, qui ne dépend pas du nombre relatif de cas et de contrôle, et qui permet de répondre à la question de l'influence des facteurs explicatifs sur l'apparition du cancer.

8.3.6 Odds ratio

L'utilisation de la fonction de lien `logit` permet de présenter une interprétation simple de l'ordre de grandeur des effets mesurés. Soit $Od(\mathrm{x})$ le rapport de la probabilité de succès $P(Z = 1;\mathrm{x})$ sur la probabilité d'échec $P(Z = 0;\mathrm{x})$:

$$Od(\mathrm{x}) = \frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})}$$

Par référence à un pari au tiercé, $Od(\mathrm{x})$ est appelée **cote** (**odds** en anglais), ou probabilité relative, de l'événement Z sous les conditions d'expérience x.

Définition 16. *On appelle* ***rapport de cotes*** *ou* ***odds ratio*** *le rapport de deux cotes calculées dans des conditions d'expérience différentes*

$$OR(\mathrm{x}_2/\mathrm{x}_1) = \frac{Od(\mathrm{x}_2)}{Od(\mathrm{x}_1)} = \frac{\pi(\mathrm{x}_2)/(1-\pi(\mathrm{x}_2))}{\pi(\mathrm{x}_1)/(1-\pi(\mathrm{x}_1))}.$$

Lorsque le modèle logistique est défini avec le lien canonique logit, la cote est le régresseur linéaire, et le rapport de cotes s'écrit :

$$OR(\mathrm{x}_2/\mathrm{x}_1) = \frac{Od(\mathrm{x}_2)}{Od(\mathrm{x}_1)} = e^{(\mathrm{x}_2-\mathrm{x}_1)\theta}$$

Ainsi dans ce cas, la modification d'une variable quantitative x_j d'une unité ($x_{2j} = x_{1j} + 1$) a pour effet de modifier le logarithme de la cote d'une quantité additive θ_j, et ce quelle que soit la valeur initiale de x_j, ou de façon équivalente, modifie le rapport de cotes d'un facteur multiplicatif e^{θ_j}

$$OR(x_j) = e^{\theta_j}.$$

Il est important de noter que toutes les autres variables restent fixées pour définir le rapport de cotes de la variable x_j. Le rapport de cotes associé à la consommation d'alcool est estimé par $= e^{\widehat{\theta_2}} = e^{0,018051}$, un intervalle de confiance de niveau 95% est donné par $[e^{0,014}, e^{1,978}] =$

$[1,014; 7,228]$. La hausse d'un gramme par jour de consommation multiplie la probabilité relative de déclarer un cancer de l'œsophage par un facteur pouvant aller jusqu'à 7.

Si, de plus, la probabilité de l'événement de succès est petite pour chacune des deux cotes comparées, alors $1 - \pi(\mathrm{x}_1) \simeq 1$ et $1 - \pi(\mathrm{x}_2) \simeq 1$, d'où

$$OR(x_j) = e^{\theta_j} \simeq \frac{\pi(\mathrm{x}_2)}{\pi(\mathrm{x}_1)}$$

Le facteur multiplicatif s'applique alors directement aux probabilités, et non plus seulement aux cotes. Si la consommation d'alcool est inférieure à 30 g/jour, alors une hausse de consommation de 1g/jour multiplie la probabilité de déclarer un cancer par un facteur pouvant aller jusqu'à 7.

variable explicative qualitative Il s'agit d'étudier la réponse suivant deux niveaux différents du facteur, les valeurs des autres variables étant identiques. Suivant la condition d'identifiabilité utilisée pour estimer le modèle, par exemple CI contrôle (1er niveau) / traitement :

— pour comparer par rapport au contrôle

$$OR(niveau\ j/contrôle) = e^{\theta_j}$$

— pour comparer deux niveaux j et ℓ dont aucun n'est contrôle

$$OR(niveau\ \ell/niveau\ j) = e^{\theta_\ell - \theta_j}$$

— situations so

Dans le cas de l'exemple, la variable alcool de l'exemple des données de cancer a été traitée comme variable quantitative. C'est un choix de modélisation qui peut s'avérer discutable. D'une part, parce qu'il est fort peu réalisable d'obtenir un échantillon de personnes ayant une consommation réglée au gr/jour près, d'autre part parce qu'on a imposé une contrainte de linéarité via le régresseur. L'étude initiale dont est tiré cet exemple avait traité l'alcool en tant que variable qualitative à quatre niveaux : consommation de 0 à 39 gr/jour, de 40 à 79 gr/jour, de 80 à 119 gr/jour, et de plus de 120 gr/jour. Ainsi, le nouveau modèle est de dimension quatre, et son estimation avec le premier niveau pris comme témoin :

	Estimate	Std. Error
(Intercept)	-2.6610	0.1921
alcgp40-79	1.1064	0.2303
alcgp80-119	1.6656	0.2525
alcgp120+	2.2630	0.2721

TABLE 8.4 – Exemple des données cancer : estimation des coefficients dans le cas d'une variable qualitative.

Les estimations des probabilités d'apparition de cancer correspondant aux différent niveaux de consommation sont précisées dans le tableau 8.5. Il est possible d'étudier un rapport de cotes d'un niveau par rapport à un autre : par exemple, l'influence d'une augmentation de consommation d'un niveau de 0 à 39 g/jour à un niveau de 40 à 79 g/jour :

$$OR(40-79/0-39) = e^{\theta_2}.$$

L'application numérique donne une estimation ponctuelle $OR(40-79/0-39) = e^{1,1064} = 3,023$, et un intervalle de confiance de niveau approché 95% vaut $[2,57; 3,47]$: la probabilité relative de l'apparition de cancer est multipliée par un facteur allant de $2,5$ à $3,5$ lorsque la consommation d'alcool passe du niveau $0-39$ au niveau $40-79$ g/jour.

niveau	probabilité prédite	écart type
0-39g/day	0,0653	0,0117
40-79	0,1744	0,0183
80-119	0,2698	0,0323
120+	0,4019	0,0463

TABLE 8.5 – Exemple des données cancer : probabilités prédites en fonction du niveau de consommation

8.4 Tests de rapport de vraisemblance

La vraisemblance maximale estimée dans un modèle constitue une mesure naturelle de la qualité de l'ajustement du modèle d'étude de paramètre θ. Le test de rapport de vraisemblance permet la comparaison de l'estimation dans deux modèles emboîtés.

8.4.1 Test classique du rapport de vraisemblance

Le test de rapport de vraisemblance permet de tester H_0, un modèle ω de dimension $dim(\omega)$, contre H_1, un modèle Ω de dimension $dim(\Omega)$. Si les hypothèses sont emboîtées, c'est à dire si $\omega \subset \Omega$, alors la loi de la statistique du rapport de vraisemblance TRV, sous l'hypothèse H_0, est asymptotiquement une loi du Khi-deux à $r = dim(\Omega) - \dim(\omega)$ degrés de liberté :

$$TRV = 2(l_n(\widehat{\theta}_\Omega) - l_n(\widehat{\theta}_\omega)) \xrightarrow{\mathcal{L}} \chi^2(r),$$

avec $l_n(\widehat{\theta}_\Omega)$ la valeur maximale de la log-vraisemblance dans Ω et $l_n(\widehat{\theta}_\omega)$ celle dans ω. La région de rejet de ce test est

$$\mathcal{R}_\alpha = \left\{TRV > q_{\chi^2(r)}(1-\alpha)\right\} \quad \text{avec } P(\mathcal{R}_\alpha) \simeq \alpha,$$

où $q_{\chi^2(r)}(1-\alpha)$ est le quantile de la loi du Khi-deux à r degrés de liberté, et le test de niveau approché α.

Ce test est utile pour tester la significativité d'un régresseur, d'un ensemble de régresseur ou de la régression dans son ensemble.

8.4.2 Déviance

Soit $\mathcal{M}_p$ un modèle de régression logistique d'observations $(Y_i)_{i=1,\dots,n}$, d'espérance $\mu_i = \mathrm{E}(Y_i)$, de prédicteur linéaire $\mathrm{x}_i\theta$, et de fonction de lien g :

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\theta_j = \eta_i(\theta) = \mathrm{x}_i\theta.$$

La qualité d'ajustement peut-être évaluée par le critère de **déviance**, fondé sur la comparaison entre la vraisemblance du modèle d'étude et la vraisemblance du modèle dit **saturé**. Le modèle saturé, noté $\mathcal{M}_S$, pose les mêmes hypothèses concernant la loi de Y_i mais ne contraint pas les espérances μ_i à varier selon une relation particulière. Il y a donc dans le modèle saturé autant de paramètres que d'espérances μ_i distinctes. C'est le modèle définissant la structure de l'espérance la moins contrainte. C'est pourquoi il est également appelé *maximal model* ou *full model* en anglais. Le modèle saturé n'a pas le même comportement selon que l'on analyse des **données**

individuelles (tous les x_i sont différents) ou des **données groupées** (des x_i sont identiques, et on les résume par le couple (n_k, x_k)), et cette distinction aura son importance lors de l'étude des propriétés de la déviance.

Définition 17. *La **déviance** d'un modèle $\mathcal{M}_p$ est la statistique définie comme le logarithme du rapport de vraisemblance entre le modèle saturé $\mathcal{M}_S$ et le modèle $\mathcal{M}_p$:*

$$D(\mathcal{M}_p) = 2[l_n(\widehat{\theta}_S, Y) - l_n(\widehat{\theta}; Y)] \tag{8.10}$$

*où $l_n(\widehat{\theta}_S, Y)$ (resp. $l_n(\widehat{\theta}; Y)$) désigne la log-vraisemblance maximale dans $\mathcal{M}_S$ (resp. $\mathcal{M}_p$). Elle est parfois appelée déviance **résiduelle** du modèle $\mathcal{M}_p$.*

Remarque : la **déviance nulle** mesure la plus grande différence de vraisemblance possible : celle entre le modèle saturé et le modèle ne contenant qu'un seul paramètre (modèle i.i.d).

8.4.3 Test de déviance

La statistique de déviance est donc liée au rapport de vraisemblance entre le modèle saturé $\mathcal{M}_S = (\Omega)$ et le modèle d'étude $D(\mathcal{M}_p) = (\omega)$. Sa loi pourrait donc être définie par (8.10), mais elle va en fait dépendre de la nature du modèle saturé : en effet, l'asymptotique n'est pas la même en données individuelles ou en données groupées.

Cas de données groupées

Les données groupées se rencontrent dans le cas d'observations répétées suivant les mêmes conditions d'expérience. Les données groupées sont typiques de régresseurs qualitatifs, comme pour l'analyse de la variance, ou dans l'exemple des données de cancer (Section 8.1.2).

Les n espérances des observations Y_i sont alors naturellement structurées en K groupes définis par K combinaisons différentes des variables explicatives. Au sein d'un groupe k, les x_i sont identiques, et les observations sont donc de même espérance et identiquement distribuées. Les Y_i sont alors traditionnellement notés Y_{kj}, où $k = 1, \ldots, K$ est l'indice du groupe et $j = 1, \ldots, n_k$ l'indice de répétition dans le groupe.

Comme il y a au maximum K espérances différentes, le modèle saturé spécifie un paramètre d'espérance par groupe : les espérances μ_k sont estimées par $\sum_j Y_{kj}/n_k$. Contrairement au cas des données individuelles, le modèle saturé en données groupées est un modèle paramétrique à part entière, qui réduit la complexité initiale des données à un nombre fixe $K < n$ de paramètres. L'asymptotique se définit lorsque les tailles n_k, $k = 1, \ldots, K$, des K groupes tendent vers l'infini à des vitesses comparables, c'est-à-dire lorsqu'il existe K valeurs finies strictement positives $a_1, \ldots, a_K$ telles que

$$\lim_{n\to\infty} \frac{n_k}{n} = a_k > 0, \quad k = 1, \ldots, K.$$

Dans ce cas, la déviance admet la loi asymptotique $\chi^2(K - p)$ pour $p < K$.

Théorème 9. *Dans le cas de données groupées, sous l'hypothèse que le modèle $\mathcal{M}_p$ est adéquat,*

$$D(\mathcal{M}_p) \xrightarrow{\mathcal{L}} \chi^2(K - p).$$

Le test de déviance permet de tester l'adéquation (H_0) d'un modèle $\mathcal{M}_p$ de dimension p contre sa non-adéquation (H_1 : les observations ne suffisent pas à expliquer correctement le modèle). Il est de région de rejet

$$\mathcal{R}_\alpha = \{D(\mathcal{M}_p) > q_{\chi^2(K-p)}(1 - \alpha)\}, \quad \text{avec } P(\mathcal{R}_\alpha) \simeq \alpha,$$

où $q_{\chi^2(K-p)}(1-\alpha)$ est le quantile de la loi du Khi-deux à $K-p$ degrés de liberté. La loi de la déviance étant asymptotique, le test est de niveau approché α.

Dans le cas de l'exemple de cancer, le modèle saturé comporte $K=12$ paramètres, que nous pouvons choisir égaux aux $\pi_k^S = \mathbb{P}(Z_k=1)$. Ainsi, la maximisation de la log-vraisemblance (8.3) en π_k^S entraîne $\widehat{\pi}_k^S = Y_k/n_k$ et sa valeur maximale vaut

$$l_n(\widehat{\theta}_S, \mathbf{Y}) = \sum_{k=1}^{K} Y_k \log\left(\frac{Y_k}{n_k - Y_k}\right) + \sum_{k=1}^{K} n_k \log\left(\frac{n_k - Y_k}{n_k}\right) + c(\mathbf{Y}).$$

Dans le modèle de dimension $p<K$, notons $\widehat{\pi}_k$ la probabilité estimée de $\{Z_k=1\}$ dans le groupe k, et $\widehat{Y}_k = n_k\widehat{\pi}_k$ la valeur "ajustée" de Y_k. Nous avons :

$$l_n(\widehat{\theta}; \mathbf{Y}) = \sum_{k=1}^{K} Y_k \log\left(\frac{\widehat{Y}_k}{n_k - \widehat{Y}_k}\right) + \sum_{k=1}^{K} n_k \log\left(\frac{n_k - \widehat{Y}_k}{n_k}\right) + c(\mathbf{Y}).$$

D'où l'expression de la déviance

$$D(\mathcal{M}_p) = 2\sum_{k=1}^{K}\left[Y_k \log\left(\frac{Y_k}{\widehat{Y}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - Y_k}{n_k - \widehat{Y}_k}\right)\right]. \tag{8.11}$$

Cas de données individuelles

Dans le cas de données individuelles, chaque réponse Y_i est associée à un cas de covariable x_i distinct de tous les autres. En général, cette situation s'observe lorsque les variables explicatives sont quantitatives et sans répétition, par exemple dans l'exemple de fragilité d'un alliage (Section 8.1.1). Le modèle saturé, par opposition à un modèle à $p<n$ paramètres, pose que les espérances des variables réponse Y_i sont sans lien entre elles : il est donc inutile d'expliciter un régresseur avec des variables explicatives. Le paramétrage se fait directement avec les n espérances :

$$\mathbb{E}(Y_i) = \mu_i, \quad i=1,\ldots,n,$$

où $\mu = (\mu_1,\ldots,\mu_n)$ est maintenant le vecteur de n paramètres à estimer. Ce modèle procure un ajustement parfait des données ($\widehat{Y}_i = \widehat{\mu}_i = Y_i$), mais il n'est pas très utile car il ne permet pas de calculer la variance de chaque $\widehat{\mu}_i$ en un x_i observé, ni de fournir de prédiction de Y en un x non observé. De plus, le nombre n de paramètres du modèle tend lui aussi vers l'infini, et la déviance peut donc diverger : le paramètre de la loi du χ^2 est en $n-p$.

Dans le cas de données individuelles, on ne connaît pas en général la loi asymptotique de la déviance. Le test de **Hosmer et Lemeshow** (cf Myers et al. (2012) par exemple) est une tentative de réponse dont le principe est de reformer des groupes d'observations de conditions proches.

L'expression de la déviance en données individuelles de l'exemple 8.1.1 est immédiate en remplaçant K par n et n_k par 1 dans l'équation 8.11.

8.4.4 Retour sur statistique du rapport de vraisemblance

Nous avons écrit la déviance comme le logarithme d'un rapport de vraisemblance. Ce dernier peut à son tour s'écrire comme une différence de déviance. En effet, pour deux modèles emboîtés $\omega \subset \Omega$,

$$D(\omega) - D(\Omega) = 2(l_n(\widehat{\theta}_\Omega; \mathrm{Y}) - l_n(\widehat{\theta}_\omega; \mathrm{Y})) = TRV.$$

Le test de rapport de vraisemblance est donc un test de différence de déviances. Ainsi, même dans le cas de données individuelles, la déviance peut servir à comparer le modèle d'étude avec un modèle de taille raisonnablement plus grande : la statistique de test est celle du rapport de vraisemblance entre les deux modèles. C'est pour cette raison que le test de rapport de vraisemblance est parfois appelé test de déviance.

8.5 Autres tests

D'autres tests, basés sur d'autres statistiques, sont également utilisés en régression logistique.

8.5.1 Test de Wald

Le test de Wald (Propriété 2) ou la delta méthode (Section 7.4.1) s'utilisent en régression logistique comme en régression non-linéaire.

8.5.2 Test du Khi2 de Pearson

Le résultat de l'algorithme des MCPI (8.7) produit la statistique de Pearson,

$$Khi^2 = \sum_i \frac{(Y_i - \widehat{Y}_i)^2}{\widehat{\text{var}}(Y_i)}. \tag{8.12}$$

Dans le cas de données groupées et pour un modèle adéquat, cette statistique suit asymptotiquement une loi du Khi-deux à r degré de libertés, où r est la différence de dimension entre le modèle saturé et le modèle d'étude. Cette statistique sert dans un test d'adéquation de région de rejet

$$\mathcal{R}_\alpha = \{Khi^2 > q_{\chi^2(r)}(1-\alpha)\}, \quad \text{avec } P(\mathcal{R}_\alpha) \simeq \alpha.$$

La loi de la déviance étant asymptotique, le test est de niveau approché α. Notons que la statistique de déviance et celle du Khi-deux de Pearson sont asymptotiquement égales.

8.5.3 Test du score

Toujours dans le contexte de modèles emboîtés $\omega \subset \Omega$, la statistique du score est définie par

$$S = [U(\widehat{\theta}_\omega)]'[I_n(\widehat{\theta}_\omega)]^{-1}U(\widehat{\theta}_\omega),$$

où $\widehat{\theta}_\omega$ est l'estimateur du maximum de vraisemblance contraint sous H_0, et les expressions de U et I_n utilisent la formulation d'un paramètre de Ω. On a, sous H_0,

$$S \xrightarrow{\mathcal{L}} \chi^2(r),$$

avec r le nombre de restrictions imposées sur θ par H_0. La région de rejet du test du score est

$$\mathcal{R}_\alpha = \left\{S > q_{\chi^2(r)}(1-\alpha)\right\} \quad \text{avec } \mathbb{P}(\mathcal{R}_\alpha) \simeq \alpha.$$

Ce test peut s'utiliser dans les mêmes contextes qu'un test de rapport de vraisemblance.

8.6 Outils de validation

En addition au test de l'adéquation du modèle, les représentations graphiques et l'étude des résidus sont des outils complémentaires de la validation du modèle.

8.6.1 Adéquation

Avant de commencer à utiliser le résultat d'une estimation, ou à affiner la définition d'un modèle, il est important de savoir si le modèle étudié comporte suffisamment de paramètres pour être correctement utilisable, c'est à dire s'il est adéquat. Le test emblématique est celui de la déviance, et donc du rapport de vraisemblance. Le test du Khi-deux de Pearson ou le test du score peuvent également être utilisés.

Tous ces tests ne sont théoriquement valides qu'en données groupées. En données individuelles, il est éventuellement possible de faire une comparaison avec un modèle raisonnablement plus grand. Des tests spécifiques peuvent aussi être définis : en régression logistique, le test d'Hosmer-Lemeshow regroupe des observations pour mimer des répétitions.

Si la décision est de conserver H_0, le modèle est adéquat, décision prise sans en connaître l'erreur (de seconde espèce). La décision est de rejeter H_0 est prise au risque α, le modèle ne convient pas, et il faut en investiguer la raison : il manque peut-être des régresseurs ; si ce n'est pas le cas, une modélisation en **sur-dispersion** peut permettre de modéliser un excès de variance.

8.6.2 Représentation de l'ajustement

La représentation graphique de l'ajustement peut se tracer dans l'échelle des observations $(\mathrm{x}_i, \widehat{\mu}_i)$ ou dans l'échelle du lien $(\mathrm{x}_i, g(\widehat{\mu}_i) = \widehat{\eta}_i)$ pour laquelle il y a linéarité : la figure 8.2 représente ces deux types pour l'exemple de la régression exponentielle. Mais elles ne sont adaptées que s'il n'y a qu'un seul régresseur. Dans le cas de plusieurs régresseurs, il est alors possible de tracer le graphe $(y_i, \widehat{y}_i)$, qui doit s'étirer sur la première bissectrice.

8.6.3 Résidus

Comme en régression linéaire, plusieurs types de résidus sont à disposition. Ceux-ci peuvent être utilisés pour affiner la qualité du modèle.

Différents résidus

Le **résidu brut** de l'observation i est défini comme la différence entre l'observation y_i et son espérance $\mu_i = g^{-1}(\mathrm{x}_i\theta)$. Il est estimé par

$$y_i - \widehat{y}_i = y_i - \widehat{\mu}_i \tag{8.13}$$

Les résidus bruts sont en général difficiles à utiliser car hétéroscédastiques : en effet, la variance de Y_i vaut $\pi(1-\pi_i)$ et dépend de son espérance μ_i. De plus, comme en linéaire, la variance du résidu dépend du plan d'expérience X.

Le **résidu de Pearson**, ou **résidu du Khi-deux**, standardise le résidu brut par l'écart-type de Y_i. Il est estimé par :

$$\widehat{r}_P = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}, \tag{8.14}$$

où $V(\widehat{\mu}_i)$ est la variance de Y_i dans laquelle μ_i est remplacée par son estimation $\widehat{\mu}_i$. Notons le lien avec la statistique de Pearson que nous avons déjà rencontrée (8.12) :

$$\sum \widehat{r}_P^2 = Khi^2.$$

Si $Y_i - \mu_i$ est bien de variance $V(\mu_i)$, l'estimation du résidu brut $Y_i - \widehat{\mu}_i$ est de variance dépendant du plan d'expérience et vaut $(1 - h_{ii})V(\mu_i)$. h_{ii} est le i-ème terme diagonal de la matrice

$H = \mathrm{X}(\mathrm{X}'W\mathrm{X})^{-1}\mathrm{X}'W$, où W a été définie en (8.4). Nous retrouvons une expression connue des moindres carrés pondérés. Ainsi, comme en régression linéaire, le résidu de Pearson peut être normalisé par le terme $\sqrt{1-h_{ii}}$:

$$\widehat{r}_N = \frac{y_i - \widehat{\mu}_i}{\sqrt{(1-h_{ii})V(\widehat{\mu}_i)}} \tag{8.15}$$

Ces résidus sont alors approximativement de variance unité.

Nous avons vu que la déviance permet de mesurer un écart entre le modèle saturé et le modèle d'étude. Elle s'écrit de façon générale $D(\mathcal{M}_p) = \sum_i d_i^2$, où d_i est la contribution de chaque observation à cet écart. Ainsi, le **résidu de déviance** de l'observation i est

$$r_D = signe(y_i - \widehat{y}_i)\sqrt{d_i^2}. \tag{8.16}$$

Le terme de signe permet d'avoir des résidus de même signe que ceux de Pearson. Les résidus de déviance peuvent également être normalisés par le terme $\sqrt{1-h_{ii}}$.

