Analyse statistique des données d'expression

ALAIN BACCINI[1], PHILIPPE BESSE[1], SÉBASTIEN DÉJEAN[1],
PASCAL MARTIN[2], CHRISTÈLE ROBERT-GRANIÉ[3] & MAGALI SAN CRISTOBAL[4]

Version décembre 2008 — mises à jour et compléments :
`http ://math.univ-toulouse.fr/biostat/`

(1) Institut de Mathématiques de Toulouse – UMR CNRS 5219
Laboratoire de Statistique et Probabilités
Université de Toulouse
(2) Laboratoire de Pharmacologie-Toxicologie – (3) Station d'amélioration génétique des animaux
(4) Laboratoire de génétique cellulaire
Institut National de la Recherche Agronomique

Table des matières

Avant-propos

Motivations

Le développement des moyens informatiques de stockage (bases de données) et de calcul permet le traitement et l'analyse d'ensembles de données très volumineux. De plus, le perfectionnement des interfaces offre aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples des outils logiciels. Dans ce contexte, le biologiste dispose d'un corpus relativement sophistiqué de techniques statistiques utilisables sur les données d'expression des gènes produites par PCR, macro ou microarrays (biopuces). Les logiciels commerciaux ou non offrent des éventails plus ou moins larges d'accès à ces techniques dans une présentation plus ou moins explicite voire "boîte noire". Intentionnellement ce cours a fait le choix d'illustrer les techniques par un logiciel, le plus complet et le plus explicite possible : R. Même s'il ne semble pas le plus simple d'utilisation par rapport à certains produits commerciaux privilégiant une interface graphique "conviviale", son utilisation incite à l'indispensable compréhension des méthodes et de leurs limites. Il fait bien admettre qu'il ne suffit pas d'obtenir des résultats, il faut leur donner du sens. Rien ne nous semble en effet plus dangereux que des résultats ou des graphiques obtenus à l'aide de quelques clics de mulot dont ni les techniques, ni les options, ni leurs limites ne sont clairement explicitées ou contrôlées par l'utilisateur. Il est par ailleurs risqué de se laisser enfermer par les seules méthodes et options offertes par "un" logiciel. En pratique, le réagencement ou la réorganisation de quelques commandes R offrent une combinatoire très ouvertes de possibilités contrairement à un système clos de menus prédéfinis. Il offre par ailleurs, grâce à de nombreuses boîtes à outils librement accessibles et continuellement mises à jour, un ensemble exhaustif des techniques et de leurs options ainsi que des interfaces à des gestionnaires de bases de données ou des outils spécifiques à l'étude des biopuces (Bioconductor).

Objectifs généraux

Ce cours se place en aval d'une présentation des problèmes de planification, d'acquisition et de transformation (traitement d'image, normalisation) des données d'expression. D'autres cours et références existent sur ces points (voir page web), ils ne sont pas développés ici même s'ils sont tout aussi influents sur la pertinence et la qualité des résultats obtenus. Les méthodes sélectionnées sont celles paraissant les plus adaptées à la représentation graphique des données d'expression et à la construction de modèles explicatifs. Il s'agit de rechercher les représentations graphiques les plus éclairantes pour la compréhension de ce type de données, de leurs structures, puis de rechercher ou d'inférer des hypothèses spécifiques.

Ce cours se propose donc d'introduire deux grandes familles de méthodes sous une forme homogène, synthétique et relativement *intuitive* en privilégiant la mise en œuvre pratique aux développements théoriques. Sont ainsi traités des exemples simples, académiques, et d'autres plus

complexes mais provenant d'expériences réelles de mesures d'expressions.

i. Techniques statistiques *exploratoires* dites *multidimensionnelles* recouvrant d'une part les méthodes *factorielles* et d'autre part les méthodes de *classification* ou apprentissage non-supervisé.

ii. Méthodes statistiques dites *inférentielles* et de *modélisation* : tests, tests multiples et le *modèle linéaire* sous différentes formes (régression, analyse de variance, modèle mixte).

D'autres techniques plus récentes avec un objectif de discrimination et issues de la théorie de l'apprentissage (agrégation de modèles, support vector machine...) ont volontairement été laissées de côté. Elles nécessiteraient plus de temps pour être abordées.

Ce déroulement pédagogique linéaire ne doit pas faire perdre de vue que la réalité d'une analyse est plus complexe et nécessite différentes étapes en boucle afin, par exemple, de contrôler l'influence possible des choix parfois très subjectifs opérés dans les étapes de normalisation pour éventuellement les remettre en cause.

L'objectif principal est donc de faciliter la mise en œuvre, la compréhension et l'interprétation des résultats des techniques décrites pour en faciliter une *utilisation pertinente et réfléchie* à l'aide d'un logiciel (R) largement répandus dans la communauté scientifique. Ce cours ne peut se concevoir sans une mise en œuvre pratique au cours de séances de travaux dirigés sur machine.

Remerciements

Un grand merci à Agnès Bonnet, Heinrick Laurell, Pascal Martin, Gwenola Tosser-Klopp et Nathalie Viguerie pour les discussions scientifiques autour de leurs données respectives.

Chapitre 1

Introduction

1 Objectifs

Toute étude sophistiquée d'un corpus de données et leur modélisation sont précédées d'une étude *exploratoire* à l'aide d'outils, certes rudimentaires mais robustes, en privilégiant les représentations graphiques. C'est la seule façon de se familiariser avec des données et surtout de dépister les sources de problèmes :

- valeurs manquantes, erronées ou atypiques,
- modalités trop rares,
- distributions "anormales" (dissymétrie, multimodalité, épaisseur des queues),
- incohérences, liaisons non linéaires.
- ...

C'est ensuite la recherche de pré-traitements des données afin de les rendre conformes aux techniques de modélisation ou d'apprentissage qu'il sera nécessaire de mettre en œuvre afin d'atteindre les objectifs fixés :

- transformation : logarithme, puissance, centrage, réduction, rangs... des variables,
- codage en classe ou recodage de classes,
- imputations ou non des données manquantes,
- réduction de dimension, classification et premier choix de variables,
- classification ou typologie des observations.

Attention, le côté rudimentaire voire trivial de ces outils ne doit pas conduire à les négliger au profit d'une mise en œuvre immédiate de méthodes beaucoup plus sophistiquées, donc beaucoup plus sensibles aux problèmes cités ci-dessus. S'ils ne sont pas pris en compte, ils réapparaîtront alors comme autant d'*artefacts* susceptibles de dénaturer voire de fausser toute tentative de modélisation.

Plus précisément, ces méthodes descriptives ne supposent, *a priori*, aucun modèle sous-jacent, de type probabiliste. Ainsi, lorsqu'on considère un ensemble de variables quantitatives sur lesquelles on souhaite réaliser une Analyse en Composantes Principales, il n'est pas nécessaire de supposer que ces variables sont distribuées selon des lois normales. Néanmoins, l'absence de données atypiques, la symétrie des distributions sont des propriétés importantes des séries observées pour s'assurer de la qualité et de la validité des résultats.

La démarche traditionnelle consiste ensuite à enchaîner sur des techniques dites d'inférence statistique visant à tester les hypothèses retenues. Selon le nombre de variables explicatives ou à expliquer, leur nature qualitative ou quantitative, différents types de modèles et tests associés sont à considérer.

En théorie, contrairement à l'approche exploratoire, l'approche inférentielle nécessite une hypothèse probabiliste sur la distribution des observations ou des erreurs qui est, le plus souvent, l'hypothèse de *normalité* : la loi est supposée *gaussienne*. En pratique, cette hypothèse n'est de toute façon guère prouvable, les tests effectués sur les résidus estimés sont peu puissants (risque d'accepter à tord l'hypothèse mal maîtrisé). Cette hypothèse est néanmoins implicitement utilisée par les logiciels qui produisent systématiquement les résultats de tests. Plus rigoureusement, ces résultats sont justifiés par les propriétés des distributions asymptotiques des estimateurs, propriétés qui ne sont pas développées dans ce cours. En conséquence, du moment que les échantillons sont de taille "raisonnable", hypothèse on non de normalité, les distributions des estimateurs et donc les statistiques de test sont considérées comme valides.

En revanche, d'autres aspects des hypothèses, inhérentes aux méthodes développées et qui, en pratique, conditionnent fortement la qualité des estimations, doivent être évalués avec soin : *tests multiples, linéarité, colinéarité, homoscédasticité, points influents* ou atypiques (outliers). Les différents *diagnostics* ainsi que le problème du choix des variables explicatives, c'est-à-dire du *choix de modèle*, sont également fondammentaux.

2 Contenu

Ce cours se propose tout d'abord d'introduire brièvement les techniques permettant de résumer les caractéristiques (tendance centrale, dispersion, diagramme en boîte , histogramme, estimation non paramétrique) d'une variable statistique ou les relations entre variables de même type quantitatif (coefficient de corrélation, nuage de points), ou qualitatif (χ^2, Cramer, Tchuprow) ou de types différents (rapport de corrélation, diagrammes en boîtes parallèles). Les notions présentées sont illustrées sur des jeux de données d'expression.

Après cette approche uni puis bi-dimensionnelle, les techniques multidimensionnelles[1] sont décrites et illustrées. Elles diffèrent selon le type des variables considérées mais permettent toutes de réduire la dimension par un ensemble de *facteurs* afin de résumer un tableau $(n \times p)$ de grande dimension et révéler ses caractéristiques. L'analyse en composantes principales (ACP) pour les variables quantitatives ; l'analyse des correspondances simples ou multiples (AFCM) pour les variables qualitatives ainsi que l'analyse factorielle discriminante sont laissées de côté. L'analyse canonique compare deux tableaux quantitatifs correspondant aux observations de deux groupes de variables sur les mêmes individus. Les méthodes de classification (hiérarchiques ou par réallocation dynamique) déterminent une variable qualitative définissant une partition de l'ensemble des données. D'autres techniques sont plus spécifiques, le positionnement multidimensionnel ou ACP sur tableau de distances est adapté à des données particulières mais permet également de structurer un ensemble de variables trop important.

Les outils inférentiels sont ensuites introduits en insistant tout particulièrement sur le modèle linéaire et ses adaptations : régression linéaire simple pour introduire les concepts principaux : les tests et diagnostics, son extension à la régression linéaire multiple d'une variable à expliquée quantitative par plusieurs autres explicatives également quantitatives. Le cas de variables explicatives qualitatives est toujours un modèle de type linéaire : ANOVA ou analyse de variance. Enfin un dernier modèle est introduit prenant en compte des variables explicatives aléatoires. Il s'agit du modèle mixte.

Ce cours ne couvre pas, loin s'en faut, l'éventail des techniques statistiques utilisables pour détecter les gènes pertinents (différentiellement exprimés) en relation avec d'autres variables bio-

[1] Elles constituent un ensemble communément appelé en France "Analyse de Données".

logiques. Il manque la description des techniques de discrimination classiques (analyse discriminante décisionnelle, régression logistique) et surtout celles récentes émergeant à la frontière de la Statistique et de l'Informatique (machine learning) dans le cadre de la théorie de l'apprentissage (Vapnik 1999). Suppport Vector machine (SVM), bagging, boosting, random forest en sont les principales représentantes (cf. Hastie et col. 2001 pour une revue ou Besse 2003 pour une introduction et Baccini et col. (2005) pour une première utilisation.