Utilisation des résidus

La construction des résidus normalisés les rend approximativement d'espérance nulle et de variance unité. Mais rappelons que les données doivent être groupées pour définir leur loi. Nous ne pouvons donc rien dire de plus en données individuelles. En données groupées (8.11) en revanche, il y a K résidus et qui suivent asymptotiquement une loi normale centrée réduite : leur analyse est alors semblable à celle faite dans le modèle linéaire.

Les graphes des résidus, comme en régression linéaire, peuvent permettre de détecter données aberrantes, des points leviers, un oubli de variables explicatives, non prise en compte d'un effet non-linéaire.

Notons enfin que bien que les résidus de Pearson et les résidus de déviance soient asymptotiquement équivalents, les résidus de déviance se comportent en général mieux à distance finie et sont donc souvent préférés.

8.7 Généralisation

La régression logistique permet de répondre à certaines limitations de la régression linéaire. C'est un cas de régression non-linéaire très particulier, qui permet de prendre en compte des observations binaires. Elle fait partie de la famille des modèles linéaires généralisés, qui sont définis par des observations indépendantes telles que :

1. la loi de la réponse fait partie d'une famille exponentielle de lois : comme par exemple les lois de Bernoulli, Poisson, multinomiale ou exponentielle
2. la présence d'un régresseur linéaire des covariables : $\mathrm{x}\theta$
3. une fonction de lien entre le régresseur linéaire et l'espérance de la loi.

Les modèles linéaires généralisés étendent le modèle de régression linéaire, donnent un cadre à une nouvelle classe de modèles qui possèdent de bonnes propriétés et sont largement utilisés dans les applications. Le terme de modèle linéaire généralisé (*generalized linear model* en anglais) est dû à Nelder et Wedderburn (1972), lorqu'ils ont étendu des méthodes sur les modèles binomiaux aux modèles de régression dans les familles exponentielles (Poisson, multinomiale par exemple). Une généralisation immédiate de la régression logistique est la régression multinomiale où la réponse observée est qualitative.

Chapitre 9

Classification

9.1 Introduction

Nous n'avons abordé pour l'instant la prédiction en régression logistique que sous l'angle des valeurs moyennes : il s'est agit de prédire $\widehat{\pi}_(\mathrm{x})$, l'espérance du phénomène sous la condition d'expérience x. La prédiction d'une valeur individuelle d'une variable catégorielle s'appelle classification

Définition 18. *La **classification (supervisée)** est la définition d'une règle de décision qui permet d'affecter une observation à une classe à partir d'une référence d'observations dont les classes sont déjà connues.*

Cette définition ne se limite donc pas au cas binaire. Cette méthode d'apprentissage est à distinguer de la **classification non supervisée** (ou **clustering**) pour laquelle il n'existe pas d'observations dont les classes sont déjà connues. Le travail est donc plus aisé (et le problème mieux posé) en classification supervisée. Il existe plusieurs méthodes de classification supervisée, basées par exemple sur la régression logistique, l'analyse discriminante ou les arbres de classification

En régression linéaire ou non linéaire, la valeur individuelle d'une nouvelle observation $Y_{\mathrm{x}_{n+1}}$ est prédite par $\widehat{Y}^p_{\mathrm{x}_{n+1}} = m(\mathrm{x}_{n+1}, \widehat{\theta})$. L'erreur de prévision est $\widehat{\varepsilon}^p_{n+1} = Y_{\mathrm{x}_{n+1}} - \widehat{Y}^p_{\mathrm{x}_{n+1}}$, d'espérance nulle et de variance $\mathrm{x}_{n+1}\,\mathrm{var}(\widehat{\theta})\,\mathrm{x}'_{n+1} + \sigma^2$. Mais en régression logistique, la fonction de régression $m(\mathrm{x}, \widehat{\theta}) = \pi(\mathrm{x}\widehat{\theta}) = \widehat{\pi}_{\mathrm{x}} \in]0;1[$ ne peut être directement utilisée pour estimer $\widehat{Y}_{\mathrm{x}}$, puisque $\widehat{Y}_{\mathrm{x}}$ ne prend que deux valeurs 0 ou 1.

Une première voie pourrait être de simuler $\widehat{Y}^p_{\mathrm{x}_{n+1}}$ suivant une loi de Bernoulli $\mathcal{B}(1, \pi_{\mathrm{x}_{n+1}})$, mais ce processus ne garantit pas la reproductibilité. La deuxième voie part de l'idée que si la probabilité $\pi(\mathrm{x}_{n+1})$ d'avoir 1 est grande (proche de 1), alors on peut prédire $\widehat{Y}^p_{\mathrm{x}_{n+1}} = 1$ avec peu d'erreur ; si elle est proche de 0, il faut prédire 0 :

$$\widehat{Y}_{\mathrm{x}_{n+1}} = \mathbb{1}_{\{\pi(\mathrm{x})>s\}} = h(s, \pi_{\mathrm{x}})$$

La règle de décision $h(s, \pi_{\mathrm{x}})$ est construite à l'aide d'un seuil s à déterminer.

9.2 Règle de Bayes

Il existe une règle uniformément meilleure que les autres, que l'on appelle règle de Bayes.

Proposition 6. *Soit a_0 le coût de choisir 1 à la place de 0, et a_1 le coût de choisir 0 à la place de 1. Alors le prédicteur*

$$h^*(\mathrm{x}) = \mathbb{1}_{\{\pi(\mathrm{x})>\frac{a_0}{a_1+a_0}\}}$$

minimise le risque quadratique $R(h)$, coût moyen de la perte quadratique associée à la décision h :

$$R(h) = \mathbb{E}[\ell(Y, h(X))] \text{ où } \ell(Y, h(X)) = (h(X) - Y)^2 a_Y$$

La décision h^* est appelée ***prédicteur de Bayes****.*

Preuve. Notons que

$$\ell(Y, h(X)) = (h(X) - Y)^2 a_Y = \mathbb{1}_{h(X)=0}\mathbb{1}_{Y=1}a_1 + \mathbb{1}_{h(X)=1}\mathbb{1}_{Y=0}a_0$$

Le coût moyen inconditionnel est donc

$$a_0\mathbb{P}(Y = 0 \cap \widehat{Y}^p = 1) + a_1\mathbb{P}(Y = 1 \cap \widehat{Y}^p = 0) = a_0\mathbb{E}(\mathbb{1}_{Y=0}\mathbb{1}_{h(\mathrm{x})=1}) + a_1\mathbb{E}(\mathbb{1}_{Y=1}\mathbb{1}_{h(\mathrm{x})=0})$$

Le coût moyen conditionnel à $\{X = x\}$ est

$$\begin{aligned}
C(h(\mathrm{X})|\mathrm{X} = \mathrm{x}) &= a_0\mathbb{E}(\mathbb{1}_{Y=0}\mathbb{1}_{h(\mathrm{X})=1}|\mathrm{X} = \mathrm{x}) + a_1\mathbb{E}(\mathbb{1}_{Y=1}\mathbb{1}_{h(\mathrm{x})=0}|\mathrm{X} = \mathrm{x}) \\
&= a_0\mathbb{1}_{h(\mathrm{x})=1}\mathbb{E}(\mathbb{1}_{Y=0}|\mathrm{X} = \mathrm{x}) + a_1\mathbb{1}_{h(\mathrm{x})=0}\mathbb{E}(\mathbb{1}_{Y=1}|\mathrm{X} = \mathrm{x}) \\
&= a_0\mathbb{1}_{h(\mathrm{x})=1}(1 - \pi(\mathrm{x})) + a_1\mathbb{1}_{h(\mathrm{x})=0}\pi(\mathrm{x}) \\
&= a_0\mathbb{1}_{h(\mathrm{x})=1}(1 - \pi(\mathrm{x})) + a_1(1 - \mathbb{1}_{h(\mathrm{x})=1})\pi(\mathrm{x}) \\
&= a_0\mathbb{1}_{h(\mathrm{x})=1} + (a_1 - (a_0 + a_1)\mathbb{1}_{h(\mathrm{x})=1})\pi(x)
\end{aligned}$$

Soit $h^*(\mathrm{x})$ un autre prédicteur, alors

$$\begin{aligned}
C(h(\mathrm{x})) - C(h^*(\mathrm{x})) &= a_0[\mathbb{1}_{h(\mathrm{x})=1} - \mathbb{1}_{h^*(\mathrm{x})=1}] - \pi(\mathrm{x})(a_0 + a_1)[\mathbb{1}_{h(\mathrm{x})=1} - \mathbb{1}_{h^*(\mathrm{X})=1}] \\
&= \underbrace{[\mathbb{1}_{h(\mathrm{x})=1} - \mathbb{1}_{h^*(\mathrm{x})=1}]}_{A}\underbrace{[a_0 - \pi(\mathrm{x})(a_0 + a_1)]}_{B}
\end{aligned}$$

Si $h^*(\mathrm{x})$ est le prédicteur de Bayes,
— si $\pi(\mathrm{x}) > a_0/(a_0 + a_1)$, alors, $\mathbb{1}_{h^*(\mathrm{x})=1} = 1$. on a $A \leq 0$ et $B \leq 0$
— si $\pi(\mathrm{x}) \leq a_0/(a_0 + a_1)$, alors, $\mathbb{1}_{h^*(\mathrm{x})=1} = 0$. on a $A \geq 0$ et $B \geq 0$
La différence est donc toujours positive. Ceci étant valable pour tout x, le coût moyen inconditionnel est bien minimisé par la règle de Bayes. ◊

Ceci permet de voir que le prédicteur de Bernoulli évoqué en introduction a une erreur moyenne supérieure à celle du prédicteur de Bayes.

Dans le cas d'une étude statistique, $\pi(x)$ est inconnue et sera estimée par $\widehat{\pi}(x)$ sur l'échantillon.

9.2.1 Cas particulier $a_0 = a_1 = 1$

Proposition 7. *Soit $a_0 = a_1 = 1$. Si ε^* est le coût moyen inconditionnel associé à la règle de décision de Bayes, et $\tilde{\varepsilon}$ le coût moyen inconditionnel associé au prédicteur de Bernoulli $\widehat{Y}_{\mathrm{x}} \sim \mathcal{B}(1, \pi(\mathrm{x}))$ on a*

$$\varepsilon^* \leq \tilde{\varepsilon} \leq 2\varepsilon^*$$

On en déduit que si ε^* est petit, le prédicteur de Bernoulli reste un bon prédicteur, même s'il est aléatoire.

Preuve. Il reste à monter la deuxième inégalité. Si $a_0 = a_1 = 1$, $h^*(\mathrm{x}) = \mathbb{1}_{\pi(\mathrm{x})>1/2}$

$$\varepsilon^*(\mathrm{x}) = C(h^*(\mathrm{x})) = \mathbb{1}_{h^*(\mathrm{x})=1} + (1 - 2\mathbb{1}_{h^*(\mathrm{x})=1})\pi(\mathrm{x})$$

— si $\pi(\mathrm{x}) > 1/2$, alors, $\mathbb{1}_{h(\mathrm{X})=1} = 1$. On a $\varepsilon^*(\mathrm{x}) = 1 - \pi(\mathrm{x}) \leq \pi(\mathrm{x})$
— si $\pi(\mathrm{x}) \leq 1/2$, alors, $\mathbb{1}_{h(\mathrm{X})=1} = 0$. On a $\varepsilon^*(\mathrm{x}) = \pi(\mathrm{x}) \leq 1 - \pi(\mathrm{x})$

Donc, $\varepsilon^*(\mathrm{x}) = \min(\pi(\mathrm{x}), 1 - \pi(\mathrm{x}))$. Pour le prédicteur de Bernoulli : $\tilde{h}(\mathrm{x}) = 1$ avec probabilité(conditionnelle) $\pi(\mathrm{x})$

$$\begin{aligned}
\tilde{\varepsilon}(\mathrm{x}) = C(\tilde{h}(\mathrm{x})) &= \mathbb{E}(\mathbb{1}_{Y=0}\mathbb{1}_{\tilde{h}(\mathrm{X})=1}|\mathrm{X} = \mathrm{x}) + \mathbb{E}(\mathbb{1}_{Y=1}\mathbb{1}_{\tilde{h}(\mathrm{x})=0}|\mathrm{X} = \mathrm{x}) \\
&= \mathbb{E}(\mathbb{1}_{\tilde{h}(\mathrm{X})=1}|\mathrm{X} = \mathrm{x})\mathbb{E}(\mathbb{1}_{Y=0}|\mathrm{X} = \mathrm{x}) + \mathbb{E}(\mathbb{1}_{\tilde{h}(\mathrm{X})=0}|\mathrm{X} = \mathrm{x})\mathbb{E}(\mathbb{1}_{Y=1}|\mathrm{X} = \mathrm{x}) \\
&= \pi(\mathrm{x})(1 - \pi(\mathrm{x})) + (1 - \pi(\mathrm{x}))\pi(\mathrm{x}) \\
&= 2\pi(\mathrm{x})(1 - \pi(\mathrm{x})) \\
&= 2\min(\pi(\mathrm{x}), 1 - \pi(\mathrm{x}))[1 - \min(\pi(\mathrm{x}), 1 - \pi(\mathrm{x}))] \\
&= 2\varepsilon^*(\mathrm{x})(1 - \varepsilon^*(\mathrm{x})) \leq 2\varepsilon^*(\mathrm{x})
\end{aligned}$$

D'où l'erreur inconditionnelle

$$\tilde{\varepsilon} = \mathbb{E}(\tilde{\varepsilon}(\mathrm{X})) \leq 2\mathbb{E}(\varepsilon^*(\mathrm{X})) = 2\varepsilon^*$$

◊

Définition 19. *On appelle* ***score de Bayes***
— *soit* $\pi(\mathrm{x}) - a_0/(a_1 + a_0)$ *dans l'échelle de la proba, qu'on compare à* 0.5
— *soit* $\mathrm{x}\theta - g(a_0/(a_1 + a_0))$ *dans l'échelle du lien qu'on compare à* 0.
Si $a_0 = a_1$, $g(a_0/(a_1 + a_0)) = 0$ *pour une fonction de lien* g *symétrique.*

Les calculs faits jusqu'à maintenant supposaient les probabilités $\pi(\mathrm{x})$ connues. Quand elles ne le sont pas, on les remplace par leurs estimées $\widehat{\pi}(\mathrm{x})$. L'erreur de prédiction sera elle-même estimée sur un échantillon test ou par validation croisée, cf cours suivant

9.2.2 Lien avec les mélanges

En utilisant le théorème de Bayes, on peut écrire

$$\pi(\mathrm{x}) = \mathbb{P}(Y = 1|\mathrm{X} = \mathrm{x}_0) = \frac{\mathbb{P}(Y = 1 \cap \mathrm{X} = \mathrm{x}_0)}{f(\mathrm{x}_0)} = \frac{f(\mathrm{X} = \mathrm{x}_0|Y = 1)\mathbb{P}(Y = 1)}{\sum_{c=0,1} f(\mathrm{X} = \mathrm{x}_0|Y = c)\mathbb{P}(Y = c)}$$

On peut donc voir le problème de classification comme l'affectation de x à l'une des composantes d'un **mélange de populations** de loi $\varphi_c(x) = \varphi(x|Y = c)$. La loi d'une observation a pour densité

$$f(x) = \sum_{c=0,1} \mathbb{P}(Y = c)\varphi_c(x) \tag{9.1}$$

Si l'échantillon d'apprentissage n'a pas la même répartition de proportion que la population, il faudra "redresser" les estimations pour les probabilités

$$\widehat{\pi}_P(\mathrm{x}) = \frac{\mathbb{P}_P(Y = 1)}{\mathbb{P}_A(Y = 1)} \frac{f_A(x)}{f_T(x)} \widehat{\pi}_A(\mathrm{x})$$

ie, modifier l'intercept du régresseur linéaire, mais pas les autres coefficients.

9.2.3 Extension

Dans le cas d'une variable réponse catégorielle à C, la règle de Bayes s'écrit

$$\widehat{Y}_{\mathrm{x}} = argmax_c \, \pi_c(\mathrm{x})$$

et

$$\pi_c(x) = \mathbb{P}(Y = c|\mathrm{X} = \mathrm{x}_0) = \frac{p(\mathrm{X} = \mathrm{x}_0|Y = c)\mathbb{P}(Y = c)}{\sum_{c'=1}^{C} p(\mathrm{X} = \mathrm{x}_0|Y = c')\mathbb{P}(Y = c')}$$

9.3 Scoring

On considère une règle de classification $h(\mathrm{x}) = \mathbb{1}_{\widehat{\eta}(\mathrm{x})>s}$ où s est le **seuil**. Par anglicisme, une règle de classification est parfois appelée **classifieur**. Le **scoring** est la méthode qui permet de construire un score pertinent pour la classification des individus d'une étude. Nous verrons en particulier comment comparer des scores (et donc des règles de classification).

9.3.1 Erreurs de classification

Deux types d'erreur peuvent être commises lors de la décision :
— **Faux Positif** : choix de $\widehat{Y} = 1$ alors que $Y = 0$, on a détecté (à tort).
— **Faux Négatif** : choix de $\widehat{Y} = 0$ alors que $Y = 1$, on n'a pas détecté (à tort).
On note FP le nombre de faux positifs, FN le nombre de faux négatifs, VP le nombre de vrais positifs (positifs et scorés positifs), VN le nombre de vrais négatifs (négatifs et scorés négatifs), POS le nombre de positifs et NEG le nombre de négatifs

	$\widehat{Y} = 1$	$\widehat{Y} = 0$	
$Y = 1$	VP	FN	POS
$Y = 0$	FP	VN	NEG

La figure 9.1 représente sur un exemple l'évolution du nombre de faux positifs, de faux négatifs et d'erreur totale d'une règle de classification construite sur un jeu de données. Quand $s = 0$, $\widehat{Y}^p = 1$ pour tout x, il n'y a pas de prédiction 0, et donc un maximum de faux positifs et aucun faux négatif. Quand s augmente, le nombre de prédits négatifs augmente également, de même que le nombre de faux négatifs. Le nombre de prédits positifs diminue, donc le nombre de faux prédits positifs aussi, pour atteindre aucun faux positif et un maximum de faux négatifs quand $s = 1$. L'erreur de classification, somme des deux, a un minimum compris entre 0.4 et 0.6. La règle de Bayes pose un seuil à 0.5.

Dans la théorie des tests, on cherche à minimiser l'erreur de seconde espèce à erreur de première espèce fixée. En classification, les faux positifs et faux négatifs sont mis sur le même pied d'égalité, et contribuent de la même façon à la minimisation de l'erreur.

Plutôt que de traiter directement avec le nombre de faux positifs et le nombre de faux négatifs, on va qualifier une règle de classification par des propriétés qui en sont déduites.

Définition 20. *On appelle*
— ***sensibilité*** *de la règle de classification, la probabilité de décider 1 (score > s) à raison*

$$\alpha(s) = \mathbb{P}(\{\widehat{\eta}(\mathrm{x}) > s\}|Y = 1) = \mathbb{P}(\widehat{Y} = 1|Y = 1)$$

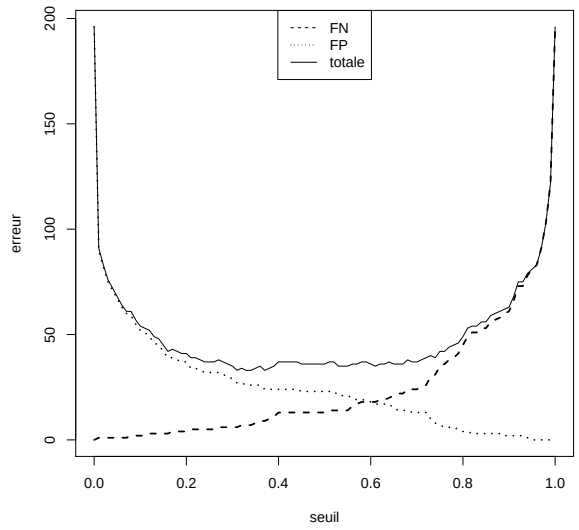


FIGURE 9.1 – Evolution des erreurs en fonction du seuil (exprimé dans l'échelle de la probabilité)

*On l'appelle également TPR (*True Positive Rate*), rappel (*recall *en anglais). C'est la puissance du test de* (H_0) *l'observation est issue de la classe* 0*, contre* (H_1) *l'observation est issue de la classe* 1*.*

— ***spécificité*** *de la règle de classification, la probabilité de décider 0 (score<*s*) à raison*

$$\beta(s) = \mathbb{P}(\{\widehat{\eta}(x) < s\}|Y = 0) = \mathbb{P}(\widehat{Y} = 0|Y = 0)$$

*On l'appelle également TNR (*True Negative Rate*) et vaut un moins le niveau du test précédent.*

La sensibilité est estimée par $\widehat{\alpha}(s) = VP/POS$. C'est l'estimation d'une proportion. On a : $VP \sim \mathcal{B}(POS, \alpha(s))$, d'où $\text{var}(VP) = POS\ \alpha(s)(1-\alpha(s))$ et

$$\text{var}\left(\frac{VP}{POS}\right) = \frac{\alpha(s)(1-\alpha(s))}{POS} \simeq \frac{\widehat{\alpha}(s)(1-\widehat{\alpha}(s))}{POS}$$

De la même façon, $\widehat{\beta}(s) = VN/NEG$.

On souhaite qu'une règle de décision soit très spécifique (pour ne pas déclarer positifs des cas qui ne le sont pas) et très sensible (pour déclarer positifs des cas qui le sont). Mais ces deux objectifs sont antagonistes...

Pour comparer l'efficacité d'une règle de classification, on peut tracer des courbes de sensibilité
— en fonction de $1-\beta(s)$: **courbe ROC**
— en fonction de la probabilité de choisir 1 (à tort ou à raison) : **courbe de lift**.

9.3.2 Courbe ROC

La courbe ROC (Receiver Operating Curve) trace la sensibilité $\alpha(s)$ (proportion de positifs détectés pour un seuil s) en fonction de $1-\beta(s)$ (proportion de scorés positifs parmi les négatifs) La courbe ROC contient quelques points caractéristiques :

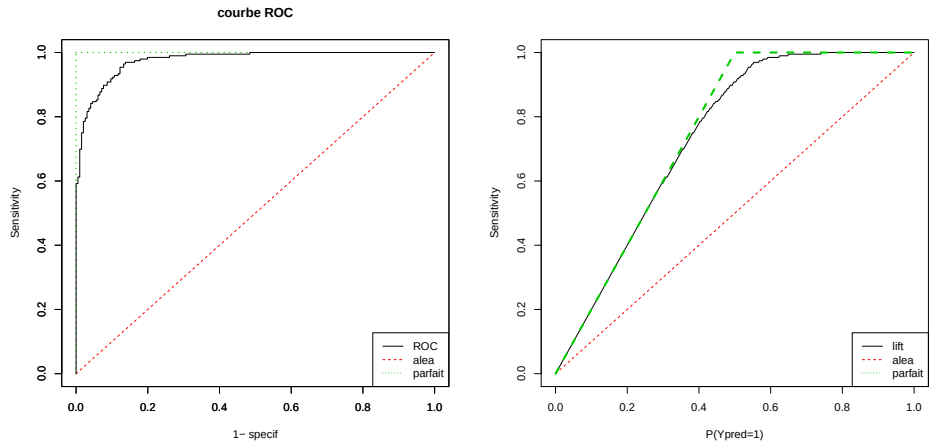


FIGURE 9.2 – Exemple de courbe ROC (à gauche) et de lift (à droite)

— le point $(0, 0)$ appartient toujours à la courbe : il correspond au seuil maximum (∞ dans l'échelle du lien ou 1 dans l'échelle de la probabilité). Le seuil est tel que rien n'est jamais prédit positif.
— le point $(1, 1)$ appartient toujours à la courbe : il correspond au seuil minimum ($-\infty$ dans l'échelle du lien ou 0 dans l'échelle de la probabilité). Le seuil est tel que tout est toujours prédit positif

Sur la figure 9.2 à gauche sont également représenté deux classifieurs particuliers :

— La première bissectrice correspond à un choix aléatoire équi-réparti entre positif et négatif. C'est le choix au hasard quand on n'essaie pas de tirer partie des covariables.
— un score parfait discrimine exactement : il existe un seuil pour lequel 100% de vrais positifs et 0% de faux positifs sont prédits. Ainsi, sa courbe est formée du segment $(0; 0) - (0; 1)$ et du segment $(0; 1) - (1; 1)$

Un score est d'autant meilleur que sa courbe ROC est proche de celle du score parfait. Ces courbes permettent donc de comparer les méthodes de classification :

— comparaison locale : la courbe d'un score est au dessus de celle d'un autre score. Mais ces tendances peuvent s'inverser tout au long de la courbe
— comparaison globale : comparaison des aires sous la courbe (estimation par la méthode des trapèzes, sur l'exemple $A = 0.95$). Un score est globalement meilleur qu'un autre si l'aire sous sa courbe ROC est plus grande que l'aire sous la courbe ROC de son concurrent.
On montre que $A = 1 - \frac{W}{n_1 n_0}$ avec W statistique non-paramétrique de Mann-Whitney, n_1 le nombre d'événements, n_0 le nombre de non-événement. Elle se prête donc à un test statistique pour tester la significativité du score par rapport à la décision due au hasard.