3 Application aux données d'expression

Pour se montrer réaliste et dépasser les exemples académiques pour lesquels tout marche (trop) bien, les données utilisées pour illustrer ce cours sont effectivement des données d'expression recueillies dans le cadre du Génopôle de Toulouse. Trois exemples de données transcriptomiques sont utilisées. Le premier a été obtenu par PCR (*Polymerase Chain Reaction*), les deux suivants par puces à ADN sur membrane de nylon (*macroarray*). Dans chaque cas, le nombre de gènes étudiés est limité : entre 36 et 871, alors que l'homme, comme la souris, en comporte environ 30000. Les données nous ont été fournies après normalisation, de sorte que ce problème ne sera pas abordé ici. De même, les gènes considérés ayant été au préalable sélectionnés, le problème de la détection des gènes sur-exprimés (ou sous-exprimés) ne sera pas abordé directement, mais seulement à travers les analyses statistiques réalisées.

3.1 Jeux de données

Nutrition chez la souris

(*T. Pineau, P. Martin, Unité de Pharmacologie-Toxicologie, INRA Toulouse)*

Pour cette étude de nutrition, nous disposons de 40 souris réparties selon un plan à 2 facteurs (génotype à 2 niveaux et régime à 5 niveaux) avec 4 observations par cellule. Les mesures effectuées sont, d'une part, les mesures d'expression de 10 gènes relevées sur *macroarray* et, d'autre part, les proportions de 21 acides gras mesurées dans le foie.

La question du praticien concerne l'existence de *corrélations entre certains gènes ou groupes de gènes et certains acides gras hépatiques*.

Cet exemple est plus particulièrement développé au cours des travaux pratiques et joue un rôle de *fil rouge* tout au long de ce document afin d'en illustrer les principaux aspects ou point de vue. Les résultats décrits ici sont repris d'un article (Baccini et col. 2005) qui en font une analyse relativement exhaustive. C'est article est à paraître dans un numéro spécial du Journal de La Société Française de Statistique consacré aux données d'expression.

Obésité humaine

(*D. Langin, N. Viguerie, Unité de recherche sur les Obésités - INSERM U586, Toulouse)*

Pour cette étude, 50 patients répartis sur 8 sites européens ont été soumis à 2 régimes différents : plus ou moins riche en lipides et glucides avec même apport calorique. Nous disposons des expressions de 36 gènes pré-sélectionnés mesurées par PCR avant le régime et au bout de 10 semaines. Ces données sont complétées par les relevés (avant et après) de quelques paramètres cliniques directement reliés à l'amaigrissement (masse, masse grasse...). Ces données sont par ailleurs traitées par Viguerie et col. (2005).

Le problème consiste à *trouver un modèle tenant compte des différents facteurs afin d'en extraire les gènes différentiellement exprimés.*

Cancer pancréatique humain

(*H. Laurell, Unité de biologie et pathologie digestive - INSERM U531, Toulouse)*

L'étude a pour but d'améliorer le diagnostic du cancer pancréatique dont le pronostic est très mauvais. Nous disposons de l'expression de 871 gènes "spécifiques" du cancer mesurée via une membrane nylon sur 65 souches différentes : 49 cellules (26 pancréatiques, 18 coliques et 5 leucémiques) et 16 tissus (3 pancréas normaux et 13 tumeurs). Ces données sont traitées par Laurell et col. (2005).

L'objectif est de *représenter au mieux ces données afin d'en extraire des informations en termes de groupe de gènes et/ou de souches*.

3.2 Spécificités

Ce cours est délibérément orienté vers un type particulier de données qui se caractérise par, en général, un nombre très important de gènes dont l'expression est observée sur un nombre relativement restreint d'échantillons biologiques. De façon formelle, le problème se pose comme l'observation d'une variable, l'*expression* (ou quantité d'ARN messager produite) dans des situations expérimentales croisant deux facteurs : le gène et le type d'échantillon biologique (tissus sain ou pathologique, culture cellulaire ...). Le premier facteur peut présenter quelques centaines voire dizaines de milliers de niveaux tandis que le second, pour des raisons évidentes de coûts, ne présente en général que quelques dizaines de niveaux.

Nous nous intéressons donc à l'analyse d'un tableau de données dont on cherche des représentations graphiques pertinentes quant à l'expression des gènes, puis une approche de type "analyse de variance" donnera un autre regard, par des tests, sur la significativité des expressions observées. La difficulté majeure rencontrée, voire même le défi au statisticien, est posé par le nombres de gènes (ou variables) considérés au regard du nombre d'échantillon biologiques. Ces valeurs sont en rupture brutale avec les habitudes prises avec des tableaux de taille raisonnable, qui font généralement jouer un rôle dissymétrique aux lignes ("individus") et colonnes ("variables") du tableau. Cette situation induit des difficultés importantes à tous les niveaux de l'analyse. Il faut noter également la présence possible d'autres variables biologiques observées sur les mêmes échantillons avec le souci de vouloir comparer ou relier expressions et valeur prises par ces variables.

Dans la plupart des méthodes considérées, de nombreux choix sont laissés à l'utilisateur qui doit les conduire en connaissance de cause ou "tâtonner" pour arriver à des représentation satisfaisante, des tests significatifs, compte tenu de ses *a priori* et surtout de ses conditions expérimentales. Ces choix doivent bien sûr être connectés à ceux relatifs aux problèmes de normalisation dus à la technique de marquage et à la présence de "gènes" témoins ou calibrés sur les biopuces.

3.3 Choix méthodologiques initiaux

Voici une tentative de présentation synthétique de ces choix. Cette liste n'est sans doute pas exhaustive, elle devra être complétée avec l'acquisition d'une meilleure expertise du traitement de ces données. Nous pouvons déjà insister sur l'indispensable dialogue entre biologiste et statisticien pour opérer ces choix en connaissance de cause tant sur les aspects techniques que sur leurs implications biologiques. Ces choix ne pourront bien sûr pas tous être discutés en détail dans le cadre restreint de ce cours et nous nous proposons d'en illustrer les principaux sur les jeux de données rencontrés.

Transformations

Les données traitées sont issues des procédures de normalisation afférentes aux techniques de marquage ou peuvent encore subir des transformations. Voici les plus courantes en pratique :

logarithme cette fonction corrige une distribution de variable trop dissymétrique (skewness) et réduit l'influence de grandes valeurs qui pourraient être atypiques. Ceci se justifie en considérant que dans certains systèmes naturels, des effets peuvent être modélisés par des facteurs multiplicatifs plutôt qu'additifs.

centrage les données se présentent sous la forme d'une matrice, il est habituel, par exemple lors d'une analyse en composantes principales, de *centrer* les colonnes. Chaque variable est translatée de la valeur de sa moyenne empirique qui devient donc nulle. L'information liée à la "moyenne" peut être utile en soi mais est rarement très informative : cela concerne l'expression moyenne d'un gène pour toutes les puces ou celle d'une puce pour tous les gènes. On verra que le rôle des lignes et colonnes ou la distinction entre variables et individus n'étant pas toujours explicite, il peut être intéressant de procéder à un double centrage à la fois en lignes et en colonnes du tableau des données.

réduction dans le même ordre d'idée, l'unité de mesure utilisée n'est pas toujours à prendre en compte surtout si elle change d'une variable à l'autre ou encore si les variances sont très hétérogènes. Pour éliminer l'effet des variances dépendant directement des choix d'unité de mesure, il est d'usage de *réduire*, c'est-à-dire de diviser par son écart-type chacune des variables qui deviennent ainsi des quantités sans unité. Attention, pour des données d'expression, cette transformation n'est pas toujours pertinente. En ramenant à un les variances de gènes, les effets de sur ou sous-expressions de certains d'entre-eux sont en effet éliminés.

marges unitaires une autre façon d'éliminer une unité de mesure consiste à diviser les lignes (ou les colonnes) d'un tableau par ses marges ou sommes des valeurs en lignes (ou en colonnes). C'est la pratique courante lorsque le tableau contient des effectifs : table de contingence et cela conduit à l'analyse des correspondances. Pour les raisons évoquées ci-dessus (sur et sous expressions), cette approche ne semble pas appropriée aux données d'expression.

rangs lorsque les données sont parsemées de valeurs atypiques sans qu'aucune transformation fonctionnelle (logarithme, puissance) ne puisse en atténuer les effets, une façon "brutale" ou "robuste" de s'en sortir consiste à remplacer une valeur par son rang dans la séquence ordonnée. Ceci est à rapprocher des coefficients de corrélation calculés sur les rangs (Spearman).

Distances et pondérations

Il peut être utile d'introduire des pondérations sur les lignes ou colonnes du tableau des données. Cette pratique permet de redresser un échantillon lors d'un sondage. Il peut s'agir, par exemple, d'équiliblre l'importance de groupes qui se trouveraient sous représentés à cause "d'incidents techniques" ou d'affecter des poids nuls à des lignes ou colonnes dites alors supplémentaires. Ils n'interviennent pas dans les calculs mais restent représentés dans les graphiques. Par défaut les poids sont $1/n$ pour les lignes et 1 pour les variables ou colonnes. Chaque ligne (chaque colonne) est considérée comme un vecteur d'un espace vectoriel muni d'un produit scalaire induisant une norme euclidienne et donc une distance entre ces vecteurs. Par défaut, cette distance est celle classique dont le carré est la somme des carrés des écarts entre les coordonnées de deux vecteurs. Introduire des pondérations sur les lignes (les colonnes) conduit à pondérer le calcul de cette distance. La matrice de produit scalaire associée est alors une matrice diagonale faisant intervenir les pondérations (leur carré) sur la diagonale en lieu et place de la matrice identité.

D'autres matrices carrées symétriques définies positives sont également utilisables de façon plus générale. Citons l'inverse de la variance résiduelle ou intra en analyse discriminante, la matrice diagonale des inverses des fréquences marginales en analyse des correspondances qui définissent encore des distances euclidiennes de même que la matrice terme général $\sqrt{1 - cor(X^j, X^k)^2}$. D'autres matrices définissent des dissemblances entre variables : $1 - cor(X^j, X^k)$ faisant intervenir la corrélation linéaire (Pearson) ou celle calculée sur les rangs (Spearman).

Factorisation et projections

Beaucoup des méthodes proposées proposent la recherche de *facteurs* associés à la construction de nouvelles variables décorellées obtenues par combinaison linéaires des variables initiales et optimisant un critère : la variance pour l'analyse en composantes principales. La décomposition ainsi obtenue a-t-elle un sens pour les données considérées ? Combien de facteurs sont nécessaires pour "résumer l'information" et fournir des représentations graphiques pertinentes des nuages de points (individus et variables) dans cette nouvelle base ? Sur le plan mathématique, ces facteurs sont simplement les vecteurs propres associés aux plus grandes valeurs propres d'une matrice (variance, corrélation, produits scalaire...) carrée symétrique positive relativement à des métriques à définir dans les espaces vectoriels des individus et des variables.

Classification

Une approche classique dans toute discipline scientifique consiste à faire de la taxinomie c'est-à-dire à rechercher des classes homogènes des objets étudiés (gènes, échantillons biologiques) au sens d'un critère qui se définit par une matrice de distances ou dissemblances. Le choix de ce critère est évidemment prépondérant pour la signification et l'interprétation des résultats.

Test multiples

La pratique statistique usuelle vise à tester une hypothèse H_0 : le gène considéré n'a pas d'expression différentielle significative. Cela conduit à calculer une *statistique* de test dont la valeur est comparée aux quantiles de la loi de probabilités (Student ou Fisher) sous-jacente à cette statistique. Plus précisément, si la valeur calculée de la statistique de test est supérieure à un α-quantile (par exemple, $\alpha = 5\%$), on dit que l'hypothèse H_0 est rejeté avec un risque de première espèce de 5%. EN d'autres termes et pour cet exemple, nous avons moins de 5 chances sur 100 de nous tromper en affirmant que le gène en question est différentiellement exprimé.

Le problème qui se pose alors est celui dit des *faux positifs* dus à la très grande multiplicité des tests. En effet, en réalisant simultanément autant de tests que de gènes, par exemple 1000, rien que du fait du hasard, il est naturel de trouver qu'en moyenne, 5% (soit ici 50) des statistique de ces tests dépassent la valeur critique sans pour autant que les gènes se soient réellement exprimés d'un point de vue biologique. Ce sont les 5% d'erreurs associés au risque de première expèce induisant donc des faux positifs. Evidemment des correctifs sur les valeurs seuils sont apportés pour tenir compte de la multiplicité des tests. Bonferronni est la plus classique mais, très contraignante, elle semble peu adaptée à l'étude des données d'expression. D'autres approches sont proposées : FDR (false discovery rate, local FDR...) et une littérature très volumineuse est consacrée à ce problème. L'utilisateur est donc confronté au choix d'une stratégie de correction des valeurs critiques. Les autres corrections, bibliographie.

Chapitre 2

Description statistique élémentaire

1 Introduction

l'objectif des outils de Statistique descriptive élémentaire est de fournir, si possible graphiquement, des résumés synthétique de séries de valeurs, adaptés à leur type (qualitatives ou quantitatives), et observées sur une population ou un échantillon.

Dans le cas d'une seule variable, Les notions les plus classiques sont celles de médiane, quantile, moyenne, fréquence, variance, écart-type définies parallèlement à des représentations graphiques : diagramme en bâton, histogramme, diagramme-boîte, graphiques cumulatifs, diagrammes en colonnes, en barre ou en secteurs.

Dans le cas de deux variables, on s'intéresse à la corrélation, au rapport de corrélation ou encore à la statistique d'un test du χ^2 associé à une table de contingence. Ces notions sont associées à différents graphiques comme le nuage de points (scatterplot), les diagrammes-boîtes parallèles, les diagrammes de profils ou encore en mosaïque.

Les définitions de ces différentes notions se trouvent dans n'importe quel ouvrage élémentaire de Statistique[1], nous nous proposons simplement de rappeler dans ce chapitre certains outils moins classiques mais efficaces et présents dans la plupart des logiciels statistiques. Cela nous permettra également d'illustrer les premières étapes exploratoires à réaliser sur un jeu de données.

2 Decription d'une variable

2.1 Cas quantitatif

Une variable quantitative prend des valeurs entières ou réelles, elle est dite alors discrète ou continue. Cette propriété ayant des incidences sur la nature de sa distribution et donc sur les graphiques associés. Nous nous intéresserons surtout aux variables continues.

La distribution d'un variable statistique quantitative est résumée par différents indicateurs empiriques de *tendance centrale* (moyenne $\overline{x} = \sum_{i=1}^{n} w_i x_i$, médiane) ou de *dispersion* (écart-type σ, intervalle inter-quartiles). D'autres indicateurs s'intéressent à la dissymétrie (skeeness, associée au moment d'ordre 3) ou encore à l'aplatissement (kurtosis à partir du moment d'ordre 4)

Deux graphiques permettent de rendre compte précisément de la nature de la distribution. La statistique de Kolmogorov est la plus couramment utilisée pour tester l'adéquation à une loi (normale).

[1] Un support de cours accessible à la page `www-sv.cict.fr/lsp/Besse`.

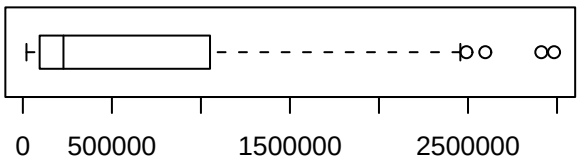


FIG. 2.1 – *Obésité : Diagramme-boîte illustrant la distribution dissymétrique de l'expression d'un gène.*

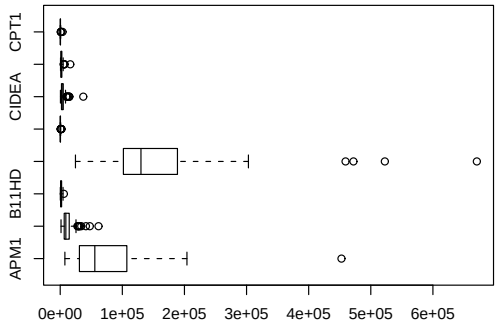


FIG. 2.2 – *Obésité : Diagrammes-boîtes parallèles représentant simultanément les distributions de plusieurs gènes.*

Diagramme-boîte (box-and-whiskers plot)

Il s'agit d'un graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane.

Histogramme

Dans le cas d'un échantillon, on cherche à approcher par une estimation empirique le graphe de la densité de la loi théorique associée à la population. L'*histogramme* en est un exemple. Une fois déterminée un découpage en classes de l'ensemble des valeurs et les fréquences f_ℓ d'occurrences de ces classes, un histogramme est la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_\ell = b_\ell - b_{\ell-1}$) et dont les hauteurs sont les quantités $\frac{f_\ell}{b_\ell - b_{\ell-1}}$, appelées *densités de fréquence*. L'aire du ℓ-ème rectangle vaut donc f_ℓ, fréquence de la classe correspondante.

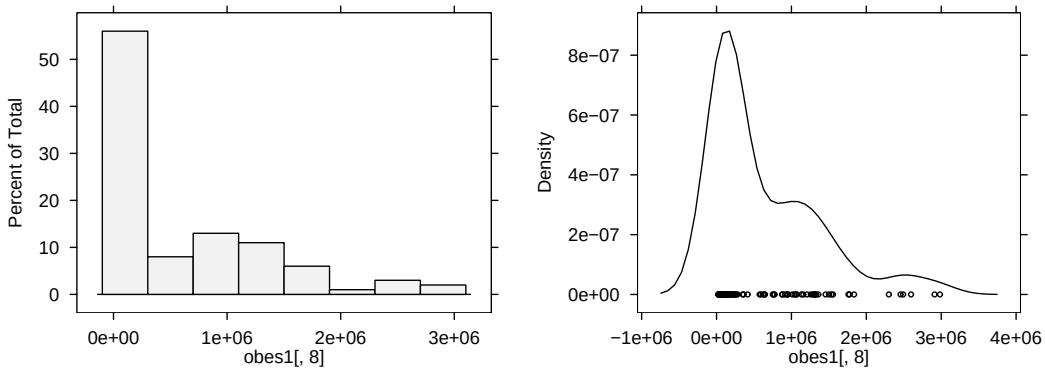


FIG. 2.3 – *Obésité : Histogramme et estimation fonctionnelle par la méthode du noyau de la distribution de l'expression d'un gène.*

Estimation fonctionnelle

La qualité de l'estimation d'une distribution par un histogramme dépend beaucoup du découpage en classe. Malheureusement, plutôt que de fournir des classes d'effectifs égaux et donc de mieux répartir l'imprécision, les logiciels utilisent des classes d'amplitudes égales et tracent donc des histogrammes parfois peu représentatifs. Ces 20 dernières années, à la suite du développement des moyens de calcul, sont apparues des méthodes d'estimation dites *fonctionnelles* ou *non-paramétriques* qui proposent d'estimer la distribution d'une variable ou la relation entre deux variables par une fonction construite point par point (noyaux) ou dans une base de fonctions *splines*. Ces estimations sont simples à calculer (pour l'ordinateur) mais nécessitent le choix d'un paramètre dit de *lissage*.

L'estimation de la densité par la méthode du noyau se met sous la forme générale :

$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda}\right)$$

où λ est le paramètre de lissage optimisé par une procédure automatique qui minimise une approximation de l'erreur quadratique moyenne intégrée (norme de l'espace L^2) ; K est une fonction symétrique, positive, concave, appelée *noyau* dont la forme précise importe peu. C'est souvent la fonction densité de la loi gaussienne réduite :

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$$

qui possède de bonnes propriétés de régularité. Le principe consiste simplement à associer à chaque observation un "élément de densité" de la forme du noyau K et à sommer tous ces éléments. Un histogramme est une version particulière d'estimation dans laquelle l"'élément de densité" est un "petit rectangle" dans la classe de l'observation.

2.2 Cas qualitatif

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées *modalités*. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac ou une classe d'âge), la variable est dite *ordinale*. Dans le cas

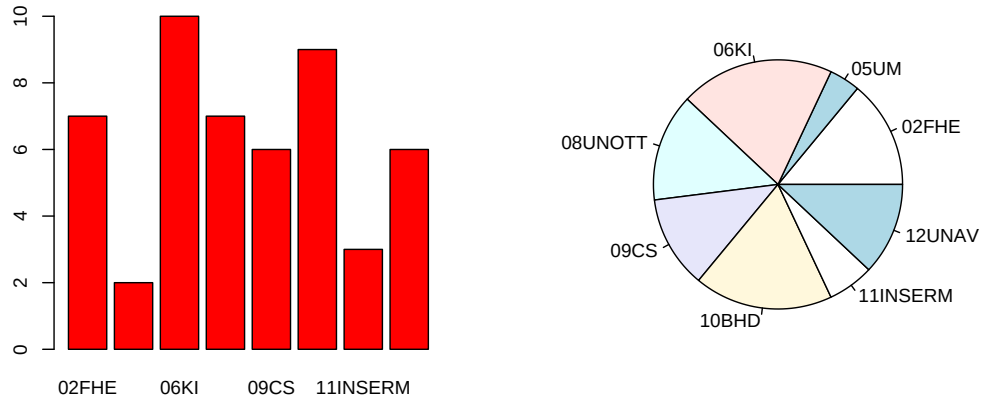


FIG. 2.4 – *Obésité : Diagramme en barres et camembert de la répartition des centres.*

contraire (par exemple, la profession dans une population de personnes actives ou la situation familiale) la variable est dite *nominale*.

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont les diagrammes en colonnes, en barre, en secteurs. Tous visent à représenter la répartition en effectif ou fréquences des individus dans les différentes classes ou modalités.

3 Liaison entre variables

Dans cette section, on s'intéresse à l'étude simultanée de deux variables X et Y. L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors *liaison*. Dans certains cas, cette liaison peut être considérée *a priori* comme *causale*, une variable X expliquant l'autre Y ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations et une liaison n'entraîne pas nécessairement une causalité. Sont ainsi introduites les notions de covariance, coefficient de corrélation linéaire, régression linéaire, rapport de corrélation, indice de concentration, khi-deux et autres indicateurs qui lui sont liés. De même, nous présentons les graphiques illustrant les liaisons entre variables : nuage de points (*scatter-plot*), diagrammes-boîtes parallèles, diagramme de profils, tableau de nuages (*scatter-plot matrix*).