Les courbes ROC permettent de comparer des classifieurs, qu'ils proviennent de régression logistique ou d'autres méthodes de classification.

9.3.3 Courbe de lift

C'est une variante de la courbe ROC, souvent utilisée en marketing. Elle est aussi utilisée en économétrie sous le nom de courbe de Lorentz ou courbe de puissance. Elle représente la sensibilité $\alpha(s)$ en fonction de la proportion des individus scorés positifs (qu'ils le soient à tort

ou à raison), ie en fonction de

$$\begin{aligned} P_s(\widehat{Y}=1) &= P_s(\widehat{Y}|Y=0)P(Y=0) + P_s(\widehat{Y}|Y=1)P(Y=1) \\ &= (1-\beta(s))(1-p) + \alpha(s)p \end{aligned}$$

où p est la proportion de clients risqués ou d'acheteurs potentiels dans la population totale.

La courbe de lift est sous la courbe ROC : en effet, pour une même ordonnée, l'abscisse de lift est plus grande (si le score est meilleur que le score aléatoire, $\alpha(s) > 1-\beta(s)$) :

$$\underbrace{P_s(\widehat{Y}=1)}_{x_{lift}(s)} = \underbrace{(1-\beta(s))}_{x_{ROC}(s)}(1-p) + \underbrace{\alpha(s)}_{\geq 1-\beta(s)} p$$

$\alpha(s) \geq 1-\beta(s)$ pour les règles de décisions meilleures que la décision aléatoire. On a donc

$$x_{lift}(s) \geq (1-\beta(s))(1-p) + (1-\beta(s))p = 1-\beta(s) = x_{ROC}(s)$$

et

$$\mathcal{A}_{lift} = \int \alpha d\{p\alpha + (1-p)(1-\beta)\} = p\int \alpha d\alpha + (1-p)\int \alpha d\{1-\beta\} = \frac{p}{2} + (1-p)\mathcal{A}_{ROC}$$

Si $\mathcal{A}_{ROC} > 0.5$, alors $\mathcal{A}_{ROC} - \mathcal{A}_{lift} = p(\mathcal{A}_{ROC} - 0.5)$ donc $\mathcal{A}_{ROC} > \mathcal{A}_{lift}$:

Pour obtenir la courbe de lift, on classe les individus par score décroissant, puis on les regroupe en classes (centiles par exemple), on calcule le pourcentage de vrais acheteurs dans chaque centile, et on dresse la courbe cumulative de ces pourcentages : en effet,

$$P(\widehat{Y}=1|Y=1) = \frac{P(\widehat{Y}=1 \cap Y=1)}{P(Y=1)} \simeq \frac{VP/(POS+NEG)}{POS/(POS+NEG)} = \frac{VP}{POS}$$

- Si le score est parfait, $\mathcal{A}_{ROC} = 1$, $\mathcal{A}_{lift} = 1 - p(1-0.5) = 1 - p/2$. La courbe de lift est formée de deux segments de droite d'extrémités $(0;0) - (p,1) - (1;1)$
- Si p est très faible, $P_s(\widehat{Y}=1) \simeq 1-\beta(s)$. Courbes de lift et de score sont très proches
- On peut comparer deux modèles en comparant leurs courbes de lift.

Définition 21. *Un point de coordonnées $(n;m)$ sur la courbe signifie que $n\%$ des individus ayant le plus fort score concentre $m\%$ des acheteurs ou des risques.*

*On appelle **lift à** $n\%$ le quotient m/n où (m,n) appartient à la courbe de lift. On peut ainsi tracer le diagramme de lift.*

La courbe de lift est utilisée pour définir l'efficacité d'un mailing par exemple :

- On fixe la taille de la cible (par exemple 20% des individus de la base de taille N).
  - Si on envoie les courriers au hasard, 20% des individus positifs seront atteints, soit $0.2Np$ si p est la probabilité de positifs potentiels
  - Si on suit le score mis en place (figure 9.2 à droite), environ 40% des individus positifs seront atteints, soit $0.4Np$: on a gagné $0.2Np$ clients supplémentaires par rapport au hasard
  - le taux de retour passe de $0.2Np/(0.2N) = p$ à $0.4Np/(0.2N) = 2p$ (2 est le lift à 0.2)
  - la part de marché passe à $0.4Np/(Np)$ => il reste encore 0.6 Np clients potentiels non découverts
- Si on fixe un objectif de part de marché : 50% des clients potentiels
  - il faut envoyer un courrier aux 35% premiers individus de plus grand score de la base $0.35N$

— Avec un tirage aléatoire, il faudrait envoyer $0.5N$ pour espérer ramener le même nombre de nouveaux clients, d'où une économie de $0.15N$
— le taux de retour passe de $0.5Np/(0.5n) = p$ à $0.5Np/(0.35N) = 1.4p$
— la part de marché vaut $0.5Np/(Np) =>$ il reste encore $0.5\ Np$ clients potentiels non découverts

L'aire sous la courbe de lift n'a pas de signification absolue, car elle dépend de la probabilité p de l'événement, mais elle doit tendre vers la courbe de lift idéale. D'où la mesure de performance d'un modèle prédictif définie par le ratio de Gini (accurate ratio) :

Définition 22. *On appelle* ***indice de Gini*** *le rapport suivant :*

$$\begin{aligned}\frac{\textit{surface entre la courbe de lift réelle et la diagonale}}{\textit{surface entre la courbe de lift idéale et la diagonale}} &= \frac{\mathcal{A}_{lift} - 0.5}{1 - p/2 - 0.5} \\ &= \frac{p/2 + (1-p)\mathcal{A}_{ROC} - 0.5}{(1-p)/2} \\ &= 2\mathcal{A}_{ROC} - 1\end{aligned}$$

C'est le double de la surface entre la courbe ROC et la diagonale. Il est indépendant de p comme $\mathcal{A}_{ROC}$ et coïncide avec la statistique de Somers.

9.3.4 Autres indicateurs liés à la classification

On définit également
— Le taux de faux négatifs, ou FNR *False Negative Rate* : $FNR = FN/POS = 1 - TPR$, c'est l'équivalent de l'erreur de seconde espèce
— Le taux de faux positifs, ou FPR *False Positive Rate* : $FPR = FP/NEG = 1 - TNR$, c'est l'équivalent de l'erreur de première espèce.
— La précision (*precision* en anglais) est le rapport du nombre de classés par le nombre de scorés positifs $PREC = VP/(VP + FP)$
— Le FDR (*False Discovery Rate* en anglais) est le rapport des faux positifs par le nombre de scorés positifs $PREC = FP/(VP + FP) = 1 - PREC$
— L'*accuracy* est le taux de bien classés $ACC = (VP + VN)/n$
— la prévalence (*prevalence* en anglais) est le taux de positifs de l'échantillon POS/n. Noter que la prévalence de l'échantillon n'est pas toujours celle de la population entière, en particulier si le tirage a été pondéré.

Ces indicateurs sont résumés dans le tableau suivant

	$\widehat{Y} = 1$	$\widehat{Y} = 0$			
$Y = 1$	VP	FN	POS	$\alpha = TPR = VP/POS$ puissance, recall, sensibilité	$FNR = FN/POS = 1 - TPR$ erreur de 2nde espèce
$Y = 0$	FP	VN	NEG	$FPR = FP/NEG = 1 - TNR$ erreur de 1ère espèce	$\beta = TNR = VN/NEG = 1 - FPR$ spécificité
	precision $\frac{VP}{VP+FP}$				
	FDR $\frac{FP}{VP+FP}$				

A partir de ces indicateurs, on définit le score F, moyenne harmonique de la precision

$PREC = VP/(VP + FP)$ et de la sensibilité (recall) $\alpha = VP/(VP + FN)$:

$$F = 2\frac{1}{\frac{1}{PREC} + \frac{1}{\alpha}} = 2\frac{PREC \times RECALL}{PREC + RECALL}$$

Le score F est compris entre 0 et 1 et la classification est d'autant meilleure qu'il est proche de 1.

9.3.5 Méthodologie d'une étude de score

Celle-ci se fait en plusieurs étapes, qui peuvent être longues quand on travaille sur les données de l'entreprise

- Déterminer les objectifs de l'étude
- Faire l'inventaire et la préparation des données (étude uni et bi-variée)
- Constituer la base d'analyse
- Élaborer le modèle prédictif (choix des variables, ..., validation)
- Utilisation du score sur le lieu de vente (choix de groupe, ...)
- Déploiement du score
- Suivi

Cette méthodologie s'applique pour la construction de différents types de score :

- score d'**appétence** : propension à consommer, score d'affinité. C'est l'étude de la probabilité qu'un client connu soit intéressé par un produit
- score (de comportement) **de risque** : étude de la probabilité de rencontrer un accident de traitement (calculé après quelques mois de présence dans la banque, prend en compte le fonctionnement des comptes).
- Le croisement des deux premiers score permet la définition du score de **pré-acceptation** (appétence positive et risque négatif)
- score d'**octroi** : pour un nouveau client ou un client ayant eu une activité faible : on ne dispose pas ou pas assez de données historisées, et le risque est calculé en temps réel, sur la base de données déclaratives fournies par le client (CSP, ...) et de données de géomarketing (niveau de vie et habitudes de consommation dans la zone d'habitation du client)
- score de **recouvrement** : évalue le montant susceptible d'être récupéré sur un compte ou un contentieux
- score d'**attrition** : étudie la probabilité de quitter un service (banque, assurance..). Ce cas est délicat à modéliser, il y a de nombreuses façons de partir : diminuer les versements, les flux. Le score ne devient en général très fiable qu'au moment du départ. L'attrition d'un produit est en général plus facile à mettre en place que l'attrition d'un client.

9.4 Autres méthodes

La régression logistique est largement utilisée en classification. Mais il existe d'autres méthodes paramétriques (par exemple la discrimination) ou non paramétriques (K plus proches voisins, SVM, réseaux de neurones) qui feront l'objet d'un autre cours.

La méthode de discrimination fait l'hypothèse que les données proviennent d'un mélange de populations (gaussiennes p-variées) qui ont chacune leur caractéristiques propres. Le modèle de mélange a été défini par l'équation 9.1.

Suivant la classe d'appartenance c, $\mathrm{x} = (x_1, \ldots, \mathrm{x}_p) \sim \mathcal{N}_p(\mu_c, \Sigma_c)$ suit une gaussienne différant en moyenne et variance

$$\varphi_c(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_c|^{1/2}} \exp -\frac{1}{2}[(\mathrm{x}-\mu_c)'\Sigma_c^{-1}(\mathrm{x}-\mu_c)]$$

9.4.1 Analyse Discriminante Linéaire (LDA)

On considère que toutes les gaussiennes ont même variance $\Sigma_c = \Sigma$ et on calcule

$$\pi_c(\mathrm{x}) = \mathbb{P}(Y_c = 1|\mathrm{X} = \mathrm{x}) = \frac{\mathbb{P}(Y=c)\exp -\frac{1}{2}[(\mathrm{x}-\mu_c)'\Sigma^{-1}(\mathrm{x}-\mu_c)]}{\sum_\ell \mathbb{P}(Y=\ell)\exp -\frac{1}{2}[(\mathrm{x}-\mu_\ell)'\Sigma^{-1}(\mathrm{x}-\mu_\ell)]}$$

Les dénominateurs sont communs et peuvent être ignorés. De même, il suffit de ne conserver dans le classifieur que les termes qui dépendent de c. Notons $\pi_c = \mathbb{P}(Y = c)$. Le classifieur associé à chaque classe s'écrit donc

$$\delta_c(\mathrm{x}) = \mathrm{x}'\Sigma^{-1}\mu_c - \frac{1}{2}\mu_c'\Sigma^{-1}\mu_c + \log(\pi_c)$$

On cherche c rendant maximum $\pi_c(\mathrm{x})$. Si $C > 2$, il faut faire $C(C-1)/2$ comparaisons deux à deux des numérateurs, de frontières $\delta_c(\mathrm{x}) = \delta_\ell(\mathrm{x})$ qui sont linéaires en x.

Comme on ne connait pas Σ, μ_c et π_c, ces paramètres sont estimés par la méthode des moments, ou par maximum de vraisemblance.

Attention, il faut faire attention attention au fait que si $\mathbb{P}(Y = 1)$ est faible, la règle qui prédit toujours 0 (s'il n'y a que deux niveaux) indépendamment du reste aura une erreur faible (égale à la proportion d'observations $Y = 1$, et le classifieur trivial (aléatoire) pourra avoir un taux d'erreur que peu différent de celui d'une LDA. Dans un tel cas, il faut que la règle arrive à baisser le taux d'erreur parmi les individus $Y = 1$, ie autoriser éventuellement un plus fort taux de faux positifs pour arriver à découvrir les vrais positifs.

9.4.2 Analyse Discriminante Quadratique (QDA)

Chaque classe a maintenant sa propre matrice de covariance et le classifieur pour chaque classe est de la forme :

$$\delta_c(\mathrm{x}) = -\frac{1}{2}\mathrm{x}'\Sigma_c^{-1}\mathrm{x} + \mathrm{x}'\Sigma_c^{-1}\mu_c - \frac{1}{2}\mu_c'\Sigma^{-1}\mu_c - \frac{1}{2}\log|\Sigma_c| + \log(\pi_c)$$

Les frontières $\delta_c(\mathrm{x}) = \delta_\ell(\mathrm{x})$ sont quadratiques en x.

QDA a $Cp(p+1)/2 + Cp$ paramètres à estimer contre $p(p+1)/2 + Cp$ pour LDA. LDA est donc moins flexible, mais aussi substantiellement moins de variance, ce qui peut améliorer la prédiction

Maintenant, si l'hypothèse de variance commune est erronée, on peut le payer au prix d'un fort biais.

Donc, s'il y a peu de données, on préfèrera LDA pour diminuer la variance. Sur un grand jeu de données, on peut s'essayer à QDA, puisque il y a suffisamment d'information pour estimer la variance, ou parce que l'hypothèse de variance commune n'est pas tenable

9.4.3 KNN

La classification par k plus proches voisins (k nearest neighbors) est une méthode de classification non paramétrique. On choisit un nombre k, puis on affecte l'observation en regardant les ses k plus proches voisins par votre majoritaire (et tirage aléatoire s'il y a des ex-aequo). Le problème est le choix de k, cf cours suivant.

Chapitre 10

Choix de modèles

Nous avons jusqu'à présent, pour étudier les propriétés des estimateurs, supposé disposer de tous les régresseurs du "vrai" modèle et de la "vraie" fonction de régression. Dans les applications réelles, la variable à expliquer ou prédire est identifiée, et un certain nombre de régresseurs sont candidats, parmi lesquels il faut choisir. De plus, différents types de modèles (linéaire, non linéaires, non paramétriques...) peuvent aussi être mis en compétition. Comment choisir entre ces modèles ? Un modèle est vu ici comme l'ensemble des densités qui répondent à des hypothèses de modélisation données. Le modèle ajusté est l'estimation du modèle à partir des données d'estimation. Une famille de modèles est un ensemble de modèles partageant les mêmes hypothèses, mais avec des complexité différentes. Par exemple, l'ensemble des régressions polynomiales est une famille de modèles (emboîtés dans ce cas) indexée par le degré du polynôme, représentant sa complexité.

On peut choisir un modèle à partir d'une information a priori, par exemple une information causale d'un phénomène physique. Une autre façon est de retenir un modèle explicatif par le moyen de tests de sous-modèles emboîtés (Fisher, Wald, rapport de vraisemblance...). Mais les modèles à comparer ne sont pas forcément emboîtés, voire n'appartiennent pas forcément à la même famille de modèle. Et un bon modèle explicatif n'est pas forcément un bon modèle prédictif. En effet, la variabilité due à l'estimation des paramètres influe sur celle de la prédiction : il peut ainsi arriver qu'un "faux" modèle donne de meilleures prédictions que le "vrai" modèle (voir l'exemple jouet de Azais et Bardet (2005)). Quel est donc "le" "meilleur" modèle ? Il est important de préciser le but de la régression, et d'utiliser un critère adapté (Mallows, 1973) :

— **Description** : la réponse Y est d'autant mieux décrite que l'erreur résiduelle SCR est faible, ou que le coefficient de détermination R^2 est fort dans le cas du modèle linéaire. Ce critère va amener à retenir le modèle avec le plus de régresseurs.

— **Estimation de la fonction de régression** : il s'agit de déterminer θ (ou la fonction de régression $f(x;\theta)$) le plus précisément possible. Si des composantes sont oubliées, l'estimateur sera biaisé ; s'il y en a trop, sa variance sera plus importante. L'erreur quadratique moyenne associée à l'estimation de θ peut être un critère permettant un compromis entre le biais et la variance, qu'il faut savoir estimer. C'est l'axe de la tradition statistique, *statistical learning*.

— **Prévision de nouvelles observations** : il s'agit de trouver un modèle minimisant le risque d'erreur de réponse sur de nouvelles observations, même si le modèle considéré est peu explicatif, voire n'est pas le modèle de génération des données. Quand la règle de décision n'est pas liée à un modèle, on passe de l'apprentissage statistique à l'apprentissage

automatique ou *machine learning*, dont la visée est essentiellement prédictive.

Le choix des variables doit répondre à deux objectifs contradictoires
- Le nombre de variables doit être réduit pour que le modèle soit facilement interprétable
- L'ensemble des variables doit être suffisamment grand pour que l'ajustement soit correct.

Le choix doit donc être à la fois parcimonieux (peu de paramètres), mais fournissant un bon ajustement. Il existe plusieurs critères de sélection, et tous ne conduisent pas au même choix de modèle : le choix du modèle est guidé par une synthèse des performances fournis par différentes méthodes de sélection. Il faut d'ailleurs distinguer
- le **choix de modèle**, qui estime les performances de différents modèles dans le but d'en choisir "un" "meilleur"
- l'**établissement de la performance d'un modèle** (*model assessment*), i.e. l'estimation du risque sur le modèle retenu.

Les sections de ce chapitre étudient l'influence d'un oubli ou d'un ajout de covariables, l'estimation de la performance d'un modèle, l'étude de critères de choix et les procédures de sélection.

Bibliographie Ce chapitre est une introduction à la problématique du choix de modèle, en prenant l'exemple de la régression. Sa structure est largement inspirée de celle de Cornillon et Matzner-Løber (2007). Azais et Bardet (2005) précise les détails techniques et le comportement asymptotique des critères.

10.1 Conséquences d'un choix incorrect de variables

Une mauvaise paramétrisation de la moyenne peut avoir plusieurs origines (Guyon, 2001) :
- trop de régresseurs ont été retenus : on dit qu'il y a sur-paramétrisation
- il manque des régresseurs : on dit qu'il y a sous-paramétrisation
- plus généralement, certains régresseurs ont été oubliés et d'autres sont superflus.

10.1.1 Sur-paramétrisation

Il n'y a pas d'oubli de variables, mais certaines sont superflues. On suppose donc que le vrai modèle (inconnu) est un modèle linéaire (ω) de dimension $|\omega| = q$, et qu'il est estimé dans $\Omega(\supset \omega)$ de dimension p :

(ω) : $E(Y) = X_1\theta_1$ (vrai modèle, inconnu), $dim(\theta_1) = q$, $dim(X_1) = n \times q$

(Ω) : $E(Y) = X\theta$ (modèle de travail), avec $\theta = (\theta_1', \theta_2')'$ de dimension p et $X = [X_1 X_2]$ de dimension $n \times p$, matrice résultant de la concaténation de X_1 et de celle des $r = p - q$ régresseurs supplémentaires X_2. Le vrai paramètre dans Ω est $\theta^* = (\theta_1', 0')'$.

En remarquant que

$$(X'X)^{-1} = \begin{pmatrix} \nabla^{-1} & -\nabla^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}\nabla^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}\nabla^{-1}A_{12}A_{22}^{-1} \end{pmatrix},$$

où $A_{ij} = X_i'X_j$ et $\nabla = A_{11} - A_{12}A_{22}^{-1}A_{21}$, l'EMC de θ^* dans Ω est $\hat{\theta}_\Omega = (X'X)^{-1}X'Y$, il est sans biais,

$$\mathbb{E}(\hat{\theta}_\Omega) = (X'X)^{-1}X'\mathbb{E}(Y) = (X'X)^{-1}X'X\theta^* = \theta^*$$

mais sa variance ∇^{-1} pour les q premières composantes est supérieure à $(X_1'X_1)^{-1}$, celle dans (ω). En effet, $A_{11} - A_{12}A_{22}^{-1}A_{21} \leq A_{11}$ puisque $A_{12}A_{22}^{-1}A_{21}$ est symétrique définie positive,

d'où $\nabla^{-1} \geq A_{11}^{-1} = (X_1'X_1)^{-1}$. Maintenant, si le plan d'expérience X_2 est orthogonal au plan d'expérience X_1, alors $A12 = 0$, et la variance n'est pas augmentée.

D'autre part, $\hat{\sigma}^2(\Omega) = \frac{SCR(\Omega)}{n-(p+q)}$ estime aussi sans biais σ^2, mais $\text{var}(\hat{\sigma}^2(\Omega)) = \frac{2\sigma^4}{n-(p+q)} \geq \text{var}(\hat{\sigma}^2(\omega)) = \frac{2\sigma^4}{n-p}$.

Proposition 8. *La sur-paramétrisation ne biaise pas les estimateurs, mais elle en diminue la précision.*

10.1.2 Sous-paramétrisation

C'est le cas d'un oubli de régresseurs. On suppose que le "vrai" modèle (inconnu) est (Ω) de dimension $|\Omega| = p$, mais que le modèle de travail est le modèle $\omega(\subset \Omega)$ de dimension $q < p$:

(ω) : $E(Y) = X_1\theta_1$ (modèle de travail)

(Ω) : $E(Y) = X\theta$ (vrai modèle, inconnu), avec $\theta = (\theta_1', \theta_2')'$ de dimension p et $X = [X_1 X_2]$ de dimension $n \times p$, matrice résultant de la concaténation de X_1 et de celle des $r = p - q$ régresseurs sur-numéraires X_2.

Les paramètres correspondant aux régresseurs oubliés sont pris en compte dans θ_2. L'EMC de θ_1 dans (ω) est biaisé :

$$\begin{aligned} E(\hat{\theta}_1) &= (X_1'X_1)^{-1}X_1'\mathbb{E}(Y) \\ &= (X_1'X_1)^{-1}X_1'(X_1\theta_1 + X_2\theta_2) \\ &= (X_1'X_1)^{-1}X_1'X_1\theta_1 + (X_1'X_1)^{-1}X_1'X_2\theta_2 \\ &= \theta_1 + (X_1'X_1)^{-1}X_1'X_2\theta_2 \end{aligned}$$

sans compter l'oubli de la partie θ_2. De même, la variance de $\widehat{\sigma}_\omega$ est un estimateur biaisé de σ^2. Comme dans le cas de sur-paramétrisation, si X_2 est orthogonale à X_1, alors $A_{12} = 0$, et l'estimateur des q premières composantes n'est pas biaisé. Il l'est bien sûr pour les $p - q$ composantes oubliées. Ce biais est acceptable si l'estimation dans (ω) diminue fortement la variance. C'est le cas par exemple dans une situation de presque-colinéarité : si deux régresseurs sont presque colinéaires, la variance associée à chacun d'entre eux est très forte. On peut donc avoir intérêt à en supprimer l'un des deux, avec une faible augmentation du biais, mais une forte diminution de la variance.

Proposition 9. *La sous-paramétrisation biaise les estimateurs, mais peut en augmenter la précision.*

Bien sûr, les deux situations précédentes peuvent se produire (oubli de régresseur et régresseurs surnuméraires). Et une solution pour réduire le biais (ajouter des régresseurs) va à l'encontre de celle pour réduire la variance (diminuer le nombre de régresseurs).

10.2 Performance d'un modèle

Quelle performance veut-on mesurer ? Si le critère de choix est la variance, le modèle retenu aura peu de paramètres à estimer. Si c'est le biais, il faudra au contraire prendre le plus de paramètres. L'erreur quadratique moyenne est un critère permettant de réaliser un compromis entre ces deux phénomènes, tandis que l'erreur erreur quadratique moyenne de prédiction aura pour objectif de minimiser l'erreur de prédiction d'une valeur individuelle. Nous verrons leur définition, puis les moyens de les estimer.

Remarque Il existe d'autres critères de qualité, comme par exemple la dissemblance de Kullback, qui sont à la base de la définition de critère de sélection de modèle par vraisemblance pénalisée, mais nous ne développerons pas cet aspect.

10.2.1 Erreur quadratique moyenne

L'erreur quadratique moyenne est un critère permettant de réaliser un compromis entre ces deux phénomènes. Mais c'est un critère probabiliste, qui fait intervenir le paramètre inconnu. Une difficulté va être son estimation.

Définition 23. *L'erreur quadratique moyenne d'un estimateur $\widehat{\theta}$ de θ est le risque quadratique d'utiliser $\widehat{\theta}$ à la place de θ :*

$$\begin{aligned} EQM(\widehat{\theta}) &= \mathbb{E}(||\theta - \widehat{\theta}||^2) \\ &= \mathbb{E}(||\theta - \mathbb{E}(\widehat{\theta})||^2) + \mathbb{E}([\widehat{\theta} - \mathbb{E}(\widehat{\theta})]'[\widehat{\theta} - \mathbb{E}(\widehat{\theta})]) \\ &= ||Biais(\widehat{\theta})||^2 + tr(\mathrm{var}(\widehat{\theta})) \end{aligned}$$

On y reconnaît la somme d'un terme de biais et d'un terme de variance. L'EQM permet donc de comparer les estimateurs d'un même paramètre. Dans le cas de la régression, il est classique de traiter le problème de choix de variables par l'intermédiaire de la valeur ajustée $\widehat{Y}$ plutôt que par l'intermédiaire de $\widehat{\theta}$.

Définition 24. *Dans le modèle de régression $Y = F(X, \theta) + \varepsilon$, l'**erreur quadratique moyenne** de $\widehat{m}_\omega$ est le risque quadratique de cet estimateur calculé dans le modèle ω pour estimer la fonction de régression $\mathbb{E}(Y) = F(X, \theta)$ du modèle ayant généré l'échantillon :*

$$\begin{aligned} EQM(\widehat{m}_\omega) &= \mathbb{E}||F(X, \theta) - \widehat{m}_\omega||^2 \\ &= ||biais(\widehat{m}_\omega)||^2 + tr(\mathrm{var}(\widehat{m}_\omega)) \end{aligned} \tag{10.1}$$

Dans le cas du modèle linéaire, $\widehat{m}_\omega = X_\omega \widehat{\theta}_\omega$ est la projection H_ω de Y sur (ω), d'où $tr(\mathrm{var}(X_\omega \widehat{\theta}_\omega)) = \sigma^2 tr(H_\omega) = |\omega|\sigma^2$, avec $|\omega|$ la dimension de (ω) :

$$EQM(\widehat{m}_\omega) = ||biais(\widehat{m}_\omega)||^2 + |\omega|\sigma^2$$

Nous avons vu que le biais diminue avec $|\omega|$ et la variance augmente dans ce cas de façon linéaire avec $|\omega|$. EQM permet donc de définir un compromis biais-variance, que l'on cherchera à minimiser, cf figure 10.1.

Si la variance est relativement facile à estimer, la tâche est plus difficile pour le biais (dont il faudrait connaître la valeur du vrai paramètre...). Nous verrons cependant plus loin des critères pour l'estimer.

10.2.2 Erreur quadratique moyenne de prévision

L'estimation de EQM n'étant pas a priori aisée, il est classique de faire intervenir n^v nouvelles observations (X^v, Y^v). Il est alors possible de comparer ces nouvelles observations avec la prédiction qu'on peut en faire sous le modèle (ω).

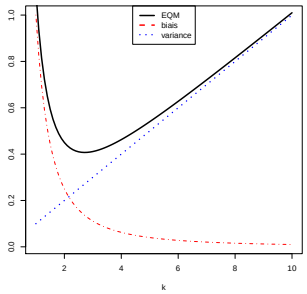


FIGURE 10.1 – Compromis Biais-Variance : avec peu de paramètres, le biais est important. Il diminue avec le nombre de paramètres alors que la variance augmente. L'EQM réalise un compromis entre les deux.

Définition 25. *Le risque quadratique des estimations $\widehat{Y}_\omega^v$ dans le modèle (ω) de nouvelles observations (X^v, Y^v) est appelé* ***erreur quadratique moyenne de prévision***

$$\begin{aligned} EQMP(\widehat{Y}_\omega^v) &= \mathbb{E}(||\widehat{Y}_\omega^v - Y^v||^2) \\ &= EQM(\widehat{Y}_\omega^v) + \mathbb{E}(||\mathbb{E}(Y^v) - Y^v||^2) \\ &- 2\mathbb{E}[(X_\omega^v\widehat{\theta}_\omega - X^v\theta)(\mathbb{E}(Y^v) - Y^v)] \end{aligned} \tag{10.2}$$

Si les nouvelles observations sont indépendantes de celles qui ont servi à calculer $\widehat{\theta}_\omega$, l'espérance du produit dans (10.2) est nulle, et

$$EQMP(\widehat{Y}_\omega^v) = EQM(\widehat{Y}_\omega^v) + n^v\sigma^2$$

les deux critères sont équivalents à minimiser. Bien sûr, si le nouvel échantillon n'est pas indépendant de l'échantillon initial, l'EQMP devient un mauvais critère puisque la méthode d'ajustement s'adapte aux données d'apprentissage, et leur réutilisation biaise l'estimation de l'erreur de prédiction : c'est ce qu'on appelle le **biais d'apprentissage**.

10.2.3 Apprentissage, validation et test

Contrairement à l'EQM, l'EQMP peut s'estimer facilement avec un estimateur empirique. La somme des carrés résiduels SCR est par exemple un estimateur empirique de l'EQMP basé sur l'échantillon initial. Mais c'est un mauvais estimateur de l'EQMP : la méthode d'ajustement s'étant adaptée aux données, elle est optimiste sur l'erreur de prédiction, et l'utilisation de SCR va conduire à retenir le plus de variables possibles, cf la courbe *échantillon d'apprentissage* de la Figure 10.2. Il faut donc posséder deux échantillons, l'un permettant d'estimer le modèle, l'autre d'estimer son erreur. Or rappelons-nous qu'il y a deux objectifs séparés :

- choix de modèle, l'estimation des performances de différents modèles dans le but d'en choisir "un" "meilleur" (ou un bien adapté)
- model assessment, établissement de la performance du modèle retenu, on dit encore performance de **généralisation**.

Si le jeu de données est suffisamment grand, la meilleure approche pour les deux problèmes est de diviser aléatoirement le jeu de données entre trois parties : échantillons d'**apprentissage**

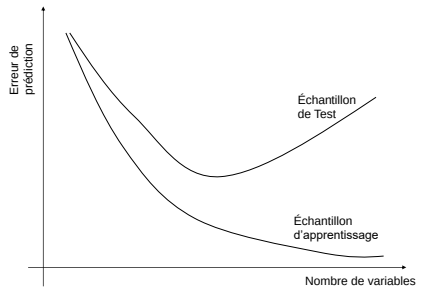


FIGURE 10.2 – Biais de la somme des carrés résiduels calculée sur l'échantillon d'apprentissage pour estimer le risque. L'estimation sur un échantillon test n'est pas biaisée.

(pour estimer les modèles), de **validation** (pour estimer leurs performances, et les comparer) et de **test** (pour estimer la performance de généralisation du modèle retenu). L'échantillon de test doit être vu comme un tabernacle, qui ne sert qu'à la fin de l'analyse. Si on l'utilise dans le processus de choix, l'erreur calculée sur l'échantillon de test sous-estimera l'erreur réelle de généralisation.

Si l'échantillon de test est trop petit, la variance de l'estimation de l'erreur de généralisation sera importante. Si l'échantillon d'apprentissage est trop petit, l'estimation de l'erreur de prédiction pourra être biaisée. Il faut donc que les deux échantillons soient suffisamment fournis pour permettre une bonne estimation à la fois des paramètres du modèle (apprentissage) et de la performance (test). Il n'y a pas de règle a priori sur la taille de ces échantillons : (3/4,1/4), (2/3,1/3), (1/2, 1/2) par exemple si on ne conserve qu'un échantillon apprentissage et un échantillon test, (1/2,1/4,1/4) pour les trois échantillons apprentissage/test/validation. Mais cela dépend des données et des méthodes ; et il faut être attentif à garder l'homogénéité entre ces échantillons.

10.2.4 Estimation de la performance sur un échantillon indépendant

L'estimateur empirique calculé sur un n^v-échantillon (X^v, Y^v) indépendant de (X, Y) et de même loi jointe que (X, Y) permet d'obtenir un estimateur non biaisé et consistant de l'EQMP. L'estimation de l'erreur moyenne de prédiction est alors calculée comme la moyenne empirique des carrés des erreurs sur l'échantillon indépendant :

$$\widehat{EQMP}(\omega) = \frac{1}{n^v}||Y^v - F(\widehat{\theta}_\omega, X^v)||^2 = \frac{1}{n^v}\sum_{i=1}^{n^v}\left(Y_i^v - f(\widehat{\theta}_\omega, \mathrm{x}_i^v)\right)^2,$$

Notons que le principe est le même pour une réponse qualitative ou catégorielle prenant K valeurs distinctes, labellisés par exemple de 1 à K. A la place de la perte quadratique, on utilise alors une perte $0-1$

$$\ell(Y, \widehat{Y}(X)) = \mathbb{1}_{Y \neq \widehat{Y}(X)}$$

ou la déviance

$$\ell(Y, \widehat{Y}(X)) = -2\sum_k \mathbb{1}_{Y=k} \log \widehat{p}_k(X) = -2 \log \widehat{p}_Y(X)$$

où $\widehat{p}_k(X)$ est la probabilité estimée qu'une observation Y soit dans la classe k sous les conditions X.

Ce principe peut s'appliquer pour définir la performance d'un modèle donné sur l'échantillon de validation à des fins de comparaison si le critère de choix est l'erreur de prédiction ; ou pour définir l'erreur de généralisation du modèle retenu (sur l'échantillon test).

10.3 Validation croisée

Idéalement, on l'a vu, quand on dispose de suffisamment de données, on scinde les données initiales en trois échantillons. Mais il existe des méthodes pour approcher l'étape de validation

— soit analytiquement, et cela mène à la définition de critères de choix de modèle qui s'affranchissent de l'échantillon de validation

— soit par simulation, en utilisant du ré-échantillonnage par validation croisée. Il est également possible de faire appel à des techniques de bootstrap, en tirant les observations suivant la loi empirique, mais ce point ne sera pas abordé.

Nous détaillons ici le principe de la validation croisée, les critères de choix seront présentés en section suivante.

Validation croisée pour remplacer l'échantillon de validation La validation croisée mime le processus de séparation entre apprentissage et validation. On définit un premier échantillon d'apprentissage, disons avec 90% des données, puis on calcule l'erreur sur les 10% restant, et on recommence l'opération neuf fois façon à ce que chaque observation soit exactement une fois dans un échantillon test. De façon plus générale, l'échantillon est partitionné en B parties de tailles n^b approximativement égales appelés plis. On parle de validation à B plis, ou B**-fold validation**. Chaque pli est à tour de rôle :

— retiré de l'échantillon d'apprentissage

— l'estimation $\widehat{\theta}_\omega^{(-b)}$ se fait sur l'échantillon tronqué

— l'erreur de prédiction est calculée sur la partie mise en réserve (X^b, Y^b)

— puis le pli est remis pour passer au suivant.

La moyenne des erreurs de prédiction cumulées sur les B parties est un estimateur asymptotiquement sans biais et consistant de l'EQMP :

$$\widehat{EQMP}(B) = \frac{1}{n}\sum_{b=1}^{B} ||Y^b - F(\widehat{\theta}_\omega^{(-b)}, X^b)||^2$$

Le cas particulier $B = n$ est celui de la validation croisée par **Leave One Out**. Dans le cas de la régression linéaire, l'estimation du critère s'exprime facilement ;

$$LOO = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i}{1-h_i}\right)^2 ,$$

où $h_i = [H_\omega]_{ii} = [X_\omega(X'_\omega X_\omega)^{-1}X'_\omega]_i$ est le i ème terme diagonal de la matrice de projection orthogonale. Cette erreur peut être approximée de la façon suivante (*Generalized Cross Validation*)

$$GCV = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{(i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i}{1-trace(H)/n}\right)^2$$

où la quantité $trace(H)$ est le nombre effectif de paramètres.

L'estimation de l'erreur de prédiction par validation croisée LOO est approximativement non biaisée à distance finie, mais peut avoir une grande variance, à cause des n échantillons d'apprentissage qui sont très similaires les uns des autres. Elle est asymptotiquement consistante. Mais elle a aussi un coût non négligeable dans les cas autres que ceux de la régression linéaire.

D'un autre côté, l'erreur calculée en validation croisée avec un moindre nombre de plis a une variance plus faible, mais le biais pourrait être un problème, dépendant de la variation de la performance de la méthode d'apprentissage en fonction de la taille de l'échantillon d'apprentissage.

Validation croisée pour remplacer l'échantillon de test La VC peut aussi être utilisée pour mimer l'échantillon de test parce que l'échantillon est trop petit pour sanctuariser cette partie des données :

— si le critère de sélection est l'erreur de prédiction, une double validation croisée est effectuée, l'une pour estimer l'erreur de chacun des modèles et les comparer, l'autre pour estimer l'erreur de généralisation

— si le critère de sélection approxime l'étape de validation (section suivante), une simple validation croisée permet d'obtenir la performance de la méthode de sélection.

10.4 Pratique du choix de modèles

Une alternative à la validation croisée pour s'affranchir de l'échantillon de validation, est l'utilisation de critères qui essaient de proposer un compromis biais variance. Tous sont adaptés à la sélection de variables (qui est un cas particulier de choix de modèles). Les critères AIC et BIC, utilisant la vraisemblance, peuvent être plus largement utilisés dans la comparaison de modèles dépendant de paramètres n'ayant pas trait à la prise en compte de variables. Rappelons que l'erreur de prédiction est également un critère convenant à tout type de comparaison de modèles, mais nous n'y revenons pas dans cette section. Avant d'aborder la présentation des critères, nous présentons différentes procédures de sélection de variables.

10.4.1 Procédures de sélection de variables

Lorsque le nombre de covariables n'est pas trop grand, on peut étudier tous les sous-modèles possibles : il y a C_{p-1}^{q} modèles à q variables explicatives et donc, 2^{p-1} modèles avec intercept. On peut calculer des critères tels que R^2, C_p ou LOO dans chacun de ces modèles, et retenir le modèle qui optimise l'un ou l'autre critère. Les quelques modèles candidats seront alors comparés par leur EQMP.

Quand la méthode exhaustive n'est pas possible, des méthodes pas à pas peuvent être mises en oeuvre :

— La méthode descendante (**backward**) procède par élimination successive de variables : à partir du modèle complet, la variable la moins influente sur le critère choisi est successivement enlevée. Si le critère est un test (Wald, Fisher, Rapport de vraisemblance...), cette méthodologie est celle des tests de type III, dont il faut définir le seuil de sortie.

— la méthode ascendante (**forward**) procède par ajout successif de variables : à partir du modèle le plus simple dont on souhaite partir (par exemple, le modèle iid), la variable la plus influente sur le critère choisi est successivement ajoutée. Si le critère est un test, cette méthodologie est celle des tests de type I, dont il faut définir le seuil d'introduction.

— La méthode mixte (**stepwise**) enchaîne les étapes ascendantes et descendantes : après une étape de forward (ajout d'une variable), l'algorithme étudie la possibilité d'en enlever une (potentiellement différente) avec une étape de backward et s'arrête lorsqu'on ne peut plus ajouter ni retrancher de variable.

De façon générale, la stratégie descendante peut être recommandée, puisqu'elle part d'un modèle qui est plus susceptible de contenir le "vrai" modèle. Mais si l'estimation est délicate dans le modèle complet, on choisira alors la stratégie ascendante. Quelle que soit la stratégie utilisée, il est important, comme nous l'avons vu, de tester le modèle obtenu en calculant son erreur de généralisation (sur un échantillon test, ou par VC).

Le reste de cette section présente des critères de choix de modèles permettant d'approcher l'étape de validation.

10.4.2 Le critère R^2

Pour chaque sous-modèle (ω) de dimension $|\omega|$, on calcule le coefficient de détermination (B.5)

$$R^2(\omega) = \frac{SCM(\mathbb{1}, \omega)}{SCR(\mathbb{1})} = 1 - \frac{SCR(\Omega)}{SCT},$$

et on sélectionne le modèle de plus grand R^2, définissant ainsi le modèle de meilleur ajustement, dont il faudra préciser la performance par un processus indépendant. Dans la mesure où le R^2 ne peut qu'augmenter si l'on ajoute un prédicteur, ce critère ne peut être employé que pour choisir un modèle parmi des sous-ensemble de prédicteurs de même taille. Pour choisir la taille d'un sous-ensemble de prédicteur, on peut pénaliser le R^2 en définissant un coefficient de détermination multiple corrigé :

$$R^2_a(\omega) = 1 - \frac{n-1}{n-|\omega|}(1 - R^2(\omega)) = 1 - \frac{n-1}{n-|\omega|}\frac{||Y - \widehat{Y}^2||}{||Y - \bar{y}\mathbb{1}||^2}$$

puis retenir le modèle (ω) qui maximise $R^2_a(\omega)$. A nouveau, il faudra par ailleurs établir la performance du modèle. Enfin, il sera sans doute intéressant de comparer des modèles qui ont des R^2_a proches, même s'ils ne sont pas égaux.

10.4.3 Test de modèles emboîtés

Nous avons vu les tests de Fisher, de Wald ou de rapport de vraisemblance, qui peuvent être utilisés soit de manière ascendante (type I), soit de manière descendante (type III), ces deux types étant identique dans le modèle linéaire à design orthogonal (par exemple en ANOVA équilibré).

Quand la statistique de Fisher est utilisée dans le test du sous-modèle (ω) contre le modèle (ω_{+1}) ayant une variable supplémentaire

$$F = \frac{SCR(\omega) - SCR(\omega_{+1})}{\widehat{\sigma}^2},$$

il faut déterminer dans quel modèle calculer l'estimateur de la variance du bruit :

1. si l'estimateur calculé dans le modèle complet contenant toutes les variables, soit $SCR(\Omega)/(n-p)$; la statistique de test suit alors une loi $\mathcal{F}(1, n-p)$
2. si l'estimateur calculé dans le modèle (ω_{+1}), soit $SCR(\omega_{+1})/(n-|\omega|-1)$; la statistique de test suit alors une loi $\mathcal{F}(1, n-|\omega|-1)$

Notons que les deux procédures sont identiques pour la première étape d'un algorithme backward mais difficilement comparables dans les autres cas.

Cette approche ne cherche pas à estimer l'erreur de prédiction, mais essaie plutôt de trouver un modèle qui s'ajuste bien aux données, afin d'essayer de la modéliser et les comprendre.

10.4.4 Le critère du C_p de Mallows

Le crotère de Mallows Si le modèle est linéaire, le risque quadratique de $\widehat{m}_\omega = X_\omega \widehat{\theta}_\omega$ pour l'estimation de $E(Y) = X\theta = m$ est, cf 10.1

$$EQM(\widehat{Y}_\omega) = E\left(\|\widehat{m}_\omega - m\|^2\right) = ||biais(\widehat{m}_\omega)||^2 + tr(\mathrm{var}(\widehat{m}_\omega)) = \|m_\omega - m\|^2 + |\omega|\sigma^2,$$

Par ailleurs,

$$\begin{aligned}
\mathbb{E}(SCR(\omega)) &= \mathbb{E}(||Y - \widehat{m}_\omega||^2) \\
&= \mathbb{E}(||Y - \widehat{m}_\omega + \mathbb{E}(Y - \widehat{m}_\omega) - \mathbb{E}(Y - \widehat{m}_\omega)||^2) \\
&= \mathbb{E}(||(I - H_\omega)(Y - m)||^2) + ||m - m_\omega||^2 \\
&= (n - |\omega|)\sigma^2 + ||m - m_\omega||^2
\end{aligned}$$

Ce faisant, il faut pouvoir inverser $\mathbb{E}$ et H_ω, la projection sur ω. Dans ce cas, c'est à dire si $SCR(\omega)$ est estimé sur un échantillon de validation indépendant de l'échantillon d'apprentissage, un estimateur du biais $||m - m_\omega||^2$ est $SCR(\omega) - (n - |\omega|)\sigma^2$. D'où l'estimateur de l'erreur quadratique moyenne

$$\widehat{EQM(\widehat{m}_\omega)} = SCR(\omega) - (n - 2|\omega|)\sigma^2.$$

Le C_p de Mallows revient à retenir le modèle (ω) minimisant

$$\frac{\widehat{EQM(\widehat{m}_\omega)}}{\sigma^2} = \frac{SCR(\omega)}{\sigma^2} + 2|\omega| - n$$

en supposant σ^2 connu. Si σ^2 n'est pas connu, il est estimé en général dans le modèle le plus grand connu, d'où la définition du critère de Mallows

Définition 26. *Le critère de Mallows d'un modèle linéaire gaussien (ω) de dimension $|\omega|$ est défini par*

$$C_p(\omega) = \frac{SCR(\omega)}{\widehat{\sigma}^2} + 2|\omega| - n$$

où $SCR(\omega)$ est estimée dans le modèle (ω) tandis que $\widehat{\sigma}^2$ est un estimateur de la variance, calculé par exemple dans le plus grand modèle à disposition.

On voit ici que la somme des carrés résiduels a été pénalisée par un terme qui dépend de la dimension du modèle, c'est à dire du nombre de régresseurs. Avec un faible nombre de régresseurs, $SCR(\omega)$ l'emporte et $C_p(\omega)$ est élevé. Quand on rajoute de (bons) régresseurs, $SCR(\omega)$ diminue drastiquement alors que l'augmentation due à la pénalisation est modérée, et $C_p(\omega)$ diminue.