3.1 Deux variables quantitatives

Nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives. Il consiste à considérer deux axes perpendiculaires, l'axe horizontal représentant la variable X et l'axe vertical la variable Y, puis à représenter chaque individu observé par les coordonnées des valeurs observées. L'ensemble de ces points donne en général une idée assez bonne

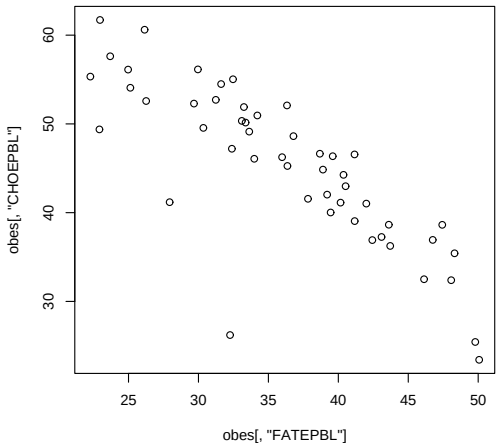


FIG. 2.5 – *Obésité : Nuage de points illustrant la liaison linéaire entre deux variables biologiques.*

de la variation conjointe des deux variables et est appelé *nuage*. On notera qu'on rencontre parfois la terminologie de *diagramme de dispersion*, traduction plus fidèle de l'anglais *scatter-plot*.

Le choix des échelles à retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables *homogènes* (représentant la même grandeur et exprimées dans la même unité) de celui des variables *hétérogènes*. Dans le premier cas, on choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites sur des axes orthonormés, soit de choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente (c'est en général cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques).

Indice de liaison

le coefficient de corrélation linéaire est un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. Il est défini à partir de la covariance qui généralise à deux variables la notion de variance :

$$\begin{aligned} \text{cov}(X,Y) &= \sum_{i=1}^{n} w_i [x_i - \overline{x}][y_i - \overline{y}] \\ &= [\sum_{i=1}^{n} w_i x_i y_i] - \overline{x}\,\overline{y}. \end{aligned}$$

La covariance est une forme bilinéaire symétrique qui peut prendre toute valeur réelle et dont la variance est la forme quadratique associée. Elle dépend des unités de mesure dans lesquelles sont exprimées les variables considérées ; en ce sens, ce n'est pas un indice de liaison "intrinsèque". C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (parfois appelé coefficient de Pearson ou de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types :

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Le coefficient de corrélation est égal à la covariance des variables centrées et réduites respectivement associées à X et Y : $\text{corr}(X,Y) = \text{cov}(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$. Par conséquent, $\text{corr}(X,Y)$ est indépendant des unités de mesure de X et de Y. Le coefficient de corrélation est *symétrique* et prend ses valeurs entre -1 et +1.

3.2 Une variable quantitative et une qualitative

Notations

Soit X la variable qualitative considérée, supposée à r modalités notées

$$x_1, \ldots, x_\ell, \ldots, x_r$$

et soit Y la variable quantitative de moyenne $\overline{y}$ et de variance σ_Y^2. Désignant par Ω l'échantillon considéré, chaque modalité x_ℓ de X définit une sous-population (un sous-ensemble) Ω_ℓ de Ω : c'est l'ensemble des individus, supposés pour simplifier de poids $w_i = 1/n$ et sur lesquels on a observé x_ℓ ; on obtient ainsi une *partition* de Ω en m classes dont nous noterons $n_1, \ldots, n_m$ les cardinaux (avec toujours $\sum_{\ell=1}^m n_\ell = n$, où $n = card(\Omega)$).

Considérant alors la restriction de Y à Ω_ℓ ($l = 1, \ldots, m$), on peut définir la moyenne et la variance partielles de Y sur cette sous-population ; nous les noterons respectivement $\overline{y_\ell}$ et σ_ℓ^2 :

$$\overline{y_\ell} = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} Y(\omega_i)\,;$$

$$\sigma_\ell^2 = \frac{1}{n_\ell} \sum_{\omega_i \in \Omega_\ell} [Y(\omega_i) - \overline{y_\ell}]^2.$$

Boîtes parallèles

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des boîtes parallèles ; il s'agit, sur un même graphique doté d'une échelle unique, de représenter pour Y un diagramme-boîte pour chacune des sous-populations définies par X. La comparaison de ces boîtes donne une idée assez claire de l'influence de X sur les valeurs de Y, c'est-à-dire de la liaison entre les deux variables.

Formules de décomposition

Ces formules indiquent comment se décomposent la moyenne et la variance de Y sur la partition définie par X (c'est-à-dire comment s'écrivent ces caractéristiques en fonction de leurs valeurs partielles) ; elles sont nécessaires pour définir un indice de liaison entre les deux variables.

$$\begin{aligned} \overline{y} &= \frac{1}{n} \sum_{\ell=1}^r n_\ell \overline{y_\ell}\,; \\ \sigma_Y^2 &= \frac{1}{n} \sum_{\ell=1}^r n_\ell (\overline{y_\ell} - \overline{y})^2 + \frac{1}{n} \sum_{\ell=1}^r n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2\,. \end{aligned}$$

Le premier terme de la décomposition de σ_Y^2, noté σ_E^2, est appelé *variance expliquée* (par la partition, c'est-à-dire par X) ou *variance inter* (between) ; le second terme, noté σ_R^2, est appelé *variance résiduelle* ou *variance intra* (within).

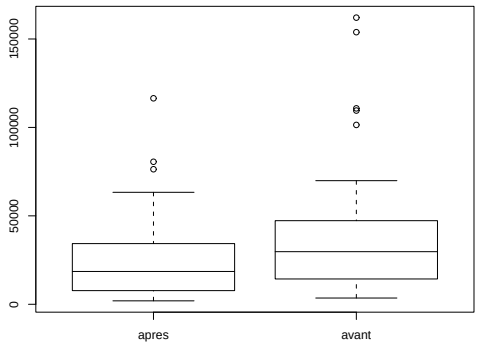


FIG. 2.6 – *Obésité : diagrammes-boîtes parallèles illustrant les différences de distribution des expressions d'un gène avant et après régime.*

Rapport de corrélation

Il s'agit d'un indice de liaison entre les deux variables X et Y qui est défini par :

$$s_{Y/X} = \sqrt{\frac{\sigma_E^2}{\sigma_Y^2}} ;$$

X et Y n'étant pas de même nature, $s_{Y/X}$ n'est pas symétrique et vérifie $0 \leq s_{Y/X} \leq 1$. Cet encadrement découle directement de la formule de décomposition de la variance. Les valeurs 0 et 1 ont une signification particulière intéressante.

3.3 Deux variables qualitatives

Notations

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur n individus. On suppose que la première, notée X, possède r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

Ces données sont présentées dans un tableau à double entrée, appelé *table de contingence*, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre $n_{\ell h}$ d'observations conjointes des modalités x_ℓ de X et y_h de Y ; les quantités $n_{\ell h}$ sont appelées les *effectifs conjoints*.

Une table de contingence se présente donc sous la forme suivante :

	y_1	$\cdots$	y_h	$\cdots$	y_c	sommes
x_1	n_{11}	$\cdots$	n_{1h}	$\cdots$	n_{1c}	n_{1+}
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell h}$	$\cdots$	$n_{\ell c}$	$n_{\ell +}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rh}	$\cdots$	n_{rc}	n_{r+}
sommes	n_{+1}	$\cdots$	n_{+h}	$\cdots$	n_{+c}	n

Les quantités $n_{\ell+}(\ell = 1, \ldots, r)$ et $n_{+h}(h = 1, \ldots, c)$ sont appelées les *effectifs marginaux* ; ils sont définis par $n_{\ell+} = \sum_{h=1}^{c} n_{\ell h}$ et $n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}$, et ils vérifient $\sum_{\ell=1}^{r} n_{\ell+} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on peut définir les notions de fréquences conjointes et de fréquences marginales.

Représentations graphiques

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'*adapter* les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas et que l'on appelle les *profils*.

Indices de liaison

Lorsque tous les profils-lignes sont égaux, ce qui est équivalent à ce que tous les profils-colonnes soient égaux et que

$$\forall (\ell, h) \in \{1, \ldots, r\} \times \{1, \ldots, c\} \ : n_{\ell h} = \frac{n_{\ell+} n_{+h}}{n} \, ,$$

on dit qu'il n'existe aucune forme de liaison entre les deux variables considérées X et Y. Par suite, la mesure de la liaison va se faire en évaluant l'écart entre la situation observée et l'état de non liaison défini ci-dessus.

Khi-deux

Il est courant en statistique de comparer une table de contingence observée, d'effectif conjoint générique $n_{\ell h}$, à une table de contingence donnée a priori (et appelée *standard*), d'effectif conjoint générique $s_{\ell h}$, en calculant la quantité

$$\sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - s_{\ell h})^2}{s_{\ell h}} \, .$$

De façon naturelle, pour mesurer la liaison sur une table de contingence, on utilise donc l'indice appelé khi-deux (chi-square) et défini comme suit :

$$\chi^2 = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \frac{n_{\ell+} n_{+h}}{n})^2}{\frac{n_{\ell+} n_{+h}}{n}} = n \left[\sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{n_{\ell h}^2}{n_{\ell+} n_{+h}} - 1 \right] .$$

Le coefficient χ^2 est toujours positif ou nul et il est d'autant plus grand que la liaison entre les deux variables considérées est forte. Malheureusement, il dépend aussi des dimensions r et c de la table étudiée, ainsi que de la taille n de l'échantillon observé ; en particulier, il n'est pas majoré. C'est la raison pour laquelle on a défini d'autres indices, liés au khi-deux, et dont l'objectif est de palier ces défauts.

Autres indicateurs

Nous en citerons trois.

- Le *phi-deux* : $\Phi^2 = \frac{\chi^2}{n}$. Il ne dépend plus de n, mais dépend encore de r et de c.

- Le coefficient T de Tschuprow :

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(r-1)(c-1)}}}.$$

  On peut vérifier : $0 \leq T \leq 1$.
- Le coefficient C de Cramer :

$$C = \sqrt{\frac{\Phi^2}{d-1}},$$

  avec : $d = \inf(r, c)$. On vérifie maintenant : $0 \leq T \leq C \leq 1$.

4 Vers le cas multidimensionnel

L'objectif des prochains chapitres de ce cours est d'exposer les techniques de la statistique descriptive multidimensionnelle. Or, sans connaître ces techniques, il se trouve qu'il est possible de débuter une exploration de données multidimensionnelles en adaptant simplement les méthodes déjà étudiées.

4.1 Matrices des covariances et des corrélations

Lorsqu'on a observé simultanément plusieurs variables quantitatives (p variables, $p \geq 3$) sur le même échantillon, il est possible de calculer d'une part les variances de toutes ces variables, d'autre part les $\frac{p(p-1)}{2}$ covariances des variables prises deux à deux. L'ensemble de ces quantités peut alors être disposé dans une matrice carrée ($p \times p$) et symétrique, comportant les variances sur la diagonale et les covariances à l'extérieur de la diagonale ; cette matrice, appelée matrice des variances-covariances (ou encore matrice des covariances) sera notée $\mathbf{S}$. Elle sera utilisée par la suite, mais n'a pas d'interprétation concrète. Notons qu'il est possible de vérifier que $\mathbf{S}$ est semi définie positive.