En particulier, quand (ω) contient toutes les variables du modèle, alors $SCR(\omega)$ est un estimateur sans biais de $(n - |\omega|)\sigma^2$ et $C_p(\omega)$ vaut approximativement $|\omega|$.

Quand l'ajout de régresseurs ne fait plus diminuer suffisamment $SCR(\omega)$ pour compenser l'augmentation de la pénalisation, le $C_p(\omega)$ se remet à augmenter. On cherche donc un modèle (ω) qui minimise le $C_p(\omega)$ de Mallows.

Mais attention, il faut garder en tête que cette interprétation n'est valable que si le $C_p(\omega)$ est calculé avec d'autres données que celles qui permettent le choix de (ω). De plus, le C_p de Mallows ou des critères sous-pénalisés peuvent amener à des résultats catastrophiques quand le nombre de variables est grand, et dans le cas d'une recherche complète.

10.4.5 La log-vraisemblance pénalisée : critères AIC et BIC

La qualité de l'ajustement d'un modèle peut être évaluée par la log-vraisemblance maximale $\log L(\omega)$ calculée dans le modèle (ω). Mais si $\omega \subset \Omega$, la log-vraisemblance dans (Ω) sera plus importante que celle dans (ω) : on observe le même phénomène de sur-ajustement que pour le SCR. Il est alors possible de pénaliser la vraisemblance pour le contrebalancer. Les critères s'écrivent en général sous la forme

$$-2\log L(\omega) + 2|\omega| pen(n)$$

Deux critères sont classiquement utilisés :

— **AIC** (Akaike Information Criterion) : (ω) minimise

$$AIC(\omega) = -2\log L(\omega) + 2|\omega|$$

Dans le cas gaussien : $AIC(\omega) = cte + n\log\frac{SCR(\omega)}{n} + 2|\omega|$.

— **BIC** (Bayesian Information Criterion) : (ω) minimise

$$BIC(\omega) = -2\log L(\omega) + |\omega|\log(n)$$

Dans le cas gaussien : $BIC(\omega) = cte + n\log\frac{SCR(\omega)}{n} + |\omega|\log(n)$.

Ces critères s'appliquent à un modèle linéaire gaussien, mais aussi aux modèles non-linéaires pour lesquels on dispose d'une vraisemblance.

Avec ces définition, on souhaite un BIC ou AIC le plus faible possible : c'est l'opposée de la vraisemblance qui est pénalisée. Certains logiciels prennent la convention inverse, il faut toujours être vigilant sur les conventions de chacun.

Notons que le critère BIC possède une pénalisation en $\log n/2$ qui devient rapidement plus importante que celle de AIC. BIC aura tendance a sélectionner moins de régresseurs que AIC sur les grands échantillons.

La confrontation de deux visions Ces deux critères asymptotiques ont des objectifs différents

— vision *efficacité* : chercher à estimer sans biais le risque, et choisir un estimateur qui le minimise. C'est le cas du C_p de Mallows (quand σ^2 est connu), ou de l'AIC. Quand le nombre de paramètres et le nombre de modèles restent finis quand le nombre d'observations tend vers l'infini, le C_p de Mallows est asymptotiquement efficace si σ^2 est connu, même si le vrai modèle n'est pas dans la collection. Mais remplacer σ par son estimateur calculé sur les données ne donne pas un bon estimateur de l'erreur de prédiction.

— vision *consistance* : trouver un modèle qui soit le plus proche possible du "vrai" modèle ayant généré les données. Sa consistance a été démontrée dans de nombreux cas si le "vrai" modèle fait partie de la liste des modèles en compétition. Sinon, BIC aura tendance à donner des modèle de trop grande complexité.

10.4.6 Méthodes non asymptotiques

Les méthodes présentées ici sont en général asymptotiques, elles ne sont donc en général valables que si le nombre de modèles reste "raisonnable". Elles ne sont plus valides quand celui-ci est fonction de la taille de l'échantillon (par exemple, définir une fonction étagée sur un intervalle donné, sans connaître ni le nombre, ni le lieu des ruptures). Il est alors nécessaire de pénaliser les critères avec la complexité de la famille de modèles. Cette problématique fait l'objet de la théorie non asymptotique du choix de modèle.

Chapitre 11

Méthodes de régularisation

Nous avons rencontré dans les chapitres précédents un cas où le rang de la matrice du plan d'expérience était inférieur au nombre de ses colonnes, typiquement dans le cas de variables qualitatives où des colonnes sont liées. Le problème avait été résolu en posant une contrainte d'identifiabilité $C\theta = 0$, ce qui revenait à "régulariser" $X'X$ en la remplaçant par $X'X + C'C$ qui est inversible.

Maintenant, le rang de X peut être inférieur au nombre de colonnes parce que le nombre de variables est supérieur (voire très supérieur) au nombre d'observations : on est alors dans un cadre de **grande dimension**. Enfin, des variables explicatives très corrélées peuvent entraîner, même si la situation reste identifiable, des valeurs propres de la matrice du plan d'expérience faibles, et donc une variance de l'estimation forte : la variance de $\widehat{\theta}$ peut fortement augmenter si $X'X$ est proche d'une matrice non inversible, entraînant une introduction aux méthodes de **régularisation** permettant de pallier le problème de dégénérescence (ou presque dégénérescence) du rang de X. Dans le cadre de la régression, on parle également de **régression biaisée**, parce que ces méthodes engendrent un biais sur l'estimateur. Nous présenterons ces méthodes dans le cadre de la régression linéaire, mais ils s'étendent à d'autres cas, en particulier celui de la régression logistique.

Bibliographie James et al. (2013), Cornillon et Matzner-Løber (2007)

11.1 Régression Ridge

La régression ridge a été proposée par Hoerl et Kennard (1970). Rappelons que les valeurs propres λ_j de $X'X$ sont positives ou nulles. S'il y a dégénérescence du rang, les λ_j ordonnées par ordre décroissant sont quasi-nulles à partir d'un certain rang r.

Or $X'X$ et $X'X + \kappa Id_p$ ont les mêmes vecteurs propres, mais des valeurs propres différentes : si u_j est un vecteur propre de $X'X$ associé à la valeur propre λ_j, alors

$$(X'X + \kappa Id_p)u_j = X'Xu_j + \kappa u_j = (\lambda_j + \kappa)u_j$$

c'est un vecteur propre de $X'X + \kappa Id_p$ associé la la valeur propre $\lambda_j + \kappa$ respectivement, $j = 1, \ldots, p$, ce qui amène à la définition de l'estimateur ridge

$$\widehat{\theta}_{ridge}(\kappa) = (X'X + \kappa Id_p)^{-1}X'Y$$

Le problème est maintenant la détermination κ. Mais avant de l'aborder, notons que la régression ridge peut être vue comme un problème d'optimisation sous contraintes ℓ_2 :

$$\tilde{\theta} = \arg \min_{\theta \in \mathbb{R}^p; ||\theta||^2 \leq \delta} ||Y - X\theta||^2$$

et n'a donc d'intérêt que si δ est petit, car sinon (si δ est suffisamment grand) $\tilde{\theta}$ est l'EMCO. Ce problème d'optimisation se résout en utilisation le lagrangien

$$L(\theta, \kappa) = ||Y - X\theta||^2 + \kappa(||\theta||^2 - \delta)$$

En le dérivant

$$\begin{aligned} -2X'(Y - X\widehat{\theta}_{ridge}) + 2\tilde{\kappa}\widehat{\theta}_{ridge} &= 0 \\ ||\widehat{\theta}_{ridge}||^2 - \delta &= 0 \end{aligned}$$

d'où $\tilde{\theta} = \widehat{\theta}_{ridge} = (X'X + \tilde{\kappa} Id_p)^{-1} X'Y$ avec

$$\tilde{\kappa} = (\widehat{\theta}_{ridge} X'Y - \widehat{\theta}'_{ridge} X'X\widehat{\theta}_{ridge})/\delta$$

Ce couple définit bien un minimum puisque le hessien

$$\begin{pmatrix} 2(X'X + \kappa Id_p) & 2\theta \\ 2\theta & 0 \end{pmatrix}$$

est une matrice symétrique semi définie positive.

11.1.1 Propriétés

— L'estimateur ridge est biaisé, on parle d'ailleurs de **régression biaisée** :

$$\begin{aligned} \mathbb{E}(\widehat{\theta}_{ridge}) - \theta &= \mathbb{E}[(X'X + \kappa Id_p)^{-1} X'Y] - \theta \\ &= (X'X + \kappa Id_p)^{-1} X' \mathbb{E}[Y] - \theta \\ &= (X'X + \kappa Id_p)^{-1} (X'X + \kappa Id_p - \kappa Id_p)\theta - \theta \\ &= -\kappa (X'X + \kappa Id_p)^{-1} \theta \end{aligned}$$

— Variance :

$$\begin{aligned} \text{var}(\widehat{\theta}_{ridge}) &= \text{var}((X'X + \kappa Id_p)^{-1} X'Y) \\ &= (X'X + \kappa Id_p)^{-1} X' \,\text{var}(Y) X (X'X + \kappa Id_p)^{-1} \\ &= \sigma^2 (X'X + \kappa Id_p)^{-1} X'X (X'X + \kappa Id_p)^{-1} \end{aligned}$$

les valeurs propres de $(X'X + \kappa Id_p)^{-1}$ étant inférieures à celles de $(X'X)^{-1}$, la variance $\text{var}(\widehat{\theta}_{ridge})$ est donc inférieure à $(X'X)^{-1} X'X (X'X)^{-1} = (X'X)^{-1}$, qui est celle de l'estimateur des moindres carrés.

— Risque

$$\begin{aligned} \mathbb{E}[\widehat{\theta}_{ridge} - \theta]\mathbb{E}[\widehat{\theta}_{ridge} - \theta]' + \text{var}(\widehat{\theta}_{ridge}) &= \kappa^2 (X'X + \kappa Id_p)^{-1} \theta\theta' (X'X + \kappa Id_p)^{-1} \\ &\quad + \sigma^2 \kappa^2 (X'X + \kappa Id_p)^{-1} X'X (X'X + \kappa Id_p)^{-1}) \\ &= (X'X + \kappa Id_p)^{-1} [\kappa^2 \theta\theta' + \sigma^2 X'X] (X'X + \kappa Id_p)^{-1} \end{aligned}$$

d'où

$$\begin{aligned} EQM(\widehat{\theta}_{ridge}) &= tr((X'X+\kappa Id_p)^{-1}[\kappa^2\theta\theta' + \sigma^2 X'X](X'X+\kappa Id_p)^{-1}) \\ &= \sum_{j=1}^{p} \frac{\sigma^2\lambda_j + \kappa^2[P'\theta]_j^2}{(\lambda_j+\kappa)^2} \end{aligned}$$

où P est la matrice de passage telle que $X'X = Pdiag(\lambda_j)P'$. Or, $tr(EQM(\widehat{\theta}_{MC})) = \sigma^2 tr((X'X)^{-1}) = \sigma^2\sum_{j=1}^{p}\lambda_j^{-1}$. Si certaines des valeurs propres λ_j sont nulles (ou proches de zéro), la trace de l'EQM est infinie pour l'estimateur non régularisé.

Propriété 10. *La régression ridge est plus précise que celle des MC quand $\kappa \leq 2\sigma^2/\theta'\theta$, et ce pour chaque composante de θ.*

Preuve. Comme on veut comparer composante par composante, on part de la définition matricielle

$$\begin{aligned} EQM(\widehat{\theta}_{MC}) &= \sigma^2(X'X)^{-1} \\ &= \sigma^2(X'X+\kappa Id_p)^{-1}(X'X + 2\kappa Id_p + \kappa^2(X'X)^{-1})(X'X+\kappa Id_p)^{-1} \end{aligned}$$

d'où

$$EQM(\widehat{\theta}_{MC}) - EQM(\widehat{\theta}_{ridge}) = \kappa\sigma^2(X'X+\kappa Id_p)^{-1}\underbrace{(2(Id_p - \frac{\kappa}{2\sigma^2}\theta\theta') + \kappa(X'X)^{-1})}_{M}(X'X+\kappa Id_p)^{-1}$$

qui est une matrice définie positive si et seulement si la matrice $M = 2(Id_p - \frac{\kappa}{2\sigma^2}\theta\theta') + \kappa(X'X)^{-1}$. Maintenant, si u est un vecteur de dimension p, alors uu' est une matrice symétrique donc diagonalisable, de valeurs propres positives ou nulles. Sa trace $tr(uu') = tr(u'u) = ||u||^2 > 0$. De plus, u est vecteur propre associé à la valeur propre $u'u = tr(uu')$

$$(uu')u = u(u'u) = (u'u)u = ||u||^2 u$$

Il y a donc $p-1$ valeurs propres nulles. Soit P la matrice de passage dans la base rendant uu' diagonale : $uu' = PDP'$ avec $D = diag(1-u'u, 0, \ldots, 0)$. On a $Id_p - uu' = PP' - PDP' = Udiag(1-u'u, 1, \ldots, 1)$. Les valeurs propres sont donc toutes positives dès que $u'u \leq 1$. Cette propriété appliquée à la matrice $Id_p - \frac{\kappa}{2\sigma^2}\theta\theta'$ donne la condition suffisante

$$\frac{\kappa}{2\sigma^2}\theta'\theta < 1$$

$\Diamond$

11.1.2 Régression ridge : en pratique

Le coefficient d'intercept n'est en général jamais inclus dans la contrainte de norme, ainsi on commence par centrer et réduire les variables X_j et Y

$$\tilde{X}_j = (X_j - \bar{X}_j\mathbb{1}_n)/\widehat{\sigma}_j; \;\; \widehat{\sigma}_j^2 = \frac{1}{n}\sum_i (X_{ij} - \bar{X}_j)^2$$

Puis, κ étant fixé, on calcule l'estimateur ridge

$$\widehat{\theta}_{ridge}(\kappa) = (\tilde{X}'\tilde{X} + \kappa Id_p)^{-1}\tilde{X}'\tilde{Y}$$

et on en déduit les valeurs ajustées

$$\widehat{Y}_{ridge}(\kappa) = \widehat{\sigma}_Y^2[\tilde{X}\widehat{\theta}_{ridge}(\kappa)] + \bar{Y}\mathbb{1}_n$$

Le choix de κ est adaptatif, il est construit à partir des données. Plusieurs méthodes sont possibles :

- *graphiquement* : tracer $\widehat{\theta}_{ridge}(\kappa)$ en fonction de κ
  $\hookrightarrow$ $\tilde{\kappa}$ est la plus petite valeur avant laquelle les coefficients "plongent" vers 0
- *analytiquement* (Hoerl, 1975) : $\tilde{\kappa} = \frac{p\widehat{\sigma}^2}{\widehat{\theta}'_{ridge}\widehat{\theta}_{ridge}}$
- par *apprentissage/validation* : on sépare l'échantillon (X, Y) en un échantillon d'apprentissage (X_a, Y_a) que l'on centre et un échantillon de validation (X_v, Y_v). Sur l'échantillon d'apprentissage (X_a, Y_a), on calcule la régression ridge pour une grille de $\kappa \in [0; \kappa_{max}]$, ce qui permet d'obtenir les estimations de $\widehat{\theta}_{ridge}$ et on en déduit celle de la variable à expliquer $\widehat{Y}_{ridge}$

  $$\widehat{Y}_{v,ridge}(\kappa) = \widehat{\sigma}_{a,Y} \sum_{j=1}^{p} \frac{X_{v,j} - \bar{X}_{a,j}\mathbb{1}_{n_v}}{\sigma_{a,j}} \widehat{\theta}_{ridge,j}(\kappa) + \bar{Y}_a \mathbb{1}_{n_v}$$

  On choisit $\tilde{\kappa}$ minimisant l'erreur quadratique de prévision observée

  $$PRESS(\kappa) = ||\widehat{Y}_{v,ridge}(\kappa) - Y_v||^2$$

- par *validation croisée* : une fois le critère mesurant la qualité du modèle choisi, on l'applique successivement sur chacun des plis de l'échantillon. Si la validation est leave one out, il y a autant de plis que d'observations et le critère PRESS s'écrit en fonction de la diagonale de la matrice de projection . On cherchera donc κ qui minimise le critère :

  $$\tilde{\kappa} = \arg\min_{\kappa} \sum_i \left(\frac{y_i - \widehat{y}_{ridge,i}(\kappa)}{1 - h_{jj}(\kappa)} \right)^2$$

11.2 Régression lasso

La régression **lasso** utilise la norme ℓ_1 : $||\theta||_1 = \sum_j |\theta_j|$ à la place de la norme ℓ_2 de la régression ridge. Ainsi on cherche

$$\tilde{\theta} = \arg \min_{\theta \in \mathbb{R}^p; ||\theta||_1 \leq \delta} ||Y - X\theta||^2,$$

soit, résoudre le problème de pénalisation

$$\tilde{\theta}(\kappa) = \arg \min_{\theta \in \mathbb{R}^p} ||Y - X\theta||^2 - \kappa||\theta||_1.$$

Si $\kappa \geq ||X'Y||_\infty = \max_j |[X'Y]_j|$ où $[X'Y]_j$ est la j-ème colonne du vecteur $X'Y$ de $\mathbb{R}^p$, alors $\tilde{\theta}(\kappa) = 0$ est une solution. Il y a des valeurs finies de κ pour lesquelles toutes les composantes de θ sont nulles, ie il n'y a aucune variable retenue. Au fur et à mesure de la décroissance de κ, une variable est ajoutée, la plus corrélée avec Y si les données sont centrées réduites, puis d'autres sont successivement ajoutées, les coefficients des variables déjà utilisées étant alors modifiés.

11.2.1 Lasso en pratique

— Centrer et réduire les variables (X, Y) $(\tilde{X}, \tilde{Y})$ et ajuster le modèle sans coefficient constant
— Le modèle de prévision est

$$\widehat{Y}(\tilde{\kappa}) = \widehat{\sigma}_Y^2 \, \tilde{X}\widehat{\theta}(\tilde{\kappa}) \; + \; \bar{Y}\mathbb{1}_n$$

— $\tilde{\kappa}$ (ou $\tilde{\delta}$) sont choisis grâce aux données, de façon graphique, par méthode analytique ou par validation croisée

11.2.2 Comparaison Lasso/Ridge

La différence entre les deux procédures est illustrée sur un exemple figure 11.1 où les trajectoires des coordonnées de $\widehat{\theta}$ sont représentées en fonction de la norme ℓ_1 de $\widehat{\theta}$. Sur le même jeu de données, la procédure lasso sélectionne (ou dé-sélectionne) des variables, tandis que la procédure ridge augmente progressivement la valeur de chacune des coordonnées.

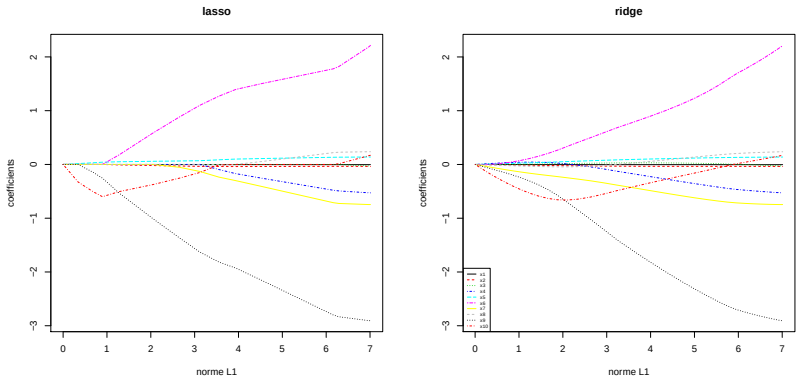


FIGURE 11.1 – Régression lasso et ridge sur un exemple

La figure 11.2 empruntée à James et al. (2013) permet d'éclairer le comportement des deux estimateurs. La solution des moindre carrés est symbolisée par le point noté $\widehat{\theta}$. Le carré bleu représente la contrainte ℓ_1 et le disque bleu la contrainte ℓ_2. Sans contrainte, c'est à dire si le carré et le disque étaient suffisamment grands, ils contiendraient l'EMC $\widehat{\theta}$ et les estimateurs ridge et lasso seraient identiques à l'EMC. Mais ce n'est pas le cas de la figure et la solution est située au point de tangence entre le disque pour ridge (ou le carré pour lasso) et une ellipse représentant l'iso-somme des carrés résiduels. La solution lasso favorise en général les cas de nullité de composante, comme sur la figure où $\tilde{\theta}_1 = 0$, et donc sélectionne des variables. Elle a un caractère plus explicatif que la régression ridge.

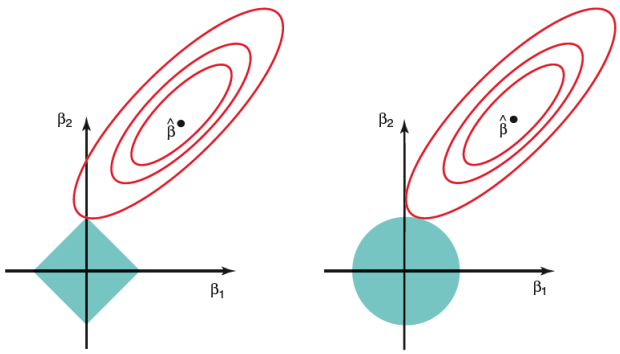


FIGURE 11.2 – Régression lasso (à gauche) et ridge (à droite)

Chapitre 12

SVM

Jusqu'à présent, nous avons défini une *modélisation* permettant d'expliquer les données, et d'en déduire un classifieur. Nous allons dans ce chapitre prendre un point de vue purement *optimisation* : il s'agit de rechercher des *frontières* de séparation (dans un premier temps linéaires) entre les classes en minimisant le taux de mauvais classement. Ceci s'apparente à la recherche d'hyperplans séparateurs. L'utilisation de l'astuce du *noyau* permet d'étendre à des frontières non linéaires dans l'espace initial (mais linéaires dans l'espace dans lequel on projette les observations). Les extensions à la régression sont possibles.

L'acronyme vient de l'anglais *Support Vector Machine* ou machines à vecteurs supports. Les vecteurs supports sont des observations particulières qui définissent les frontières de séparation entre les classes. On traitera le cas du SVM binaire (à observations $+1$ et -1), qui s'étend à plus de deux classes.

Bibliographie Hastie et al. (2001), James et al. (2013), Azencott (2022)

Note Dans tout ce chapitre, x désigne un vecteur colonne et x' sa transposée.

12.1 Classification par hyperplan

12.1.1 Séparabilité linéaire

Définition 27 (Séparabilité linéaire)**.** *Soit $\mathcal{D} = \{(x_i, y_i)]_{i=1,\dots,n}\}$ un jeu de données de n observations, $x_i \in \mathbb{R}^p, y_i \in \{1, -1\}$. $\mathcal{D}$ est linéairement séparable s'il existe un hyperplan affine de $\mathbb{R}^p$ défini par les paramètres $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$*

$$\mathcal{H} = \{x \in \mathbb{R}^p : \beta_0 + \beta' x = 0\}$$

et séparant exactement les points en deux groupes d'étiquettes y_i homogènes.

On se rend compte en regardant la figure 12.1 qu'il n'y a pas unicité de l'hyperplan séparateur s'il existe (plusieurs droites noires permettent de séparer les données). Quand un hyperplan séparateur existe, la marge du plan séparateur est la distance le séparant de l'observation la plus proche.

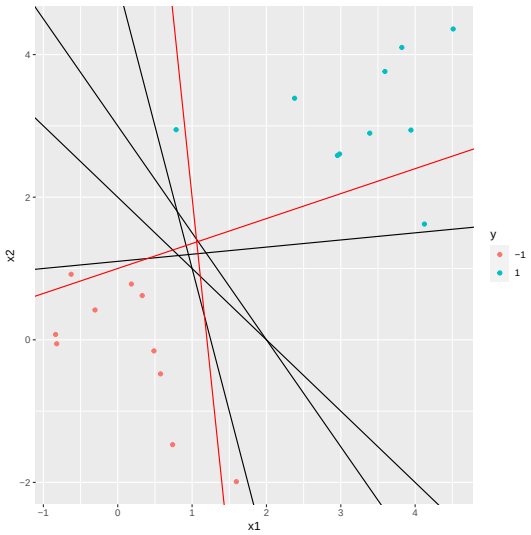


FIGURE 12.1 – Les droites sont des hyperplans de $\mathbb{R}^2$, les noires sont séparatrices des points rouges et bleus

Rappel Distance signée d'un point à une droite.
Soit $\mathcal{H} = \{x \in \mathbb{R}^p : \beta_0 + \beta' x = 0\}$ un hyperplan de $\mathbb{R}^p$.
Soit $f(x) = \beta_0 + \beta' x$

— Pour tout couple de points x_1, x_2 de $\mathcal{H}$, on a
$\beta'(x_1 - x_2) = 0$ et $\beta^* = \beta / \parallel \beta \parallel$ est le vecteur normal à la surface de $\mathcal{H}$

— Soit $x_0 \in \mathcal{H}$, la projection d'un point x sur $\mathcal{H}$

$$\beta' x_0 = -\beta_0$$

— la distance signée d'un point $x \in \mathbb{R}^p$ à $\mathcal{H}$ est :

$$\begin{aligned} \beta^{*\prime}(x - x_0) &= \frac{1}{\parallel \beta \parallel}(\beta_0 + \beta' x) \\ &= \frac{1}{\parallel \beta \parallel} f(x) \end{aligned}$$

Donc $f(x)$ est proportionnel à la distance signée de x à l'hyperplan $\mathcal{H}$ défini par $f(x) = 0$
En particulier, l'appartenance d'un point à l'un des deux sous-espaces de part et d'autre de la frontière se calcule aisément suivant le signe de $f(x)$.

12.1.2 Algorithme perceptron

L'algorithme perceptron de Rosenblatt (1958) permet d'estimer la frontière linéaire de plus grande marge quand les données sont linéairement séparables. On commence par définir la règle de classification suivante :

$$\widehat{y} = \text{signe } f(x)$$

Quand une observation est mal classée : y et $\widehat{y}$ de signe opposé. Soit $\mathcal{G}$ l'ensemble des observations mal classées. On chercher un hyperplan séparateur minimisant la distance des points

mal classés à la frontière de décision. Le critère à minimiser est :

$$D(\beta_0, \beta) = -\sum_{i \in \mathcal{G}} y_i f(x).$$

Son gradient est, quand $\mathcal{G}$ est fixé :

$$\partial D(\beta_0, \beta)/\partial \beta_0 = -\sum_{i \in \mathcal{G}} y_i$$
$$\partial D(\beta_0, \beta)/\partial \beta = -\sum_{i \in \mathcal{G}} y_i x_i$$

Quand le nombre d'observations est grand, il peut être plus efficace de résoudre ce problème de minimisation par *descente de gradient* : boucler sur l'ensemble des observations, et pour chacune, faire un pas de descente. Si l'observation est bien classée, on ne fait rien, si non, l'intercept et la pente sont modifiés. Si le nombre d'observations est très important, on peut choisir de n'en tirer uniformément qu'un sous-ensemble. Un parcours de toutes les observations (ou de toutes les observations d'un ensemble tiré aléatoirement) s'appelle une époque (*epoch*). On itère les époques jusqu'à ce qu'il n'y ait plus de mise à jour ou au bout d'un nombre d'époques fixé.

Algorithme de descente de gradient On pose $\mathrm{w} = (\beta_0, \beta')'$ et $\mathrm{x} = (1, x')' \in \mathbb{R}^{p+1}$
- Initialisation $\mathrm{w} = \mathrm{w}_0$
- Tant qu'il y a des mises à jour
  - Pour $i = 1, \ldots, n$:
    - $\widehat{y}_i = \mathrm{signe}(\mathrm{w}'\mathrm{x}_i)$
    - si $\widehat{y}_i \neq y_i$ alors : $\mathrm{w} \leftarrow \mathrm{w} + 1 y_i \mathrm{x}_i$
  - Fin pour
- Fin Tant que
- retourner w

Ainsi, le perceptron apprend de ses erreurs.

Convergence de l'algorithme Si les données sont linéairement séparables, l'algorithme du perceptron converge en un nombre fini d'étapes [Novikoff 1962]. Mais il y a des solutions multiples en cas de séparation, ce qui est un vrai inconvénient pour la prédiction de nouvelles observations. D'autre part, on note une lenteur de convergence quand l'écart de séparation est faible. Enfin, il n'y a pas convergence en cas de non séparation : il faut arrêter l'exécution après un certain nombre d'époques.

12.1.3 Classifieur à marge maximale

Afin d'améliorer la méthode précédente, Vapnik (1996) propose de rechercher un hyperplan séparateur maximisant la distance au point le plus proche de chaque classe. Ce problème a maintenant une unique solution si les données sont linéairement séparables et assure ainsi une meilleure performance de généralisation. C'est un problème d'optimisation convexe dans $\mathbb{R}^{p+1}$ sous n contraintes linéaires d'inégalité :

$$\arg \max_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \|\beta\|=1} C \text{ t.q. } y_i f(x_i) \geq C, i = 1, \ldots, n.$$

Formulation primale De façon équivalente, on l'écrit sous la forme d'une optimisation dans $\mathbb{R}^{p+1}$

$$\arg \min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2} \parallel \beta \parallel^2 \text{ t.q. } y_i f(x_i) \geq 1, i = 1, \ldots, n.$$

Soient β_0^* et β^* solutions du problème primal, la frontière de décision est

$$f^*(x) = \beta_0^* + \beta^{*'} x = 0.$$

Formulation duale C'est une optimisation dans $\mathbb{R}^n$ après introduction de n multiplicateurs de Lagrange $\alpha_i \in \mathbb{R}^+$:

$$\arg \max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i' x_j$$

$$\text{t.q. } \sum_{i=1}^n \alpha_i y_i = 0; \alpha_i \geq 0, i = 1, \ldots, n.$$

satisfaisant les conditions *Karush-Kuhn-Tucker*

- admissibilité primale : $y_i f(\mathbf{x}_i) \geq 1$
- admissibilité duale : $\alpha_i \geq 0$
- complémentarité : $\alpha_i [y_i (x_i' \beta + \beta_0) - 1] = 0$
- stationnarité : $\sum_{i=1}^n \alpha_i y_i = 0$ et $\beta = \sum_{i=1}^n \alpha_i y_i x_i$

Preuve. Pour déterminer la forme duale de ce problème d'optimisation sous contraintes, on introduit n multiplicateurs de Lagrange α_i associés aux n contraintes $y_i f(\mathbf{x}_i) \geq 1$ et on écrit le lagrangien $\mathcal{L} : \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}_+^2 \to \mathbb{R}$:

$$\mathcal{L}(\beta, \beta_0, \alpha) = \frac{1}{2} ||\beta||^2 - \sum_{i=1}^n \alpha_i [y_i (x_i' \beta + \beta_0) - 1].$$

Soit $q(\alpha) = \inf_{\beta, \beta_0} \mathcal{L}(\beta, \beta_0, \alpha)$. On en déduit le problème dual

$$\max_{\alpha} q(\alpha). \tag{12.1}$$

Or $\mathcal{L}$ est convexe en β, donc minimal quand son gradient s'annule, d'où

$$\beta = \sum_{i=1}^n \alpha_i y_i x_i.$$

De plus, $\mathcal{L}$ est affine en β_0. Son minimum vaut donc $-\infty$ sauf si

$$\sum_{i=1}^n \alpha_i y_i = 0.$$

$q(\alpha)$ est donc maximisée dans ce cas. Ainsi, le problème (12.1) devient

$$\arg \max_{\alpha \in \mathbb{R}^n} \frac{1}{2} || \sum_i \alpha_i y_i x_i ||^2 - \sum_{i=1}^n \alpha_i y_i (\sum_{j=1}^n \alpha_i \alpha_j y_i x_j x_i') - \sum_i \alpha_i y_i \beta_0 + \sum_i \alpha_i$$

$$\text{t.q. } \sum_{i=1}^n \alpha_i y_i = 0; \alpha_i \geq 0, i = 1, \ldots, n.$$

En effectuant le développement de la norme et en simplifiant, on trouve la formulation énoncée du problème dual. ◊

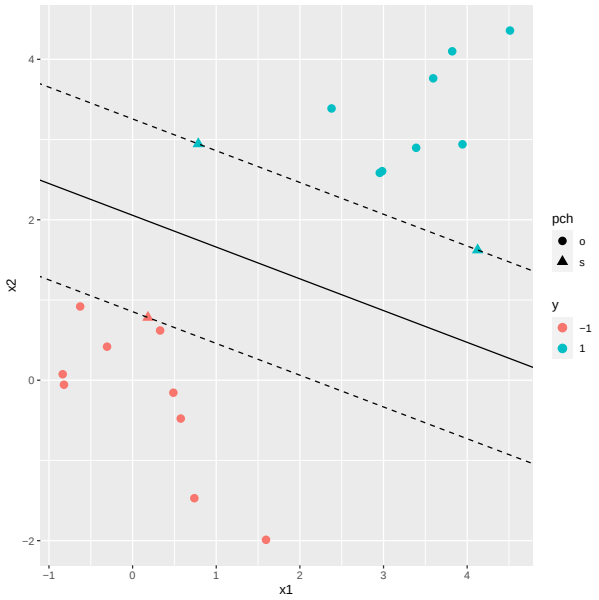


FIGURE 12.2 – Représentation du séparateur linéaire à marge rigide (trait noir plein), et des points supports (triangles). La marge est la distance entre les deux droites pointillées.

Frontière de décision Pour résoudre le problème primal, on résout le problème dual et on obtient α^*, d'où $\beta^* = \sum_i \alpha_i^* y_i x_i$. Les vecteurs supports positifs sont situés du côté de l'hyperplan positif et vérifient $x_i'\beta^* + \beta_0 = 1$ tout en minimisant leur distance à l'hyperplan séparateur soit $\min_{i;y_i>0}(x_i'\beta^*) + \beta_0 = 1$ et $\beta_0^* = 1 - \min_{i;y_i>0}(x_i'\beta^*)$

$$f^*(x) = \beta_0^* + (\sum_{i=1}^{n} \alpha_i^* y_i x_i)'x \tag{12.2}$$

Interprétation géométrique Les conditions de KKT permettent de caractériser la relation entre α^* et (β_0^*, β^*)

— la condition d'écart complémentaire signifie que

$$\alpha_i^*(y_i f^*(x_i) - 1) = 0 \; \forall i = 1, \ldots, n$$

— si $\alpha_i^* > 0$, alors $y_i f^*(x_i) = 1$: x_i est sur la frontière de la *zone d'indécision* (sans observation) de part et d'autre de l'hyperplan séparateur

— si $y_i f^*(x_i) - 1 > 0$, x_i est extérieur à la zone de séparation et $\alpha_i^* = 0$.

Définition 28. *β^* est une combinaison linéaire de points frontières, appelés vecteurs supports. Ces vecteurs supports sont les observations x_i associées à un multiplicateur de Lagrange α_i^* non nuls.*

La figure 12.2 représente sous forme de triangle les observations du jeu de données qui sont les vecteurs supports du séparateur linéaire à marge rigide : deux sont issus du groupe bleu, un du groupe rouge.

SVC

Et quand il n'existe pas de plan exactement séparateur ? : quel que soit l'hyperplan séparateur, il y aura toujours des points mal classés. On va montrer que les vecteurs supports qui définissent l'hyperplan séparateur sont les points dans la zone d'indécision et les points mal classés hors de la zone d'indécision (cf. Figure 12.3).

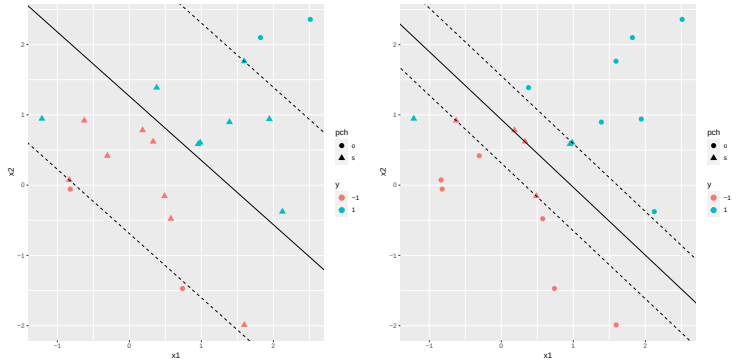


FIGURE 12.3 – Deux plans séparateurs : défini avec une grande marge (à gauche, et donc de nombreux points supports) ou une faible marge (à droite, peu de points supports). Les points supports sont les observations qui sont situés dans la zone d'indécision, entre les deux droites en pointillés ou les points mal classés hors de la zone d'indécision.

On parle de *Support Vector Classifier* ou *SVM à marge souple*. Pour optimiser les frontières de la zone d'indécision, on définit le coût d'une observation $(\mathbf{x}, y)$ d'être du mauvais côté de sa marge en utilisant la fonction de coût *hinge* :

$$\ell(f(x), y) = [1 - yf(x)]_+ = \max(0, 1 - yf(x))$$

Le coût des points mal classés pénalise l'inverse du carré de la marge :

$$\arg \min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p} \frac{1}{2} \parallel \beta \parallel^2 + \gamma \sum_i \ell(f(x_i), y_i)$$

où $\gamma \in \mathbb{R}^+$ est un hyperparamètre de la méthode. On introduit alors les variables d'*ajustement* (ou d'écart ou *slack variables* en anglais) : $\xi_i = \ell(f(x_i), y_i)$. On écrit alors le problème sous sa formulation primale.

Formulation primale

$$\arg \min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \xi \in \mathbb{R}^2} \frac{1}{2} \parallel \beta \parallel^2 + \gamma \sum_i \xi_i \text{ t.q. } \begin{cases} y_i f(x_i) \geq 1 - \xi_i, \forall i \\ \xi_i \geq 0, \forall i \end{cases}$$

Formulation duale

$$\arg \max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i' x_j$$
$$\text{t.q. } \sum_i \alpha_i y_i = 0;\ 0 \leq \alpha_i \leq \gamma, \forall i.$$

A partir des conditions de KKT, il y a deux conditions d'écart complémentaires :

$$\alpha_i^*(y_i f^*(x_i) - 1) = 0 \text{ et } (\gamma - \alpha_i^*)\ell(f^*(x), y) = 0, \quad \forall i$$

Interprétation géométrique On interprète les multiplicateurs de Lagrange, voir des exemples sur la figure 12.4 :

— $\alpha_i^* = 0$: β^* vérifie $\ell(f^*(x_i), y_i) = 0$ et $0 \geq 1 - y_i f^*(x_i)$
Dans ce cas, x_i est à l'extérieur de la zone d'indécision

— $0 < \alpha_i^* < \gamma$: $\xi_i = 1 - y_i f^*(x_i) = 0$.
Alors, x_i est vecteur support sur la frontière de la zone d'indécision

— $\alpha_i^* = \gamma$: $\xi_i = 1 - y_i f^*(x_i) > 0$
Alors, x_i est vecteur support du mauvais côté de la frontière de la zone d'indécision

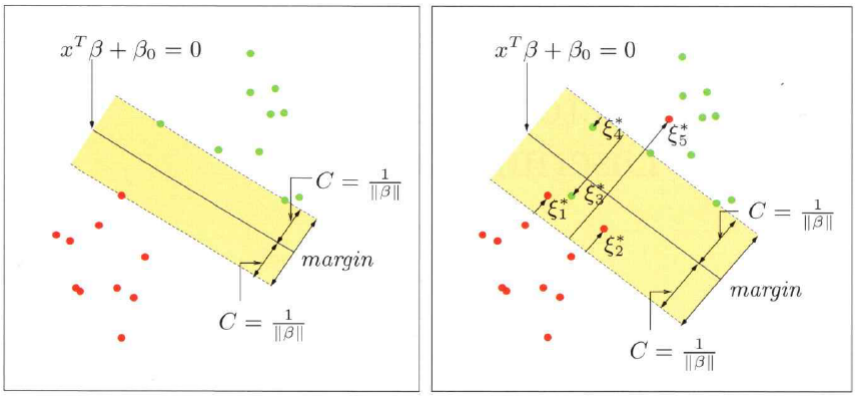


FIGURE 12.4 – A gauche, cas SVM à marge dure : il existe un hyperplan séparateur et les vecteurs supports sont à la frontière de la zone d'indécision. A droite, cas SVM à marge souple : il n'existe pas d'hyperplan séparateur ; les vecteurs supports sont les observations à l'intérieur de la zone d'indécision, qu'elles soient mal classées (observation associée à la variable ξ_3^*) ou bien classées (observations associées aux variables $\xi_1^*, \xi_2^*, \xi_4^*$) d'une part, les observations mal classées à l'extérieur de la zone d'indécision (observation associée à la variable ξ_5^*) d'autre part. Figure issue de Hastie et al. (2001)

12.2 Cas non linéaire : SVM à noyau

On espère que la projection des observations dans un espace (en général plus grand) permettra de faire apparaître des frontières linaires dans ce nouvel espace. Imaginons deux classes concentriques. Il n'y a pas d'hyperplan séparateur et les SVM seront mauvais dans cette configuration. Mais si on considère les points en cordonnées polaires, il existe une séparation linéaire ! On commence par définir la notion d'espace de redescription.

Définition 29. *On appelle espace de redescription l'espace de Hilbert $\mathcal{F}$ dans lequel il est souhaitable de décrire les données, au moyen d'une application $\phi : \mathcal{X} \to \mathcal{F}$, pour y entraîner un SVM sur les images des observations du jeu d'entraînement.*

La formulation duale du problème de minimisation à marge souple est écrit dans l'espace de

redescription :

$$\arg\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \langle\phi(x_i),\phi(x_j)\rangle_{\mathcal{F}}$$
$$\text{t.q. } \sum_i \alpha_i y_i = 0;\ 0 \leq \alpha_i \leq C, \forall i.$$

La frontière de décision est donnée par 12.2

$$f(x) = \sum_i \alpha_i^* y_i \langle\phi(x_i),\phi(x)\rangle_{\mathcal{F}} + \beta_0^*.$$

Ainsi, les images de observations dans $\mathcal{F}$ apparaissent uniquement dans des produits scalaires sur $\mathcal{F}$. On appelle *SVM à noyau* la solution du problème d'optimisation suivant

$$\arg\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \kappa(x_i,x_j)$$
$$\text{t.q. } \sum_i \alpha_i y_i = 0;\ 0 \leq \alpha_i \leq C, \forall i.$$

où κ est un noyau. La solution est de de frontière de décision

$$f(\mathbf{x}) = \sum_i \alpha_i^* y_i \kappa(x_i,x) + \beta_0^*.$$

Pour entraîner le SVM, il n'est pas besoin de faire les calculs dans $\mathcal{F}$, il suffit de connaître le noyau κ : c'est l'astuce du noyau.

12.2.1 Noyaux

Définition 30. *On appelle noyau toute fonction κ de deux variables s'écrivant sous la forme d'un produit scalaire des images dans un espace de Hilbert de ses variables. κ est une fonction continue, symétrique, semi-définie positive.*

Ainsi :

$$\forall n \in \mathbb{N}, \forall (x_1,\ldots,x_n) \in \mathcal{X}^n, \forall (a_1,\ldots,a_n) \in \mathbb{R}^n,$$
$$\sum_{i,j=1}^{n} a_i a_j \kappa(x_i,x_j) \geq 0$$

La matrice $K \in R^{n\times n}$ t.q. $k_{ij} = \kappa(x_i,x_j)$ est appelée *matrice de Gram*. Elle est semi-définie positive. Le théorème suivant permet d'indiquer que toute fonction semi-définie positive permet de définir un noyau :

Théorème 10 (Moore-Aronszajn)**.** *Pour toute fonction semi-définie positive $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, il existe un espace de Hilbert $\mathcal{F}$ et une application $\psi : \mathcal{X} \to \mathcal{F}$, telle que pour tout $x, y \in \mathcal{X}\times\mathcal{X}$, $\kappa(x,y) = \langle\psi(x),\psi(y)\rangle_{\mathcal{F}}$.*

Exemples :
- noyau quadratique : $\kappa(x,y) = (\langle x,y\rangle + c)^2$, où $c \in \mathbb{R}^+$
- noyau polynomial : $\kappa(x,y) = (\langle x,y\rangle + c)^d$, où $c \in \mathbb{R}^+$, $d \in \mathbb{N}$
- noyau radial gaussien : $\kappa(x,y) = \exp(-\tau \parallel x-y \parallel^2)$, $\tau > 0$
- neural network : $\kappa(x,y) = \tanh(k_1 < x,y > +k_2)$

Les résultats d'un SVM dépendent d'hyperparamètres , deux dans le cas du noyau gaussien : coût γ et paramètre τ du noyau. Voir l'influence du paramètre de coût γ sur la figure 12.5.

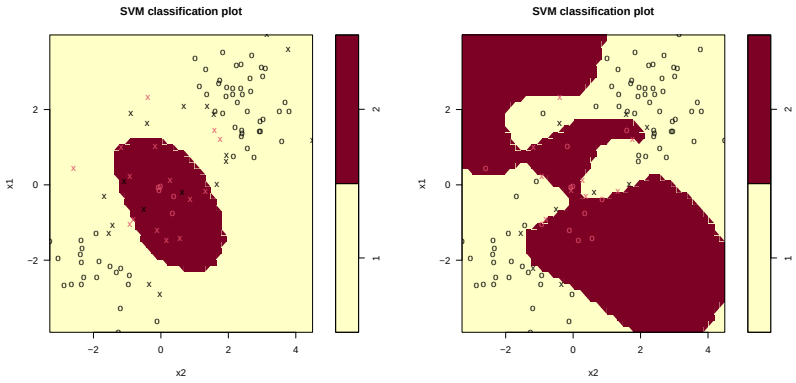


FIGURE 12.5 – Influence du coût γ sur les résultats (à paramètre du noyau constant)

Annexe A

Rappels de statistique inférentielle

Ce chapitre rappelle les principes de l'estimation ponctuelle, l'estimation par intervalle de confiance et le test d'hypothèses. Ces notions sont traitées en détail dans Delmas (2010) ou Saporta (2006) ou le cours de première année, et nous prenons pour exemple l'estimation de l'espérance d'une variable aléatoire gaussienne.

Bibliographie Lejeune (2010), Delmas (2010), Pagès (2005), Saporta (2006), Cornillon et autres (2008), Rivoirard et Stoltz (2009), cours de première année

A.1 Modélisation statistique

Afin de pouvoir inférer les propriétés d'un échantillon à celle de la loi qui aurait généré les données observées, il est nécessaire de définir un modèle

- Le rôle du probabiliste est d'étudier les propriétés d'un échantillon, et des grandeurs qui y sont liées, quand sa loi est connue.
- Le rôle du statisticien est inverse : à partir d'observations de loi inconnue, il décide des propriétés de cette loi en calibrant le risque de la décision.

La modélisation probabiliste est à la base de toute inférence statistique. Modéliser l'expérience, c'est proposer une loi théorique pour l'échantillon $X = (X_1, \ldots, X_n)$.