De la même manière, on peut construire la matrice symétrique $p \times p$, comportant des 1 sur toute la diagonale et, en dehors de la diagonale, les coefficients de corrélation linéaire entre les variables prises deux à deux. Cette matrice est appelée matrice des corrélations, elle est également semi définie positive, et nous la noterons $\mathbf{R}$. Elle est de lecture commode et indique quelle est la structure de corrélation des variables étudiées.

4.2 Tableaux de nuages

Notons $X^1, \ldots, X^p$ les p variables quantitatives considérées ; on appelle tableau de nuages le graphique obtenu en juxtaposant, dans une sorte de matrice carrée $p \times p$, p^2 sous-graphiques ; chacun des sous-graphiques diagonaux est relatif à l'une des p variables, et il peut s'agir, par exemple, d'un histogramme ; le sous-graphique figurant dans le bloc d'indice $(j, j'), j \neq j'$, est le nuage de points réalisé avec la variable X^j en abscisses et la variable $X^{j'}$ en ordonnées. Dans certains logiciels anglo-saxons, ces graphiques sont appelés *splom* (Scatter PLOt Matrix). Le tableau de nuages, avec la matrice des corrélations, fournit ainsi une vision globale des liaisons entre les variables étudiées.

5 Problèmes

Les quelques outils de ce chapitre permettent déjà de se faire une première idée d'un jeu de données mais surtout, en préalable à toute analyse, ils permettent de s'assurer de la fiabilité des

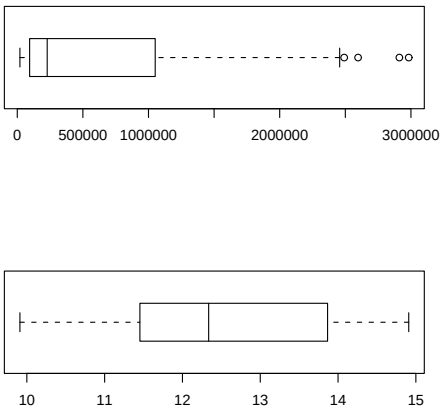


FIG. 2.7 – Obésité : La simple transformation ($\log(x)$), de l'expression d'un gène, résout bien les problèmes posés par l'allure "log-normale" de sa distribution avec son cortège de valeurs atypiques.

données, de repérer des valeurs extrêmes atypiques, éventuellement des erreurs de mesures ou de saisie, des incohérences de codage ou d'unité.

Les erreurs, lorsqu'elle sont décelées, conduisent naturellement et nécessairement à leur correction ou à l'élimination des données douteuses mais d'autres problèmes pouvant apparaître n'ont pas toujours de solutions évidentes.

- Le mitage de l'ensemble des données ou absence de certaines valeurs en fait partie. Faut-il supprimer les individus incriminés ou les variables ? Faut-il compléter, par une modélisation et prévision partielles, les valeurs manquantes ? Les solutions dépendent du taux de valeurs manquantes, de leur répartition (sont-elles aléatoires) et du niveau de tolérance des méthodes qui vont être utilisées.
- La présence de valeurs atypiques peut influencer sévèrement des estimations de méthodes peu robustes car basées sur le carré d'une distance. Ces valeurs sont-elles des erreurs ? Sinon faut-il les conserver en transformant les variables ou en adoptant des méthodes robustes basées sur des écarts absolus ?
- Même sans hypothèse explicite de normalité des distributions, il est préférable d'avoir affaire à des distributions relativement symétriques. Une transformation des variables par une fonction monotone (log, puissance) est hautement recommandée afin d'améliorer la symétrie de leur distribution ou encore pour linéariser (nuage de points) la nature d'une liaison.

6 Exemple : nutrition chez la souris

Comme annoncé en introduction, ce jeu de données est repris dans chaque chapitre. Dans cet exemple, la représentation des diagrammes en boîtes pour les souris, ordonnées selon le génotype et le régime suivi (Fig. 2.8) ne donne a priori aucune tendance spécifique sur le comportement de l'ensemble des gènes. Cette représentation atteste de la qualité de la production et de prétraitement des données. En effet, celles-ci ont été recueillies en utilisant une membrane par souris ; ainsi,

une quelconque anomalie sur un support, affectant l'ensemble des mesures relatives à une souris particulière, apparaîtrait nécessairement sur cette représentation. Notons seulement que quelques gènes atypiques, facilement repérables sur la figure 2.9 comme les plus surexprimés, se retrouvent dans les valeurs extrêmes pour chaque souris sur la figure 2.8.

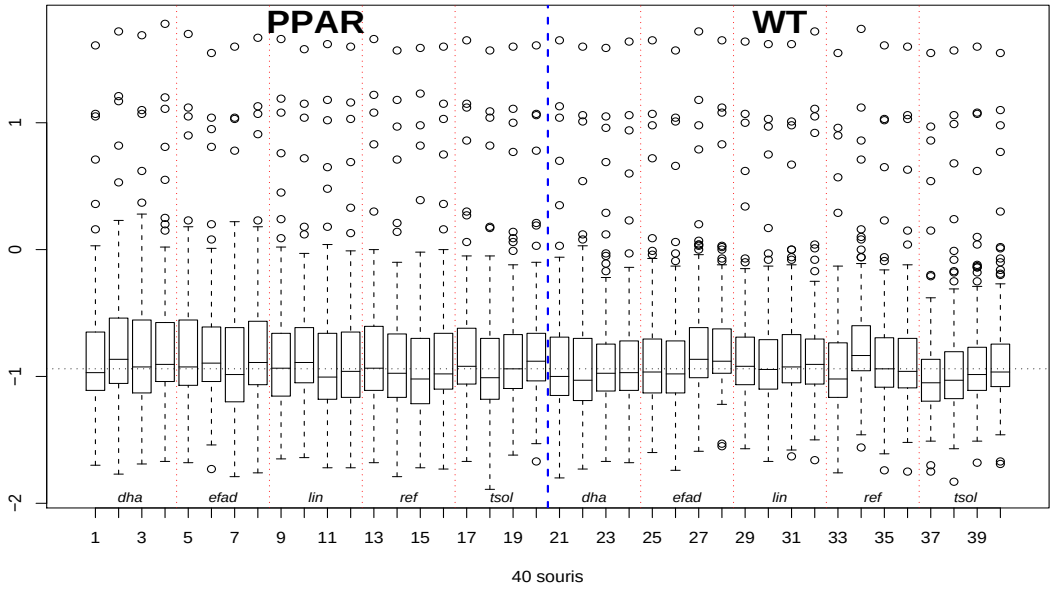


FIG. 2.8 – *Souris : iagrammes en boîtes pour les 40 souris. La ligne verticale et épaisse sépare les souris selon leur génotype. Les lignes verticales et fines séparent les souris selon le régime qu'elles ont suivi. La ligne horizontale représente la médiane de l'ensemble des valeurs.*

Les diagrammes en boîtes pour chaque gène (Fig. 2.9) révèlent des gènes dont l'expression est, sur l'ensemble des souris, nettement différentes des autres (par exemple, `16SR`, `apoA.I`, `apoE`). Les gènes des ARN ribosomiques comme le `16SR` (ARN 16s ribosomique mitochondrial), présentent, dans toutes les cellules de l'organisme, des niveaux d'expression plus élevés que tous les gènes codant des ARN messagers. Ces ARN servent en effet à la traduction des ARN messagers en protéines. Par ailleurs, on peut constater que les expressions de certains gènes varient beaucoup plus que d'autres sur l'ensemble des souris (par exemple, `FAS`, `S14` et `THIOL`). Pour ces derniers gènes, on peut supposer qu'une part de cette variabilité est due aux facteurs considérés, ce que nous essaierons de confirmer par la suite au moyen de techniques de modélisation.

L'intérêt de ces représentations réside davantage dans la vision synthétique qu'elles offrent que dans l'information biologique que l'on peut en extraire. Elles nous orientent également dans les premiers choix méthodologiques à établir avant de poursuivre l'analyse. En effet, les boîtes relatives à la distribution des gènes mettent clairement en évidence un certain nombre de gènes dont l'expression est systématiquement supérieure à celle des autres, quelles que soient les conditions expérimentales. De plus, la variabilité de ces expressions est, le plus souvent, très faible. Ce constat nous conduit à effectuer un centrage des gènes (en colonnes), afin d'éviter un effet taille lors de la mise en œuvre de techniques factorielles. En revanche, rien dans ces représentations ne nous pousse à centrer les échantillons (en lignes), ce qui, par ailleurs, ne se justifierait pas

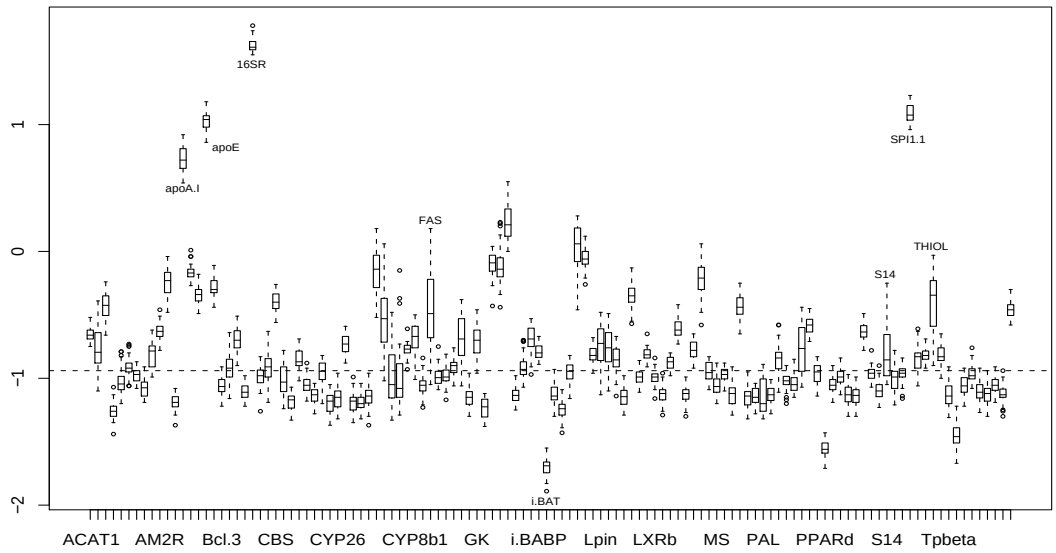


FIG. 2.9 – *Souris : diagrammes en boîtes pour les 120 gènes. Quelques gènes particuliers ont été étiquetés.*

sur le plan biologique. En effet, nous travaillons sur des données acquises via des puces dédiées sur lesquelles les gènes considérés ont été présélectionnés et sont donc, a priori, potentiellement différentiellement exprimés dans les conditions étudiées. Un centrage des échantillons serait susceptible de cacher des phénomènes biologiques. Ce raisonnement ne tiendrait pas pour une expérimentation pangénomique, où l'on pourrait supposer que globalement les gènes s'expriment de la même façon et que les surexprimés compensent les sous-exprimés.