Définition 31. *Un **modèle statistique** est la donnée d'un espace $\mathcal{X}^n$ mesuré par une tribu $\mathcal{A}^n$ et une famille de lois de probabilité $(\mathcal{P}_\theta^n)_{\theta\in\Theta}$. Le modèle associé est l'espace probabilisé noté*

$$\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathcal{P}_\theta^n, \theta \in \Theta)$$

*Quand il existe $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit **paramétrique**. Sinon, il est **non paramétrique**.*

Dans le cas paramétrique, la forme de la loi n'est connue qu'à la valeur du paramètre θ près, à inférer avec l'observation de l'échantillon. Dans le cas non paramétrique, la loi est considérée comme un élément d'un espace de dimension infinie, et il s'agira d'estimer ses coordonnées dans cette base.

Définition 32. *Un* ***échantillon*** *de loi $\mathcal{P}^n$ est un ensemble de n variables aléatoires $X_1, \ldots, X_n$ suivant la loi $\mathcal{P}^n$.*

Dans le cas d'un tirage indépendant dans une population infinie, ou le cas d'un échantillonnage avec remise dans une population finie, la loi de l'échantillon se factorise en produit des lois de chacune des composantes de l'échantillon.

Définition 33. *On appelle* ***n-échantillon i.i.d.*** *le modèle statistique de n composantes indépendantes et de même loi $\mathcal{P}_\theta$ (identiquement distribuées)*

$$\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathcal{P}_\theta^{\otimes n}, \theta \in \Theta).$$

La loi $\mathcal{P}_\theta$ est parfois appelée ***loi mère*** *de l'échantillon. On dit que l'échantillon est de* ***taille*** *n.*

Définition 34. *Une* ***observation*** *est une composante de l'échantillon, variable aléatoire X_i de loi $\mathcal{P}_\theta$. Les* ***données*** *sont les réalisations (valeurs) $x_1, \ldots, x_n$ prises par l'échantillon $X_1, \ldots, X_n$.*

Par abus de langage, on appelle parfois observation le résultat $X_i(\omega) = x_i$ de cette variable aléatoire. Le choix de la loi mère dépend du phénomène observé : si le phénomène est binaire, la loi de Bernoulli $\mathcal{B}(1, \mu)$ est un choix naturel :

$$\mathbb{P}(X = 1) = \mu; \;\; \mathbb{P}(X = 0) = 1 - \mu.$$

Si on étudie une variable quantitative, la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, de densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; \;\; x \in \mathbb{R}$$

sera souvent utilisée : par exemple, pour l'étude du rendement de parcelles de blé.

Les thèmes de la statistique paramétrique sont l'estimation ponctuelle, l'estimation par intervalle de confiance et le test d'hypothèses dont les principes sont rappelés dans ce chapitre.

A.2 Estimation ponctuelle

Définition 35. *Soit $X_1, \ldots, X_n$ un n-échantillon d'une loi $\mathcal{P}(\theta)$. Un estimateur d'une fonction déterministe $\nu(\theta)$ est une variable aléatoire T_n, application mesurable de l'échantillon*

$$T_n = t(X_1, \ldots, X_n).$$

On la note souvent $\widehat{\nu} = T_n$.

Une estimation ponctuelle de $\nu(\theta)$ est calculée à partir d'une réalisation $t_n = t(x_1, \ldots, x_n)$ de T_n.

Exemple L'estimateur empirique $\widehat{\mu}$ de l'espérance μ d'une loi est

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Il est souvent noté $\bar{X}$, et sa réalisation $\bar{x}$.

A.2.1 Propriétés d'un estimateur

Soit $\widehat{\nu}$ estimateur d'une quantité $\nu(\theta)$ dépendant de la loi $\mathcal{P}(\theta)$ (son espérance, sa variance, un quantile,...) :

— $\widehat{\nu}$ est sans biais ssi

$$Biais(\widehat{\nu}) = \mathbb{E}(\widehat{\nu} - \nu(\theta)) = E(\widehat{\nu}) - \nu(\theta) = 0$$

— $\widehat{\nu}$ est consistant ssi $\widehat{\nu}$ tend en probabilité vers $\nu(\theta)$ quand $n \to \infty$:

$$\forall \epsilon, \lim_{n\to\infty} P(|\widehat{\nu} - \nu(\theta)| > \epsilon) = 0$$

— $\widehat{\nu}$ est fortement consistant ssi

$$P(\lim_{n\to\infty} \widehat{\nu} = \nu(\theta)) = 1$$

Une condition suffisante de convergence presque sûre est

$$\forall \nu > 0, P(\limsup |\widehat{\nu} - \nu(\theta)| > 0) = 0.$$

Exemple L'estimateur empirique $\bar{X}$ de l'espérance est sans biais et fortement consistant. C'est la loi des grands nombres rappelée ci-après.

Théorème 11 (Loi faible des grands nombres)**.** *Si $(X_1, \ldots, X_n)$ est un échantillon i.i.d d'une loi $\mathcal{P}$ de carré intégrable, de variance finie σ^2, sa moyenne empirique $\bar{X}$ satisfait les propriétés suivantes :*

— *C'est un estimateur sans biais de $\mathbb{E}(X)$:*

$$Biais(\bar{X}) = \mathbb{E}(\bar{X}) - \mathbb{E}(X) = 0$$

— *de variance*

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

$(\bar{X})$ converge donc en probabilité, la suite $(\bar{X})$ est faiblement consistance.

On montre que si $\mathbb{E}(|X|)$ est finie, on peut remplacer *en probabilité* par *presque sûrement*, c'est la *loi forte des grands nombres.*

A.2.2 Risque d'un estimateur

L'analyse ne s'arrête pas à une valeur trouvée, le statisticien s'intéresse à calibrer l'erreur qu'il commet en prenant une décision.

Définition 36. *Le risque de l'estimateur $\widehat{\nu}$ est*

$$R(\widehat{\nu}) = \mathbb{E}(\ell(\widehat{\nu}, \nu)),$$

où $\ell(\widehat{\nu}, \nu)$ est une fonction de perte dans $\mathbb{R}^+$.

Exemple Le **risque quadratique** (ou **erreur quadratique moyenne**, *Mean Square Error*) de l'estimateur empirique de l'espérance μ est

$$EQM(\bar{X}) = \mathbb{E}(\bar{X} - \mu)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Il ne dépend pas de l'espérance μ, mais décroît avec le nombre d'observations n et croît avec la variance (variabilité) du phénomène observé σ^2.

- Si la variance est connue égale à σ_0^2 ou $\leq \sigma_0^2$, alors, le risque **a priori** de $\bar{X}$ peut être estimé par $EQM(\bar{X})$. Si on veut garantir $EQM(\bar{X}) \leq \epsilon^2$, on peut le faire en choisissant au moins $n_0 = (\sigma_0/\epsilon)^2$ observations.
- Si on n'a aucune idée de la valeur de σ^2, la planification n'est pas possible, mais on peut utiliser les n expériences pour estimer σ^2, par exemple, $\widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ ou $n\widehat{\sigma}^2/(n-1)$. L'estimation **a posteriori** du risque est $\widehat{\sigma}^2/n$, elle-même soumise à une erreur aléatoire.

De façon générale, EQM dépend de la variance de $\widehat{\nu}$ et de son biais (nul dans le cas de $\bar{X}$), que l'on peut voir comme "l'erreur au long cours" de $\widehat{\nu}$. En effet, en écrivant $(\widehat{\nu} - \nu) = [\widehat{\nu} - E(\widehat{\nu})] + [E(\widehat{\nu}) - \nu]$, on obtient la formule de la décomposition biais-variance :

$$EQM(\widehat{\nu}) = (Biais(\widehat{\nu}))^2 + Var(\widehat{\nu}).$$

Le risque quadratique permet de définir une relation d'ordre partiel sur les estimateurs :

Définition 37. *Un estimateur δ_1 de $\nu(\theta)$ **domine** l'estimateur δ_2 si, pour tout $\theta \in \Theta$,*

$$R_\theta(\delta_1, \nu) \leq R_\theta(\delta_2, \nu)$$

cette inégalité étant stricte pour au moins une valeur de ν.
*Un estimateur est **admissible** s'il n'existe aucun estimateur le dominant. Sinon, il est **inadmissible**.*

Il n'est pas possible en général de minimiser le risque quadratique uniformément : on recherchera donc des estimateurs optimaux dans des sous classes, par exemple celle des estimateurs sans biais : estimateurs UVMB, Uniformément de Variance Minimale parmi les estimateurs sans Biais.

A.2.3 Loi de l'estimateur

Le risque donne une information globale sur la vitesse de convergence de l'estimateur, en $\sqrt{1/n}$ pour l'estimateur de l'espérance. Ce n'est qu'une information partielle de loi de l'estimateur, qui elle est nécessaire pour la définition d'intervalles de confiance et de tests. Si la loi de l'estimateur $\bar{X}$ est connue dans le gaussien (puisque son espérance et sa variance le sont), il n'en est pas de même pour l'estimateur $\bar{X}$ de l'espérance d'une loi donnée. Dans ce cas, le théorème essentiel suivant donne le comportement asymptotique de $\bar{X}$:

Théorème 12 (de limite centrale)**.** *Si $\bar{X}$ est la moyenne empirique d'un n-échantillon d'une loi d'espérance μ et de variance finie σ^2, la suite $\sqrt{n}(\bar{X} - \mu)$ converge en loi vers $\mathcal{N}(0, \sigma^2)$.*

De façon générale, si
- la vitesse de l'estimateur est en $\sqrt{n}$,
- la convergence a lieu en loi,
- la loi limite est gaussienne,

on dit que l'estimateur est asymptotiquement normal. Un estimateur est d'autant meilleur que sa vitesse de convergence est rapide et sa loi limite concentrée autour de 0.

A.3 Construction d'estimateurs

A.3.1 Méthode des moments

Supposons que la quantité à estimer s'écrive en fonction de l'espérance d'une observation X_1 : $\nu(\theta) = \mathbb{E}[g(X_1)]$, où $g(X_1)$ est intégrable. On utilise la convergence presque sûre de la loi des grands nombres

$$\widehat{\nu} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \nu(\theta) \ \ p.s.$$

Les paramètres d'intérêt s'exprimant très souvent à partir des moments de la loi des observations, la méthode propose automatiquement des estimateurs fortement consistants pour l'estimation de ces paramètres. D'où l'estimateur empirique de l'espérance $\bar{X}$, mais aussi celui de la variance empirique

$$\widehat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2$$

quand la loi admet un moment d'ordre 2.

A.3.2 Méthode du maximum de vraisemblance

Elle s'applique dans le cas où toutes les lois du modèle sont dominées par une mesure commune (mesure de Lebesgue, de comptage, ...). Soit f_θ la densité de la loi $\mathcal{P}_\theta$ par rapport à cette mesure commune. La méthode du maximum de vraisemblance consiste à estimer θ comme la valeur $\widehat{\theta}$ maximisant la vraisemblance des observations, ie $\prod_{i=1,\ldots,n} f_\theta(X_i)$ quand les observations sont indépendantes.

A.4 Intervalle de confiance

L'estimation par intervalle de confiance permet de prendre en compte la variabilité de l'estimation, et donc sa fiabilité ou sa précision. On cherche par exemple une borne supérieure $\widehat{\nu}_{sup}$ pour un paramètre ν, pour laquelle on espère fortement (avec une forte probabilité) que ν soit inférieur à cette valeur $\widehat{\nu}_{sup}$, c'est à dire

$$\mathbb{P}[\nu \leq \widehat{\nu}_{sup}] = \mathbb{P}(\nu \in]-\infty, \widehat{\nu}_{sup}]) \geq 1-\alpha \tag{A.1}$$

où α est "petit".

Définition 38. *La variable aléatoire $\widehat{\nu}_{sup} = \widehat{\nu}_{sup}(X)$ définie par (A.1), fonction de l'échantillon X, est appelée* ***borne supérieure de confiance*** *de niveau $1-\alpha$ ou de risque α.*

Pour un tirage amenant à $\widehat{\nu}_{sup} \geq \nu$, $\widehat{\nu}_{sup}$ est effectivement une borne supérieure de ν (que l'on ne connaît pas) ; la perte associée à cette décision est nulle. Pour un tirage de amenant à $\widehat{\nu}_{sup} < \nu$, la décision va être erronée : on va dire que la borne supérieure est $\widehat{\nu}_{sup}$, alors que le vrai paramètre n'est pas dans l'intervalle $(-\infty; \widehat{\nu}_{sup}]$. C'est α qui pilote l'erreur commise : plus α est petit, plus le risque de donner une borne supérieure trop petite est faible, mais moins informatif est l'intervalle donné. A l'extrême, choisir une borne supérieure à l'infini donne une confiance maximum, mais plus du tout d'information !

Exemple Soit $X \sim \mathcal{N}(\mu, \sigma^2)$; $\bar{X}$ estime $\mu = E(X)$, et $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. On a :

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > q_{\mathcal{N}(0,1)}(\alpha)\right) \\ &= \mathbb{P}\left(\mu < \bar{X} + \frac{\sigma}{\sqrt{n}} q_{\mathcal{N}(0,1)}(1-\alpha)\right) \end{aligned}$$

soit $\widehat{\mu}_{sup} = \bar{X} + \frac{\sigma}{\sqrt{n}} q_{\mathcal{N}(0,1)}(1-\alpha)$, où $q_{\mathcal{N}(0,1)}(1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite. Si σ^2 est inconnu, il peut être estimé par son estimateur sans biais $\widehat{\sigma}^2 = \sum_i (X_i - \bar{X})^2/(n-1)$, d'où $\widehat{\mu}_{sup} = \bar{X} + \frac{\widehat{\sigma}}{\sqrt{n}} q_{\mathcal{T}(n-1)}(1-\alpha)$ où $\mathcal{T}(n-1)$ est la loi de Student à $n-1$ degrés de liberté.

Définition 39. *L'intervalle de confiance bilatéral $\{\widehat{\nu}_{inf}, \widehat{\nu}_{sup}\}$ de niveau $1-\alpha$ ou de risque α est un intervalle dont les bornes aléatoires dépendent de l'échantillon, et tel que*

$$\mathbb{P}(\widehat{\nu}_{inf} \leq \nu \leq \widehat{\nu}_{sup}) \geq 1-\alpha.$$

Remarque : Les bornes de l'intervalle sont aléatoires. On ne peut dire si la vraie valeur du paramètre appartient à l'intervalle de confiance estimé. Mais si le calcul de l'intervalle est refait sur différents échantillons indépendants, la vraie valeur du paramètre sera incluse dans l'intervalle en moyenne $(1-\alpha) \times 100\%$ des cas.

La méthodologie utilisée pour calculer l'IC de l'exemple est assez générale, il s'agit de la méthode dite **pivotale** :

- choisir un estimateur
- calculer sa loi en fonction de ν
- trouver une stat pivotale $T(\widehat{\nu})$ dont la loi ne dépend pas de ν
- exprimer les bornes de l'intervalle de confiance en fonction des quantiles de la loi pivotale et de l'estimateur

A.5 Test

Construire un test statistique, c'est définir une **règle de décision** permettant de choisir entre une hypothèse, dite **nulle** H_0, et son **alternative** H_1 en utilisant les résultats d'un échantillon. Il s'agit de répondre, par exemple, aux questions suivantes : La pièce est-elle truquée ? L'utilisation de sels d'argent augmente-t-il significativement le volume des précipitations ? Le niveau de corrosion a-t-il une influence sur la durée de vie d'une pièce mécanique ? Rappelons brièvement la méthodologie de construction d'un test :

1. Définir le **modèle** statistique des données $(\Omega, \mathcal{A}, P_\theta, \theta \in \Theta)$.
2. Définir les **hypothèses** en compétition qui traduisent la question à laquelle il faut répondre :
   - l'hypothèse **hypothèse nulle** H_0, en général sous la forme $\theta \in \Theta_0 \subset \Theta$ dans le cadre paramétrique
   - l'hypothèse **alternative** H_1, en général sous la forme $\theta \in \Theta_1 \subset \Theta$, disjoint de Θ_0.

   Si Θ_0 (ou Θ_1) est réduit à un singleton, l'hypothèse est dite simple, sinon, elle est composite.
3. Construire une statistique de test $T(X)$, et calculer sa loi sous H_0. Cette statistique doit avoir une loi différente sous H_1.

4. Définir la **règle de décision** permettant de choisir entre H_0 et H_1. C'est une fonction δ de $T(X)$ qui vaut 1 sur un ensemble $\mathcal{R}_\alpha$ appelé *région critique* ou *région de rejet* de l'hypothèse nulle (choix de H_1) et 0 sur son complémentaire (choix de H_0) :

$$\delta(T) = \mathbb{1}\{T(X) \in \mathcal{R}_\alpha\}.$$

On appelle parfois région de rejet l'événement $\{T \in \mathcal{R}_\alpha\}$ lui-même. La région de rejet est calibrée par le choix a priori du **niveau** α, permettant de maîtriser le l'erreur commise en rejetant H_0 à tort :

$$\alpha \geq \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\{T \in \mathcal{R}_\alpha\}).$$

5. Calculer la valeur observée t de la statistique de test T sur l'échantillon et prendre la décision.

Exemple test de la moyenne d'un échantillon gaussien à variance inconnue. On teste $\mu = \mu_0$ contre $\mu > \mu_0$; la statistique de test est $T = \sqrt{n}(\bar{X} - \mu_0)/\widehat{\sigma}$ qui suit une loi $\mathcal{T}(n-1)$ sous H_0. La région de rejet est $\mathcal{R}_\alpha = [q_{\mathcal{T}(n-1)}(1-\alpha); +\infty[$, telle que $\mathbb{P}_{H_0}(\{T \in \mathcal{R}_\alpha\}) = \alpha$.

A.5.1 Risques d'un test

Soit $\mathcal{R}$ la région de rejet d'un test de statistique T.

Définition 40. *On appelle*

— ***erreur ou risque de première espèce*** *la probabilité de rejeter H_0 alors qu'elle est vraie :*

$$\theta_0 \in \Theta_0 \mapsto \alpha(\theta_0) = \mathbb{P}_{\theta_0}(\{T \in \mathcal{R}\}).$$

— ***erreur ou risque de seconde espèce*** *la probabilité de conserver H_0 alors qu'elle est fausse :*

$$\theta_1 \in \Theta_1 \mapsto \beta(\theta_1) = \mathbb{P}_{\theta_1}(\{T \notin \mathcal{R}\}).$$

— ***puissance*** *la probabilité de refuser H_0 quand elle est fausse :*

$$\theta_1 \in \Theta_1 \mapsto \pi(\theta_1) = \mathbb{P}_{\theta_1}(\{T \in \mathcal{R}\}) = 1 - \beta(\theta_1).$$

A l'issue du test, les quatre situations suivantes sont possibles

	Choix H_0	Choix H_1
H_0 vraie	$1-\alpha$ bonne décision	α : erreur première espèce mauvaise décision
H_1 vraie	β : erreur deuxième espèce mauvaise décision	$\pi = 1-\beta$: puissance bonne décision

La décision du test, à partir de la valeur observée t de la statistique de test est :

— si $t \in \mathcal{R}_\alpha$, on rejette H_0 au niveau (ou au risque) α : les observations sont significativement différentes de celles attendues sous H_0. L'erreur commise en choisissant H_1 à tort est α, c'est ainsi que le test a été construit.

— si $t \notin \mathcal{R}_\alpha$, on conserve H_0 dans le test de niveau α : les données ne sont pas significatives pour choisir H_1. C'est un cas un peu moins confortable, car l'erreur de seconde espèce β commise en prenant cette décision (conserver H_0 à tort), n'est en général pas connue et peut être assez grande.

Dans le cadre de la théorie de Neyman-Pearson, l'erreur de première espèce α est calibrée $(5\%, 1\%, ...)$. L'objectif du test est le rejet de H_0, puisque le risque α de cette décision est contrôlé. On note ainsi la dissymétrie des hypothèses H_1 et H_0 : le contrôle est fait sur le risque de première espèce α, mais pas sur celui de seconde espèce β. Parmi tous les tests de niveau α, on cherchera bien sûr celui qui permet d'avoir la plus grande puissance, c'est à dire l'erreur de seconde espèce la plus faible.

Exemple pour tester la nocivité d'un nouveau médicament, il est préférable de choisir *dangereux* pour H_0, et *inoffensif* pour H_1, afin de ne pas choisir l'hypothèse d'inoffensivité sans en calibrer le risque. Peut-être qu'on décidera dangereux avec une erreur inconnue, et donc qu'on abandonnera un médicament potentiellement efficace et sans danger, mais c'est préférable à le déclarer inoffensif sans connaître l'erreur commise dans cette décision (à quel risque il pourrait être dangereux alors qu'il a étiqueté inoffensif).

Pou tester l'efficacité d'un médicament, il est préférable de choisir *inefficace* pour H_0, et *efficace* pour H_1, afin de limiter la mise sur le marché de médicaments inefficaces.

A.5.2 P-Value

La **p-value** définit le niveau critique de rejet de (H_0), ie le plus petit niveau pour lequel on rejette (H_0), étant donnée l'observation qu'on vient de faire de la statistique de test $T(\omega) = t_{obs}$. Si la région de rejet de niveau α est de la forme $\{T_n > q_\alpha\}$, la p-value est

$$\text{p-value} = \alpha(\omega) = \inf\{\alpha \in [0,1]; T_n(\omega) \in \mathcal{R}_\alpha\} = \inf\{\alpha \in [0,1]; T_n(\omega) > q_\alpha\}$$

Ce qui peut se réécrire, en remplaçant le seuil dans $\mathcal{R}$ par la valeur de la statistique observée sur les données $T_n(\omega)$

$$\text{P-value}(\omega) = \mathbb{P}_{\theta_0}\{T_n \in \mathcal{R}(T(\omega))\} = \mathbb{P}_{\theta_0}\{T_n > (T(\omega))\}.$$

Si (H_0) est composite, la p-value s'écrit

$$\text{P-value}(\omega) = \sup_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}\{T_n \in \mathcal{R}(T(\omega))\} = \sup_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}\{T_n > T(\omega)\}.$$

La p-valeur est une **variable aléatoire** de loi uniforme sous (H_0) dans le cas d'une statistique de test univariée. Dans un test de niveau α, H_0 est rejetée si $\alpha >$ p-value, conservée si $\alpha <$ p-value :

- si $0.05 >$ p-value > 0.01, le test est significatif,
- si $0.01 >$ p-value > 0.001, le test est très significatif,
- si $0.001 >$ p-value, le test est hautement significatif.

A.5.3 Propriétés d'un test

- Le test est sans biais si sa puissance est supérieure à son niveau ($\pi > \alpha$) : la probabilité de choisir H_1 à raison est supérieure à celle de choisir H_1 à tort.
- Le test est convergent si sa puissance tend vers 1 : on détecte toujours H_1.

Annexe B

Rappels de régression linéaire gaussienne

Ce chapitre rappelle les résultats essentiels de la régression linéaire

Bibliographie Azais et Bardet (2005), Pagès (2005), Cornillon et Matzner-Løber (2007), Cornillon et autres (2008), Rivoirard et Stoltz (2009)

Notations Soit X une matrice de dimension $n \times p$. Nous noterons x_i une ligne de X, X_j une colonne de X, c'est à dire $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ et $X_j = (x_{1j}, \ldots, x_{nj})'$ où la notation prime désigne la transposée d'une matrice.

B.1 Définition

On observe $(Y_1, \mathrm{x}_1), \ldots, (Y_n, \mathrm{x}_n)$. Le modèle linéaire gaussien est un modèle de régression défini matriciellement par

$$Y_{n\times 1} = X_{n\times p}\theta_{p\times 1} + \varepsilon_{n\times 1}, \quad \varepsilon_{n\times 1} \sim \mathcal{N}(0, \sigma^2 I_n), \tag{B.1}$$

avec $Y = (Y_1, \ldots, Y_n)'$ le vecteur colonne (aléatoire, de dimension n) des observations, X est la **matrice du plan d'expérience** et I_n la matrice identité de taille n. X est donnée, de taille $n \times p$; c'est la concaténation des n vecteurs lignes x_i ou celle des p variables colonnes $X_j = (x_{1j}, \ldots, x_{nj})'$.

Le modèle dépend des paramètres $\theta \in \mathbb{R}^p$ et $\sigma \in \mathbb{R}^{+*}$ et la famille de loi paramétrique est $\mathcal{P} = \mathcal{N}(X\theta, \sigma^2 I_n)$. Le paramètre de ce modèle est $\beta = (\theta, \sigma^2)$. De manière plus générale, le modèle peut être défini par

$$Y = m + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad m \in V \text{ sous espace vectoriel de } \mathbb{R}^p. \tag{B.2}$$

On ne fait plus alors référence à X, mais seulement à son image, $V = Im(X)$. Cette présentation renvoie souvent à la terminologie **modèle linéaire** gaussien.

B.2 Estimation

L'estimateur des moindres carrés et l'estimateur du maximum de vraisemblance de θ ont la même expression.

B.2.1 Estimateurs

Si X est régulière (injective)

$$\widehat{\theta} = (X'X)^{-1}X'Y, \;\; \mathbb{E}(\widehat{\theta}) = \theta, \;\; \text{var}(\widehat{\theta}) = \sigma^2(X'X)^{-1}, \;\; \widehat{\mathbb{E}(Y)} = X\widehat{\theta} = \widehat{Y}$$

$$\widehat{\sigma}^2 = SCR(\widehat{\theta})/(n-p) \text{ où } SCR(\theta) = (Y - X\theta)'(Y - X\theta) = \sum_i (Y_i - \mathrm{x}_i\theta)^2$$

Si X n'est pas injective, on résout sous une **condition d'identifiabilité** $C\theta = 0$ permettant d'obtenir une unique solution : $\widehat{\theta}_C = (X'X + C'C)^{-1}X'Y$. Dans ce cas, l'interprétation des coefficients du paramètre dépend de la CI. La **dimension** du modèle est $p = dim(Im(X))$.

B.2.2 Loi des estimateurs

Comme ε est gaussien, $\widehat{\theta}_{MV} = \widehat{\theta}_{MC}$ et $\widehat{\sigma}^2$ sont indépendants (Théorème de Cochran) et

$$\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1}), \;\; (n-p)\widehat{\sigma}^2/\sigma^2 \sim \chi^2(n-p)$$

Une forme linéaire standardisée du paramètre $L\theta$ où L est de dimension $1 \times p$ suit une loi de Student à $n - p$ degrés de liberté :

$$T = (L(X'X)^{-1}L')^{-1/2}\frac{L\widehat{\theta} - L\theta}{\widehat{\sigma}} \sim \mathcal{T}(n-p). \tag{B.3}$$

et la statistique de Wald renormalisée suit exactement une loi de Fisher

$$W/r = F_n = \frac{1}{r\widehat{\sigma}^2}(A(\widehat{\theta} - \theta))'[A(X'X)^{-1}A']^{-1}A(\widehat{\theta} - \theta) \sim \mathcal{F}(r, n-p). \tag{B.4}$$

où A matrice $r \times p$ de rang r.

Ces statistiques permettant de définir des tests de sous-modèles. La statistique de **Fisher** F peut aussi avoir une interprétation géométrique, cf B.7

B.2.3 Résidus

— **bruts** $\widehat{\varepsilon} = Y - \widehat{Y}$ sont centrés $\mathbb{E}(\widehat{\varepsilon}|X) = 0$, hétéroscédastiques $\text{var}(\widehat{\varepsilon}|X) = \sigma^2(I - H)$, décorrélés avec les valeurs estimées $\text{cov}(\widehat{Y}, \widehat{\varepsilon}|X) = 0$.

— **studentisés** par validation croisée

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \sim \mathcal{T}(n-1-p)$$

où $\widehat{\sigma}_{(i)}$ est calculé dans un échantillon privé de l'observation i, sont utilisés dans la validation du modèle.

B.2.4 Coefficient de détermination

C'est un indicateur global de la qualité de l'ajustement.

Définition 41. *Soit $\Omega = Im(X)$ un modèle contenant le régresseur $\mathbb{1}$. Le coefficient de détermination (multiple) est défini par*

$$R^2 = \frac{||\widehat{Y} - \overline{Y}\mathbb{1}||^2}{||Y - \overline{Y}\mathbb{1}||^2} = \frac{SCM(\Omega)}{SCR(\mathbb{1})} = \cos^2\alpha. \tag{B.5}$$

R^2 est un critère d'ajustement des données au modèle : $R^2 = SCM(\widehat{\theta}_{iid}, \widehat{\theta})/SCR(\widehat{\theta}_{iid})$.

B.3 Tests

B.3.1 Test de Student

C'est le test de nullité d'une forme linéaire du paramètre

$$(H_0) : L\theta = 0 \text{ contre } (H_1) : L\theta \neq 0$$

La statistique de test T définie en B.3 suit sous (H_0) une loi de Student à $n-p)$ degrés de liberté. D'où la région de rejet de niveau α

$$\mathcal{R}_\alpha = \{|T| > t_{n-p,1-\alpha/2}\}$$

où $t_{n-p,1-\alpha/2}$ est la quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-p)$ degrés de liberté.

Ce test est utile pour tester la nullité d'un paramètre de la régression $\theta_j = 0$ contre $\theta_j \neq 0$, ou l'égalité de deux paramètres $\theta_j - \theta_k = 0$ contre $\theta_j - \theta_k \neq 0$ par exemple.

B.3.2 Test de Fisher

Le test de Fisher est le test d'un sous modèle linéaire du modèle (B.2) : $Y = m+\varepsilon, m \in V$, où V est un sous espace vectoriel de R^n qu'on notera dorénavant $V = \Omega$, de dimension $dim(\Omega) < n$. Soit ω un sous espace vectoriel de Ω tel que $dim(\omega) = q < p$: on dit que ω est **emboîté** dans Ω. On souhaite tester

$$H_0 : m \in \omega \text{ contre } H_1 : m \in \Omega \backslash \omega.$$

Y se décompose de la façon suivante

$$\begin{aligned} Y &= (Y - H_\Omega Y) + (H_\Omega Y - H_\omega Y) + H_\omega Y \\ &= (Y - \widehat{Y}_\Omega) + (\widehat{Y}_\Omega - \widehat{Y}_\omega) + \widehat{Y}_\omega \end{aligned} \tag{B.6}$$

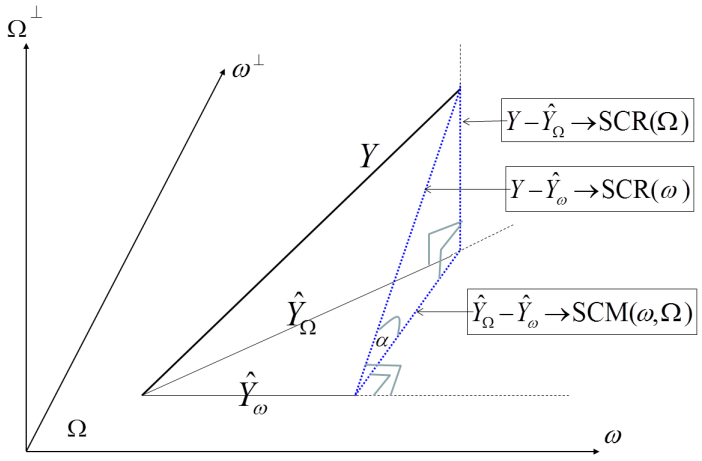


FIGURE B.1 – Illustration des projections sur Ω et ω

soit,

$$Y - \widehat{Y}_\omega = (Y - \widehat{Y}_\Omega) + (\widehat{Y}_\Omega - \widehat{Y}_\omega).$$

Or, $\widehat{Y}_\Omega$ et $\widehat{Y}_\omega$ appartiennent à Ω, et sont donc orthogonaux à $Y - \widehat{Y}_\Omega$. En appliquant le théorème de Pythagore dans le triangle en pointillé Figure B.1, il vient :

$$||Y - \widehat{Y}_\omega||^2 = ||\widehat{Y}_\Omega - \widehat{Y}_\omega||^2 + ||Y - \widehat{Y}_\Omega||^2$$
$$\mathrm{SCR}(\omega) = \mathrm{SCM}(\omega,\Omega) + \mathrm{SCR}(\Omega).$$

On reconnaît dans l'égalité précédente la somme des carrés résiduels $\mathrm{SCR}(\Omega)$ dans Ω, la somme des carrés résiduels $\mathrm{SCR}(\omega)$ dans ω, et un terme $\mathrm{SCM}(\omega, \Omega)$ appelé somme des carrés "Modèle", indiquant la perte d'ajustement résultant du choix de ω plutôt que Ω.

Quand les données sont générées sous ω, La statistique

$$F = \frac{(\mathrm{SCR}(\omega) - \mathrm{SCR}(\Omega))/(p-q)}{\mathrm{SCR}(\Omega)/(n-p)} = \frac{\mathrm{SCM}(\omega,\Omega)/(p-q)}{\mathrm{SCR}(\Omega)/(n-p)}. \tag{B.7}$$

suit une loi $\mathcal{F}(p-q, n-p)$. On en déduit la construction du test de Fisher :

Théorème 13. *Soit le modèle linéaire gaussien $Y = m + \varepsilon, m \in \Omega$, $\varepsilon \sim \mathcal{N}(0, I_n)$, de dimension $dim(\Omega) < n$. Soit ω un sous espace vectoriel de Ω tel que $dim(\omega) = q < p$, et soit F la statstique de Fisher définie par (B.7). Le test de Fisher*

$$H_0 : m \in \omega \text{ contre } H_1 : m \in \Omega \backslash \omega.$$

de région de rejet

$$\mathcal{R}_\alpha = \{F > f_{p-q,n-p,1-\alpha}\},$$

où $f_{p-q,n-p,1-\alpha}$ est le quantile de la loi de Fisher $\mathcal{F}(p-q, n-p)$, est de niveau α.

Les applications sont nombreuses :
- Test de significativité globale de la régression : ω est le modèle i.i.d. et Ω le modèle de régression.
- Test de significativité d'une composante : $\omega = \{\theta | \theta_j = 0\}$. Le test de Fisher revient alors à effectuer un test bilatéral de Student de cette composante, car si $T \sim \mathcal{T}(n-p)$, alors $T^2 \sim \mathcal{F}(1, n-p)$.
- Test d'hypothèses linéaires simultanées $\omega = \{\theta | A\theta = 0\}$

B.4 Intervalle de confiance

On peut utiliser les lois d'une combinaison linéaire du paramètre pour construire des intervalles de confiance.

B.4.1 Intervalle de confiance d'une espérance

Théorème 14. *Soit L une matrice de dimension $1 \times p$. Un **intervalle de confiance** de niveau $1-\alpha$ d'une forme linéaire de $L\theta$ est donné par*

$$\left[L\widehat{\theta} - t_{n-p}(1-\alpha/2)\widehat{\sigma}\sqrt{L'(X'X)^{-1}L}\ ;\ L\widehat{\theta} + t_{n-p}(1-\alpha/2)\widehat{\sigma}\sqrt{L'(X'X)^{-1}L}\right], \tag{B.8}$$

où $t_{n-p}(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p$ degrés de liberté.

Ces intervalles de confiance sont en particulier utilisés pour estimer une composante θ_j du paramètre, ou pour estimer l'espérance d'une observation sous une condition $L = \mathrm{x}^*$ donnée.

B.4.2 Intervalle de confiance de la prévision d'une nouvelle observation

A partir des n observations, nous avons estimé $\widehat{\theta}$. Pour prédire une nouvelle observation Y_{n+1} sous la condition d'expérience $\mathrm{x}_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$, il est naturel d'utiliser le modèle

$$Y_{n+1} = \mathrm{x}_{n+1}\theta + \varepsilon_{n+1},$$

en substituant à θ son estimation $\widehat{\theta}$. L'observation Y_{n+1} est alors prédite par

$$\widehat{Y}^p_{n+1} = \mathrm{x}_{n+1}\widehat{\theta}.$$

L'erreur de prévision est

$$\widehat{\varepsilon}^p_{n+1} = Y_{n+1} - \widehat{Y}^p_{n+1},$$

d'espérance et variance :

$$\begin{aligned}
\mathbb{E}(Y_{n+1} - \widehat{Y}^p_{n+1}) &= 0 \\
\mathrm{var}(Y_{n+1} - \widehat{Y}^p_{n+1}) &= \mathrm{var}(\mathrm{x}_{n+1}(\widehat{\theta} - \theta) + \varepsilon_{n+1}) \\
&= x_{n+1}\,\mathrm{var}(\widehat{\theta} - \theta)\mathrm{x}'_{n+1} + \sigma^2 \\
&= \sigma^2[\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1].
\end{aligned}$$

La loi de l'erreur de prévision est donc telle que

$$\frac{Y_{n+1} - \widehat{Y}^p_{n+1}}{\sigma\sqrt{\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1}} \sim \mathcal{N}(0,1).$$

D'où l'intervalle de prévision d'une valeur individuelle de niveau $1-\alpha$:

$$[\mathrm{x}_{n+1}\widehat{\theta} \pm t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1}]. \tag{B.9}$$

B.5 Variables explicatives qualitatives

B.5.1 Modèle ANOVA1

Une variable explicative qualitative à I niveaux : Soit Y_{ik} l'observation du k-ième individu du groupe $i = 1, \ldots, I$.

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}, \quad k = 1, \ldots, n_i; \quad n = \sum_{i=1}^{I} n_i; \; \varepsilon|\mathrm{x} \sim \mathcal{N}(0, \sigma^2 Id_n)$$

$\theta \in \mathbb{R}^{I+1}$, $p = dim(Im(X)) = p$. CI : $\alpha_1 = 0$, ou $\alpha_I = 0$, ou $\sum_i \alpha_i = 0$ par exemple

B.5.2 Modèle ANCOVA

Une variable explicative qualitative à I niveaux, une variable explicative quantitative t : Soit Y_{ik} l'observation du k-ième individu du groupe $i = 1, \ldots, I$.

$$Y_{ik} = \mu + \alpha_i + (\nu + \beta_i)t_{ik} + \varepsilon_{ik}, \quad k = 1, \ldots, n_I; \quad n = \sum_{i=1}^{I} n_i; \; \varepsilon|\mathrm{x} \sim \mathcal{N}(0, \sigma^2 Id_n)$$

$\theta \in \mathbb{R}^{2I+2}$, $p = dim(Im(X)) = 2I$. CI : $\alpha_1 = 0$ et $\beta_1 = 0$, ou $\alpha_1 = 0$ et $\nu = 0$ par ex.

B.6 Avec un logiciel

La fonction `lm` de R calcule en particulier la valeur de la statistique de Student de nullité de chaque coefficient. La fonction `anova` de R permet de faire des tests d'inclusion de modèle. Par exemple, elle permet de tester l'inclusion successive des variables suivant l'ordre dans lequel elles ont été utilisées pour définir le modèle. Les hypothèses des tests correspondants sont illustrés sur un exemple

`lm(y~x1+x2+x3)`

	(H_0)	(H_1)
x1	y~x2+x3	y~x1+x2+x3
x2	y~x1+x3	y~x1+x2+x3
x3	y~x1+x2	y~x1+x2+x3

`anova(lm(y~x1+x2+x3))`

	(H_0)	(H_1)
x1	iid	y~x1
x2	y~x1	y~x1+x2
x3	y~x1+x2	y~x1+x2+x3

Annexe C

Rappels d'optimisation sous contraintes

On donne f, une fonction *convexe* de $\mathbb{R}^p \to \mathbb{R}$, m fonctions *convexes* $g_i : \mathcal{X}_i \subset \mathbb{R}^p \to \mathbb{R}$ et r fonctions *affines* $h_j : \mathcal{X}_j \subset \mathbb{R}^p \to \mathbb{R}$. Un problème d'optimisation convexe sous contraintes cherche à résoudre le problème (P) suivant :

$$\begin{aligned} &\min_{x \in \mathcal{D}} f(x) \text{ sous les contraintes} \\ &\quad g_i(x) \leq 0 \; \forall i = 1, \ldots, m \\ &\quad h_j(x) = 0 \; \forall j = 1, \ldots, r \end{aligned}$$

où $\mathcal{D}$ est l'intersection des domaines des fonctions considérées. Les contraintes $g_i(x) \leq 0$ sont des contraintes d'inégalité, $h_j(x) = 0$ des contraintes d'égalité. Un point est dit *admissible* s'il appartient à $\mathcal{D}$ et vérifie toutes les contraintes. L'ensemble des points admissibles est convexe. On supposera que les fonctions f, g_i, h_j sont de classe $\mathcal{C}^1$.

Pour résoudre (P), on introduit autant de variables que de contraintes : m variables α_i et r variables β_j : ce sont les *multiplicateurs de Lagrange* ou *variables duales*. On introduit le Lagrangien de (P) qui prendre en compte ces nouvelles variables :

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{r} \beta_j h_j(x)$$

puis la *fonction duale* de Langrange de $\mathbb{R}^m \times \mathbb{R}^r \to \mathbb{R}$ définie par

$$q(\alpha, \beta) = \inf_{x \in \mathcal{X}} \mathcal{L}(x, \alpha, \beta)$$

Pour chaque couple (α, β), $q(\alpha, \beta)$ définit la plus *grande* valeur ℓ^* telle que $\ell^* \leq \mathcal{L}(x, \alpha, \beta)$. Cette fonction est concave, indépendamment de la convexité du problème (P). Le problème dual de Lagrange (D) associé au problème primal (P) le problème d'optimisation suivant

$$\begin{aligned} &\max_{\alpha \in \mathbb{R}^m, \beta \in \mathbb{R}^r} q(\alpha, \beta) \text{ sous les contraintes} \\ &\quad \alpha_i \geq 0 \; \forall i = 1, \ldots, m \end{aligned}$$

C.1 Dualité faible

Comme q est concave, c'est un problème d'optimisation convexe sous contraintes, que (P) soit convexe ou pas. Comme pour tout point x admissible $\mathcal{L}(x, \alpha, \beta) \leq f(x)$, on a

$$q(\alpha, \beta) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \alpha, \beta) \leq \min_{x \in \mathcal{D}} f(x) := p^*$$

Ainsi, si $d^* = q(\alpha^*, \beta^*)$ est solution de (D), on a $d^* \leq p^*$. Cette propriété est appelée *dualité faible*.

C.2 Dualité forte

La dualité est forte quand $p^* = d^*$.

Conditions de Slater (1950) Si (P) est convexe, ($f, g_1, \ldots, g_m$, convexes, $h_1, \ldots, h_r$ affines) et qu'il existe au moins un point admissible $\tilde{x}$ pour lesquels les contraintes d'inégalité $g_i(\tilde{x}) < 0$ soient strictement vérifiées ou qu'elles soient affines et vérifiées (condition de qualification), alors on a la dualité forte : $d^* = p^*$

Conditions de Karush-Kuhn-Tucker (1951) Si on a la propriété de dualité forte, alors x^* est un minimiseur du problème primal, si et seulement si il existe des multiplicateurs de Lagrange $\alpha^* = (\alpha_1^*, \ldots, \alpha_m^*)$ et $\beta^* = (\beta_1^*, \ldots, \beta_r^*)$ vérifiant

- admissibilité primale : pour tout $i = 1, \ldots, m$, $g_i(x^*) \leq 0$ et pour tout $j = 1, \ldots, r$, $h_j(x^*) = 0$
- admissibilité duale : pour tout $i = 1, \ldots, m$, $\alpha_i^* \geq 0$
- complémentarité des contraintes : pour tout $i = 1, \ldots, m$, $\alpha_i^* g(x^*) = 0$
- stationnarité : $\nabla_x \mathcal{L}(x^*, \alpha^*, \beta^*) = 0$

Bibliographie

J.-M. Azais et J.-M. Bardet. *Le modèle linéaire par l'exemple.* Dunod, Paris, 2005.

C.-A. Azencott. *Introduction au Machine Learning-2e éd.* Dunod, 2022.

N.-E. Breslow et N.-E. Day. Statistical methods in cancer research. 1 : The analysis of case-control studies. *IARC Publications*, 1, 1980. URL `http://www.iarc.fr/en/publications/pdfs-online/stat/sp32/`.

P. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation.* Masson, Paris, 1990.

P.-A. Cornillon et al. *R pour la statistique et la science des données.* Presses Universitaires de Rennes, 2018.

P.-A. Cornillon et autres. *Statistiques avec R.* Presses Universitaires de Rennes, Rennes, 2008.

P.-A. Cornillon et E. Matzner-Løber. *Régression Théorie et applications.* Springer, Paris, 2007.

D. Da Cunha et M. Duflo. Probabilités et statistiques. tome 2, collection mathématiques appliquées pour la maîtrise, 1983.

J.-F. Delmas. *Introduction au calcul des probabilités et à la statistique.* Les Presses de l'ENSTA, 2010.

A. Dobson. *An introduction to generalized linear models, 2nd edition.* Chapman and Hall, London, 2002.

B. Escofier et J. Pagès. *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation.* Dunod, 2008.

L. Fahrmeir et H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1) :342–368, 1985.

L. Fahrmeir, T. Kneib, S. Lang, et B. Marx. *Regression : Models, Methods and Applications.* Springer, Heidelberg, 2013.

A. R. Gallant et J. J. Goebel. Nonlinear regression with autocorrelated errors. *Journal of the American Statistical Association*, 71(356) :961–967, 1976.

X. Guyon. *Statistique et économétrie.* Ellipses, Paris, 2001.

T. Hastie, R. Tibshirani, et J. Friedman. *The elements of statistical learning : Data mining,Inference, and Prediction.* Springer, London, 2001.

L. Hubert et P. Arabie. Comparing partitions journal of classification 2 193–218. *Google Scholar*, pages 193–128, 1985.

S. Huet, A. Bouvier, M. Poursat, et E. Jolivet. *Statistical tools for nonlinear regression, 2nd edition.* Springer-Verlag, New-York, 2003.

F. Husson, S. Lê, et J. Pagès. *Analyse de données avec R.* Presses universitaires de Rennes, 2016.

G. James, D. Witten, T. Hastie, et R. Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

R. Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Statistics*, 40(2) :633–643, 1969.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā : The Indian Journal of Statistics, Series A*, pages 49–66, 2000.

L. Lebart, A. Morineau, et M. Piron. *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris, 1995.

M. Lejeune. *Statistique-La thèorie et ses applications.* Springer, 2010.

C. L. Mallows. Some comments on c p. *Technometrics*, 15(4) :661–675, 1973.

P. McCullagh et J. Nelder. *Generalized Linear Models, 2nd edition.* Chapman and hall, London, 1989.

R. H. Myers, D. C. Montgomery, G. G. Vining, et T. J. Robinson. *Generalized linear models : with applications in engineering and the sciences*, volume 791. John Wiley & Sons, 2012.

J. Nelder et R. Wedderburn. Generalized linear models. *J. R. Statist. Soc.*, 135 :370–384, 1972.

J. Pagès. *Statistique générales pour utilisateurs.* Presses Universitaires de Rennes, Rennes, 2005.

J. Pagès. *Analyse factorielle multiple avec R.* edp Sciences, Paris, 2011.

V. Rivoirard et G. Stoltz. *Statistique en action.* Vuibert, Paris, 2009.

P. J. Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 :53–65, 1987.

G. Saporta. *Probabilités, Analyse des données et Statistique, 2ème édition.* Editions TECHNIP, Paris, 2006.

R. Tibshirani, G. Walther, et T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63 (2) :411–423, 2001.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.