Chapitre 3

Analyse en Composantes Principales

1 introduction

Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique adéquat, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (ainsi remplacées par les facteurs).

C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques, appelés aides à l'interprétation, qui sont là pour aider l'utilisateur à faire l'interprétation la plus juste et la plus objective possible.

L'analyse en Composantes Principales (ACP) est un grand classique de l"'analyse des données" en France pour l'étude exploratoire ou la compression d'un grand tableau $n \times p$ de données quantitatives. Le livre de Jolliffe (2002) en détaille tous les aspects et utilisations de façon exhaustive. Elle est introduite ici comme l'estimation des paramètres d'un modèle, afin de préciser la signification statistique des résultats obtenus. L'ACP est introduite d'abord intuitivement à travers l'étude de données fictives élémentaires. Elle est ensuite plus détaillée d'un point de vue méthodologique et illustrée par deux jeux de données Le premier est constitué des moyennes sur dix ans des températures moyennes mensuelles de 32 villes françaises. La matrice initiale $\mathbf{X}$ est donc (32×12). Les colonnes sont l'observation à différents instants d'une même variable. Le deuxième jeu concerne des expressions de gènes.

L'ACP joue dans ce cours un rôle central ; cette méthode sert de fondement théorique aux autres méthodes de statistique multidimensionnelle dites *factorielles* qui en apparaissent comme des cas particuliers. Cette méthode est donc étudiée en détail et abordée avec différents niveaux de lecture. La première section présente les grands principes de façon très élémentaire, voire intuitive, tandis que les suivantes explicitent les expressions matricielles des résultats.

2 Présentation élémentaire de l'ACP

2.1 Les données

Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Nous savons comment analyser séparément chacune de ces 4 variables, soit en faisant un *graphique*, soit en calculant des *résumés numériques*. Nous savons également qu'on peut regarder les *liaisons entre 2 variables* (par exemple mathématiques et français), soit en faisant un graphique du type nuage de points, soit en calculant leur *coefficient de corrélation linéaire*, voire en réalisant la *régression* de l'une sur l'autre.

Mais comment faire une étude simultanée des 4 variables, ne serait-ce qu'en réalisant un graphique ? La difficulté vient de ce que les individus (les élèves) ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension 4 (chacun étant caractérisé par les 4 notes qu'il a obtenues). L'objectif de l'Analyse en Composantes Principales est de revenir à un espace de dimension réduite (par exemple, ici, 2) en déformant le moins possible la réalité. Il s'agit donc d'obtenir *le résumé le plus pertinent* des données initiales.

2.2 Résultats préliminaires

Tout logiciel fournit la moyenne, l'écart-type, le minimum et le maximum de chaque variable. Il s'agit donc, pour l'instant, d'*études univariées*.

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Notons au passage la grande homogénéité des 4 variables considérées : même ordre de grandeur pour les moyennes, les écarts-types, les minima et les maxima.

Le tableau suivant est la *matrice des corrélations*. Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. C'est une succession d'*analyses bivariées*, constituant un premier pas vers l'*analyse multivariée*.

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Remarquons que toutes les corrélations linéaires sont positives (ce qui signifie que toutes les variables varient, en moyenne, dans le même sens), certaines étant très fortes (0.98 et 0.95), d'autres moyennes (0.65 et 0.51), d'autres enfin plutôt faibles (0.40 et 0.23).

2.3 Résultats généraux

Continuons l'analyse par celui de la **matrice des variances-covariances**, matrice de même nature que celle des corrélations, bien que moins "parlante" (nous verrons néanmoins plus loin comment elle est utilisée concrètement). La diagonale de cette matrice fournit les variances des 4 variables considérées (on notera qu'au niveau des calculs, il est plus commode de manipuler la variance que l'écart-type ; pour cette raison, dans de nombreuses méthodes statistiques, comme en A.C.P., on utilise la variance pour prendre en compte la dispersion d'une variable quantitative).

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Les *valeurs propres* données ci-dessous sont celles de la matrice des variances-covariances.

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	40.30	1.00	

Interprétation

Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (voilà les *facteurs*) dont la colonne VAL. PR. (valeur propre) fournit la variance (en fait, chaque valeur propre représente la variance du facteur correspondant). La colonne PCT. VAR, ou pourcentage de variance, correspond

au pourcentage de variance de chaque ligne par rapport au total. La colonne PCT. CUM. représente le cumul de ces pourcentages.

Additionnons maintenant les variances des 4 variables initiales (diagonale de la matrice des variances-covariances) : $11.39 + 8.94 + 12.06 + 7.91 = 40.30$. La dispersion totale des individus considérés, en dimension 4, est ainsi égale à 40.30.

Additionnons par ailleurs les 4 valeurs propres obtenues : $28.23+12.03+0.03+0.01 = 40.30$. Le nuage de points en dimension 4 est toujours le même et sa dispersion globale n'a pas changé. Il s'agit d'un simple changement de base dans un espace vectoriel. C'est la répartition de cette dispersion, selon les nouvelles variables que sont les facteurs, ou composantes principales, qui se trouve modifiée : les 2 premiers facteurs restituent à eux seuls la quasi-totalité de la dispersion du nuage, ce qui permet de négliger les 2 autres.

Par conséquent, les graphiques en dimension 2 présentés ci-dessous résument presque parfaitement la configuration réelle des données qui se trouvent en dimension 4 : l'objectif (résumé pertinent des donnée en petite dimension) est donc atteint.

2.4 Résultats sur les variables

Le résultat fondamental concernant les variables est le tableau des **corrélations variables-facteurs**. Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les facteurs. Ce sont ces corrélations qui vont permettre de donner un sens aux facteurs (de les interpréter).

Corrélations variables-facteurs

FACTEURS -->	F1	F2	F3	F4
MATH	0.81	-0.58	0.01	-0.02
PHYS	0.90	-0.43	-0.03	0.02
FRAN	0.75	0.66	-0.02	-0.01
ANGL	0.91	0.40	0.05	0.01

Les deux premières colonnes de ce tableau permettent, tout d'abord, de réaliser le *graphique des variables* (version SAS) donné ci-dessous.

Mais, ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques).

On notera que les deux dernières colonnes ne seront pas utilisées puisqu'on ne retient que deux dimensions pour interpréter l'analyse.

Interprétation

Ainsi, on voit que le premier facteur est corrélé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif. En ce qui concerne l'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques, surtout marqué par l'opposition entre le français et les mathématiques. Cette interprétation peut être précisée avec les graphiques et tableaux relatifs aux individus que nous présentons maintenant.

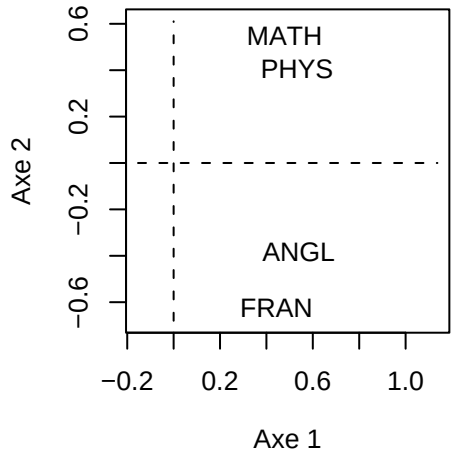


FIG. 3.1 – *Données fictives : Représentation des variables*

2.5 Résultats sur les individus

Le tableau ci-dessous contient tous les résultats importants sur les individus.

Coordonnées des individus ; contributions ; cosinus carrés

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	−8.61	−1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	−3.88	−0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	−3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	−2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	−3.03	−4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	−1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	−4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73

On notera que chaque individu représente 1 élément sur 9, d'où un poids (une pondération) de $1/9 = 0.11$, ce qui est fourni par la première colonne du tableau ci-dessus.

Les 2 colonnes suivantes fournissent les coordonnées des individus (les élèves) sur les deux premiers axes (les facteurs) et ont donc permis de réaliser le **graphique des individus**. Ce dernier permet de préciser la signification des axes, donc des facteurs.

Interprétation

On peut ainsi voir que l'axe 1 représente le résultat d'ensemble des élèves (si on prend leur score – ou coordonnée – sur l'axe 1, on obtient le même classement que si on prend leur moyenne générale). Par ailleurs, l'élève "le plus haut" sur le graphique, celui qui a la coordonnée la plus élevée sur l'axe 2, est Pierre dont les résultats sont les plus contrastés en faveur des disciplines littéraires (14 et 11.5 contre 7 et 5.5). C'est exactement le contraire pour André qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disci-

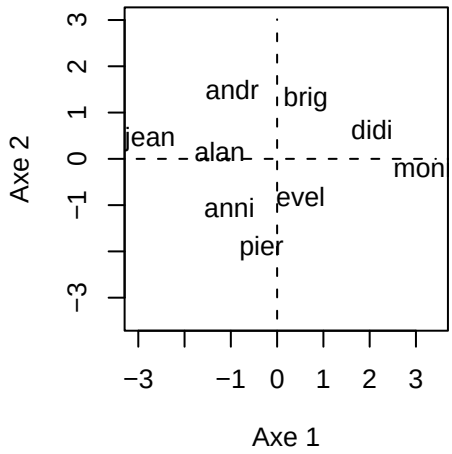


FIG. 3.2 – *Données fictives : Représentation des individus*

plines littéraires (7 et 5.5). On notera que Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines (mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1).

Les 3 colonnes suivantes du tableau fournissent des **contributions** des individus à diverses dispersions : CONT1 et CONT2 donnent les contributions des individus à la variance selon les axes 1 et 2 (rappelons que c'est la variance qui caractérise la dispersion) ; CONTG les contributions à la dispersion en dimension 4 (il s'agit de ce que l'on appelle l'**inertie** du nuage des élèves ; la notion d'inertie généralise celle de variance en dimension quelconque, la variance étant toujours relative à une seule variable). Ces contributions sont fournies en pourcentages (chaque colonne somme à 100) et permettent de repérer les individus les plus importants au niveau de chaque axe (ou du nuage en dimension 4). Elles servent en général à affiner l'interprétation des résultats de l'analyse.

Ainsi, par exemple, la variance de l'axe 1 vaut 28.23 (première valeur propre). On peut la retrouver en utilisant la formule de définition de la variance :

$$Var(C^1) = \frac{1}{9}\sum_{i=1}^{9}(c_i^1)^2$$

(il faut noter que, dans une A.C.P., les variables étant centrées, il en va de même pour les facteurs ; ainsi, la moyenne de C^1 est nulle et n'apparaît pas dans la formule de la variance). La coordonnée de Jean (le premier individu du fichier) sur l'axe 1 vaut $c_1^1 = -8.61$; sa contribution est donc :

$$\frac{\frac{1}{9}(-8.61)^2}{28.23} \times 100 = 29.19\ \%.$$

À lui seul, cet individu représente près de 30 % de la variance : il est prépondérant (au même titre que Monique) dans la définition de l'axe 1 ; cela provient du fait qu'il a le résultat le plus faible, Monique ayant, à l'opposé, le résultat le meilleur.

Enfin, les 2 dernières colonnes du tableau sont des cosinus carrés qui fournissent la (* qualité de la représentation *) de chaque individu sur chaque axe. Ces quantités s'additionnent axe par

axe, de sorte que, en dimension 2, Évelyne est représentée à 98 % (0.25 + 0.73), tandis que les 8 autres individus le sont à 100 %.

Lorsqu'on considère les données initiales, chaque individu (chaque élève) est représenté par un vecteur dans un espace de dimension 4 (les éléments – ou coordonnées – de ce vecteur sont les notes obtenues dans les 4 disciplines). Lorsqu'on résume les données en dimension 2, et donc qu'on les représente dans un plan, chaque individu est alors représenté par la projection du vecteur initial sur le plan en question. Le cosinus carré relativement aux deux premières dimensions (par exemple, pour Évelyne, 0.98 ou 98 %) est celui de l'angle formé par le vecteur initial et sa projection dans le plan. Plus le vecteur initial est proche du plan, plus l'angle en question est petit et plus le cosinus, et son carré, sont proches de 1 (ou de 100 %) : la représentation est alors très bonne. Au contraire, plus le vecteur initial est loin du plan, plus l'angle en question est grand (proche de 90 degrés) et plus le cosinus, et son carré, sont proches de 0 (ou de 0 %) : la représentation est alors très mauvaise. On utilise les carrés des cosinus, parce qu'ils s'additionnent suivant les différentes dimensions.

3 Représentation vectorielle de données quantitatives

3.1 Notations

Soit p *variables statistiques* réelles X^j $(j = 1, \ldots, p)$ observées sur n individus i $(i = 1, \ldots, n)$ affectés des poids w_i :

$$\begin{aligned} \forall i = 1, \ldots, n \ : w_i &> 0 \text{ et } \sum_{i=1}^{n} w_i = 1 \ ; \\ \forall i = 1, \ldots, n \ : x_i^j &= X^j(i), \text{ mesure de } X^j \text{ sur le } i^{ème} \text{ individu.} \end{aligned}$$

Ces mesures sont regroupées dans une matrice $\mathbf{X}$ d'ordre $(n \times p)$.

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	x_1^1	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_i^1	$\cdots$	x_i^j	$\cdots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_n^1	$\cdots$	x_n^j	$\cdots$	x_n^p

- À chaque individu i est associé le vecteur $\mathbf{x}_i$ contenant la i-ème ligne de $\mathbf{X}$ mise en colonne. C'est un élément d'un espace vectoriel noté E de dimension p ; nous choisissons $\mathbb{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice $\mathbf{M}$ lui conférant une structure d'espace euclidien : E est isomorphe à $(\mathbb{R}^p, \mathcal{E}, \mathbf{M})$; E est alors appelé *espace des individus*.
- À chaque variable X^j est associé le vecteur $\mathbf{x}^j$ contenant la j-ème colonne *centrée* (la moyenne de la colonne est retranchée à toute la colonne) de $\mathbf{X}$. C'est un élément d'un espace vectoriel noté F de dimension n ; nous choisissons $\mathbb{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice $\mathbf{D}$ diagonale des *poids* lui conférant une structure d'espace euclidien : F est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$; F est alors appelé *espace des variables*.

3.2 Interprétation statistique de la métrique des poids

L'utilisation de la métrique des poids dans l'espace des variables F donne un sens très particulier aux notions usuelles définies sur les espaces euclidiens. Ce paragraphe est la clé permettant de fournir les interprétations en termes statistiques des propriétés et résultats mathématiques.

Moyenne empirique de X^j :	$\overline{x^j}$	$=$	$\left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n.$
Barycentre des individus :	$\overline{\mathbf{x}}$	$=$	$\mathbf{X}'\mathbf{D}\mathbf{1}_n.$
Matrice des données centrées :	$\overline{\mathbf{X}}$	$=$	$\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'.$
Ecart-type de X^j :	σ_j	$=$	$(\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}.$
Covariance de X^j et X^k :	$\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k$	$=$	$\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}.$
Matrice des covariances :	$\mathbf{S}$	$=$	$\sum_{i=1}^n w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' = \overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}.$
Corrélation de X^j et X^k :	$\frac{\langle \mathbf{x}^j, \mathbf{x}^k \rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}} \|\mathbf{x}^k\|_{\mathbf{D}}}$	$=$	$\cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k).$

Attention : Par souci de simplicité des notations, on désigne toujours par $\mathbf{x}^j$ les colonnes de la matrice **centrée** $\overline{\mathbf{X}}$. On considère donc que des vecteurs "variables" sont toujours centrés.

Ainsi, lorsque les variables sont centrées et représentées par des vecteurs de F :
- la *longueur* d'un vecteur représente un *écart-type*,
- le *cosinus* d'un angle entre deux vecteurs représente une *corrélation*.

3.3 La méthode

Les objectifs poursuivis par une ACP sont :
- la représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace E_q de dimension q $(q < p)$,
- la représentation graphique des variables dans un sous-espace F_q en explicitant au "mieux" les liaisons initiales entre ces variables,
- la réduction de la dimension (compression), ou approximation de X par un tableau de rang q $(q < p)$.

Les derniers objectifs permettent d'utiliser l'ACP comme préalable à une autre technique préférant des variables orthogonales (régression linéaire) ou un nombre réduit d'entrées (réseaux neuronaux).

Des arguments de type géométrique dans la littérature francophone, ou bien de type statistique avec hypothèses de normalité dans la littérature anglo-saxonne, justifient la définition de l'ACP. Nous adoptons ici une optique intermédiaire en se référant à un modèle "allégé" car ne nécessitant pas d'hypothèse "forte" sur la distribution des observations (normalité). Plus précisément, l'ACP admet des définitions équivalentes selon que l'on s'attache à la représentation des individus, à celle des variables ou encore à leur représentation simultanée.

4 Modèle

Les notations sont celles du paragraphe précédent :
- $\mathbf{X}$ désigne le tableau des données issues de l'observation de p variables *quantitatives* X^j sur n individus i de *poids* w_i,
- E est l'espace des individus muni de la base canonique et de la métrique de matrice $\mathbf{M}$,
- F est l'espace des variables muni de la base canonique et de la métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

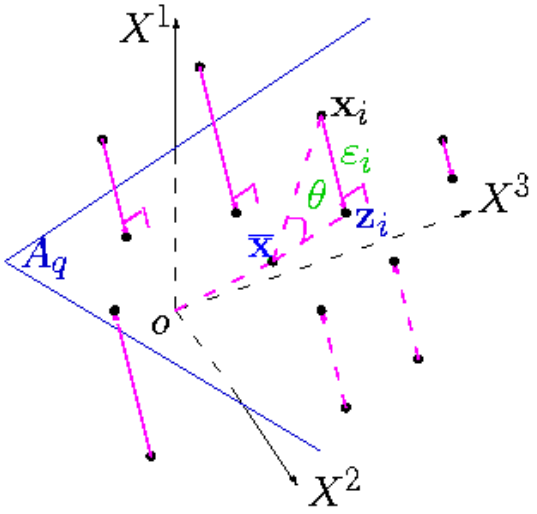


FIG. 3.3 – Principe de l'ACP dans l'espace des individus avec $p = 3$.

De façon générale, un modèle s'écrit :

$$\textbf{Observation} = \textbf{Modèle} + \textbf{Bruit}$$

assorti de différents types d'hypothèses et de contraintes sur le modèle et sur le bruit.

En ACP, la matrice des données est supposée être issue de l'observation de n vecteurs aléatoires indépendants $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, de même matrice de covariance $\sigma^2\mathbf{\Gamma}$, mais d'espérances différentes $\mathbf{z}_i$, toutes contenues dans un sous-espace affine de dimension q $(q < p)$ de E. Dans ce modèle, $E(\mathbf{x}_i) = \mathbf{z}_i$ est un paramètre spécifique attaché à chaque individu i et appelé *effet fixe*. Ceci s'écrit en résumé :

$$\begin{array}{l} \{\mathbf{x}_i\,; i = 1, \ldots, n\},\ n \text{ vecteurs aléatoires indépendants de } E, \\ \mathbf{x}_i = \mathbf{z}_i + \boldsymbol{\varepsilon}_i\,, i = 1, \ldots, n \text{ avec } \left\{ \begin{array}{l} E(\boldsymbol{\varepsilon}_i) = 0,\ \mathrm{var}(\boldsymbol{\varepsilon}_i) = \sigma^2\mathbf{\Gamma}, \\ \sigma > 0 \text{ inconnu, } \mathbf{\Gamma} \text{ régulière et connue,} \end{array} \right. \\ \exists A_q,\ \text{sous-espace affine de dimension } q \text{ de } E \text{ tel que } \forall i, \mathbf{z}_i \in A_q\ (q < p). \end{array} \tag{3.1}$$

Soit $\overline{\mathbf{z}} = \sum_{i=1}^{n} w_i \mathbf{z}_i$. Les hypothèses du modèle entraînent que $\overline{\mathbf{z}}$ appartient à A_q. Soit donc E_q le sous-espace vectoriel de E de dimension q tel que :

$$A_q = \overline{\mathbf{z}} + E_q.$$

Les paramètres à estimer sont alors E_q et $\mathbf{z}_i, i = 1, \ldots, n$, éventuellement σ ; $\mathbf{z}_i$ est la part systématique, ou *effet*, supposée de rang q ; éliminer le bruit revient donc à réduire la dimension.

4.1 Estimation

PROPOSITION 3.1. — *L'estimation des paramètres de (3.1) est fournie par l'ACP de* $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ *c'est-à-dire par la décomposition en valeurs singulières de* $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ *:*

$$\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda_k^{1/2} \mathbf{u}^k \mathbf{v}^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q'.$$

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant.
- Les $\mathbf{v}_k$, appelés *vecteurs principaux*, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{S}\mathbf{M}$ associés aux mêmes valeurs propres ; ils engendrent des s.e.v. de dimension 1 appelés axes principaux.

Les estimations sont donc données par :

$$\begin{aligned}
\widehat{\overline{\mathbf{z}}} &= \overline{\mathbf{x}}, \\
\widehat{\mathbf{Z}_q} &= \sum_{k=1}^{q} \lambda^{1/2}\mathbf{u}^k\mathbf{v}^{k'} = \mathbf{U}_q\mathbf{\Lambda}^{1/2}\mathbf{V}_q' = \overline{\mathbf{X}}\widehat{\mathbf{P}_q}', \\
\text{où } \widehat{\mathbf{P}_q} &= \mathbf{V}_q\mathbf{V}_q'\mathbf{M} \text{ est la matrice de projection} \\
&\quad \mathbf{M}\text{-orthogonale sur } \widehat{E_q}, \\
\widehat{E_q} &= \text{vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}, \\
\widehat{E_2} &\text{ est appelé plan principal,} \\
\widehat{\mathbf{z}_i} &= \widehat{\mathbf{P}_q}\mathbf{x}_i + \overline{\overline{x}}.
\end{aligned}$$

Remarques

i. Les solutions sont emboîtées pour $q = 1, \ldots, p$:

$$E_1 = \text{ vect}\{\mathbf{v}^1\} \subset E_2 = \text{ vect}\{\mathbf{v}^1, \mathbf{v}^2\} \subset E_3 = \text{ vect}\{\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3\} \subset \ldots$$

ii. Les espaces principaux sont uniques sauf, éventuellement, dans le cas de valeurs propres multiples.

iii. Si les variables ne sont pas homogènes (unités de mesure différentes, variances disparates), elles sont préalablement réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}}\mathbf{\Sigma}^{-1/2} \text{ où } \mathbf{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2), \text{ avec } \sigma_j^2 = \text{Var } (X^j) \text{ ;}$$

$\widetilde{\mathbf{S}}$ est alors la matrice $\mathbf{R} = \mathbf{\Sigma}^{-1/2}\mathbf{S}\mathbf{\Sigma}^{-1/2}$ des *corrélations*.

4.2 Définition équivalente

On considère p variable statistiques *centrées* $X^1, \ldots, X^p$. Une *combinaison linéaire* de coefficients $\mathbf{f}_j$ de ces variables,

$$\mathbf{c} = \sum_{j=1}^{p} f_j\mathbf{x}^j = \overline{\mathbf{X}}\mathbf{f},$$

définit une nouvelle variable centrée C qui, à tout individu i, associe la "mesure"

$$C(i) = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{f}.$$

PROPOSITION 3.2. — *Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'ACP de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ est aussi la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.*

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les *facteurs principaux*. Ils permettent de définir les combinaisons linéaires des X^j optimales au sens ci-dessus.

- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les *composantes principales*.
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les *variables principales* ;

$$\begin{aligned}\operatorname{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)'\mathbf{D}\overline{\mathbf{X}}\mathbf{f}^\ell = \mathbf{f}^{k\prime}\mathbf{S}\mathbf{f}^\ell \\ &= \mathbf{v}^{k\prime}\mathbf{M}\mathbf{S}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime}\mathbf{M}\mathbf{v}^\ell = \lambda_\ell \delta_k^\ell.\end{aligned}$$

- Les $\mathbf{f}^k$ sont les vecteurs propres $\mathbf{M}^{-1}$-orthonormés de la matrice $\mathbf{MS}$.
- La matrice

$$\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$$

  est la matrice des composantes principales.
- Les axes définis par les vecteurs $\mathbf{D}$-orthonormés u^k sont appelés *axes factoriels*.

5 Représentations graphiques

5.1 Les individus

Les graphiques obtenus permettent de représenter "au mieux" les distances euclidiennes inter-individus mesurées par la métrique $\mathbf{M}$.

Projection

Chaque individu i représenté par $\mathbf{x}_i$ est approché par sa projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}_i}^q$ sur le sous-espace $\widehat{E_q}$ engendré par les q premiers vecteurs principaux $\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$. En notant $\mathbf{e}_i$ un vecteur de la base canonique de E, la coordonnée de l'individu i sur $\mathbf{v}^k$ est donnée par :

$$\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{M}\mathbf{v}^k = \mathbf{e}_i'\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = c_i^k.$$

PROPOSITION 3.3. — *Les coordonnées de la projection $\mathbf{M}$-orthogonale de $\mathbf{x}_i - \overline{\mathbf{x}}$ sur $\widehat{E_q}$ sont les q premiers élément de la i-ème ligne de la matrice $\mathbf{C}$ des composantes principales.*

Mesures de "qualité"

La "qualité globale" des représentations est mesurée par la *part de dispersion expliquée* :

$$r_q = \frac{\operatorname{tr}\mathbf{S}\mathbf{M}\widehat{\mathbf{P}_q}}{\operatorname{tr}\mathbf{S}\mathbf{M}} = \frac{\sum_{k=1}^q \lambda_k}{\sum_{k=1}^p \lambda_k}.$$

Remarque. — La dispersion d'un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage $\mathcal{N}$ par rapport à son barycentre $\overline{\mathbf{x}}$ se mesure par l'*inertie*, généralisation de la variance :

$$I_g(\mathcal{N}) = \sum_{i=1}^n w_i \left\|\mathbf{x}_i - \overline{\mathbf{x}}\right\|_{\mathbf{M}}^2 = \left\|\overline{\mathbf{X}}\right\|_{\mathbf{M},\mathbf{D}}^2 = \operatorname{tr}\left(\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M}\right) = \operatorname{tr}\left(\mathbf{S}\mathbf{M}\right).$$

La qualité de la représentation de chaque x_i est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$[\cos\theta(\mathbf{x}_i - \overline{\mathbf{x}}, \widehat{\mathbf{z}_i}^q)]^2 = \frac{\left\|\widehat{\mathbf{P}_q}(\mathbf{x}_i - \overline{\mathbf{x}})\right\|_{\mathbf{M}}^2}{\left\|\mathbf{x}_i - \overline{\mathbf{x}}\right\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^q (c_i^k)^2}{\sum_{k=1}^p (c_i^k)^2}.$$

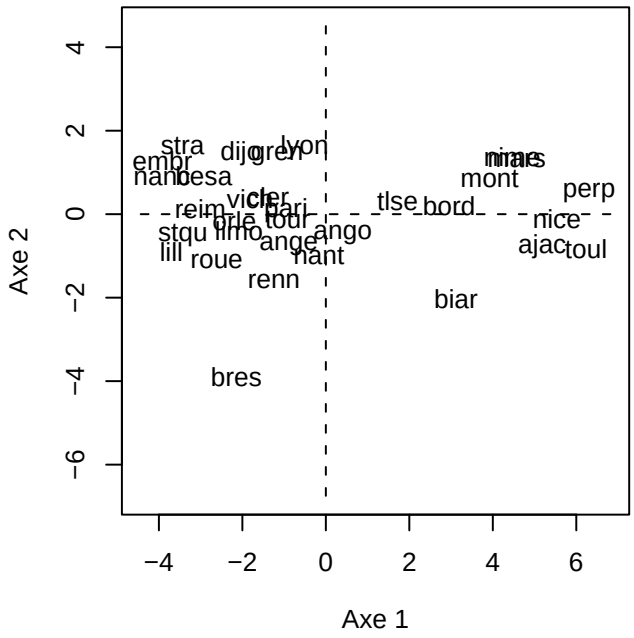


FIG. 3.4 – *Températures : premier plan des individus.*

Pour éviter de consulter un tableau qui risque d'être volumineux (n lignes), les étiquettes de chaque individu peuvent être affichées (macro SAS) sur les graphiques avec des caractères dont la *taille est fonction de la qualité*. Un individu très mal représenté est à la limite de la lisibilité.

Contributions

Les contributions de chaque individu à l'inertie de leur nuage

$$\gamma_i = \frac{w_i \, \|\mathbf{x}_i - \overline{\mathbf{x}}\|^2_{\mathbf{M}}}{\mathrm{tr}\mathbf{SM}} = \frac{w_i \sum_{k=1}^{p} (c_i^k)^2}{\sum_{k=1}^{p} \lambda_k},$$

ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k},$$

permettent de déceler les observations les plus *influentes* et, éventuellement, aberrantes. Ces points apparaissent visiblement lors du tracé des diagrammes-boîtes parallèles des composantes principales qui évitent ainsi une lecture fastidieuse de ce tableau des contributions. En effet, ils se singularisent aussi comme "outliers" hors de la boîte (au delà des moustaches) correspondant à une direction principale. Les individus correspondants, considérés comme *individus supplémentaires*, peuvent être éliminés lors d'une nouvelle analyse.

Individus supplémentaires

Il s'agit de représenter, par rapport aux axes principaux d'une analyse, des individus qui n'ont pas participé aux calculs de ces axes. Soit $\mathbf{s}$ un tel vecteur, il doit être centré, éventuellement réduit,

puis projeté sur le sous-espace de représentation. Les coordonnées sont fournies par :

$$\left\langle \mathbf{v}^k, \mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}})\right\rangle_{\mathbf{M}} = \mathbf{v}^{k\prime}\mathbf{M}\mathbf{V}_q\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}) = \mathbf{e}^{k\prime}\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}).$$

Les coordonnées d'un individu supplémentaire dans la base des vecteurs principaux sont donc :

$$\mathbf{V}_q'\mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}).$$

5.2 Les variables

Les graphiques obtenus permettent de représenter "au mieux" les corrélations entre les variables (cosinus des angles) et, si celles-ci ne sont pas réduites, leurs variances (longueurs).

Projection

Une variable X^j est représentée par la projection $\mathbf{D}$-orthogonale $\widehat{\mathbf{Q}_q\mathbf{x}^j}$ sur le sous-espace F_q engendré par les q premiers axes factoriels. La coordonnée de $\mathbf{x}^j$ sur $\mathbf{u}^k$ est :

$$\left\langle \mathbf{x}^j, \mathbf{u}^k\right\rangle_{\mathbf{D}} = \mathbf{x}^{j\prime}\mathbf{D}\mathbf{u}^k = \frac{1}{\sqrt{\lambda_k}}\mathbf{x}^{j\prime}\mathbf{D}\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = \frac{1}{\sqrt{\lambda_k}}\mathbf{e}^{j\prime}\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M}\mathbf{v}^k = \sqrt{\lambda_k}v_j^k.$$

PROPOSITION 3.4. — *Les coordonnées de la projection $\mathbf{D}$-orthogonale de $\mathbf{x}^j$ sur le sous-espace F_q sont les q premiers éléments de la j-ème ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.*

Mesure de "qualité"

La qualité de la représentation de chaque $\mathbf{x}^j$ est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$\left[\cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_q\mathbf{x}^j})\right]^2 = \frac{\left\|\widehat{\mathbf{Q}_q\mathbf{x}^j}\right\|_{\mathbf{D}}^2}{\|\mathbf{x}^j\|_{\mathbf{D}}^2} = \frac{\sum_{k=1}^q \lambda_k (v_k^j)^2}{\sum_{k=1}^p \lambda_k (v_k^j)^2}.$$

Corrélations variables×facteurs

Ces indicateurs aident à l'interprétation des axes factoriels en exprimant les corrélations entre variables principales et initiales.

$$\mathrm{cor}(X^j, C^k) = \cos\theta(\mathbf{x}^j, \mathbf{c}^k) = \cos\theta(\mathbf{x}^j, \mathbf{u}^k) = \frac{\left\langle \mathbf{x}^j, \mathbf{u}^k\right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j}v_j^k\,;$$

ce sont les éléments de la matrice $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Cercle des corrélations

Dans le cas de variables réduites $\widetilde{\mathbf{x}}^j = \sigma_j^{-1}\mathbf{x}^j$, $\left\|\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = 1$, les $\widetilde{\mathbf{x}}^j$ sont sur la sphère unité $\mathcal{S}_n$ de F. L'intersection $\mathcal{S}_n \cap F_2$ est un cercle centré sur l'origine et de rayon 1 appelé *cercle des corrélations*. Les projections de $\widetilde{\mathbf{x}}^j$ et $\mathbf{x}^j$ sont colinéaires, celle de $\widetilde{\mathbf{x}}^j$ étant à l'intérieur du cercle :

$$\left\|\widehat{\mathbf{Q}_2\widetilde{\mathbf{x}}^j}\right\|_{\mathbf{D}} = \cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_2 x_j}) \leq 1.$$

Ainsi, plus $\widehat{\mathbf{Q}_2\widetilde{\mathbf{x}}^j}$ est proche de ce cercle, meilleure est la qualité de sa représentation. Ce graphique est commode à interpréter à condition de se méfier des échelles, le cercle devenant une ellipse si elles ne sont pas égales. Comme pour les individus, la taille des caractères est aussi fonction de la qualité des représentations.

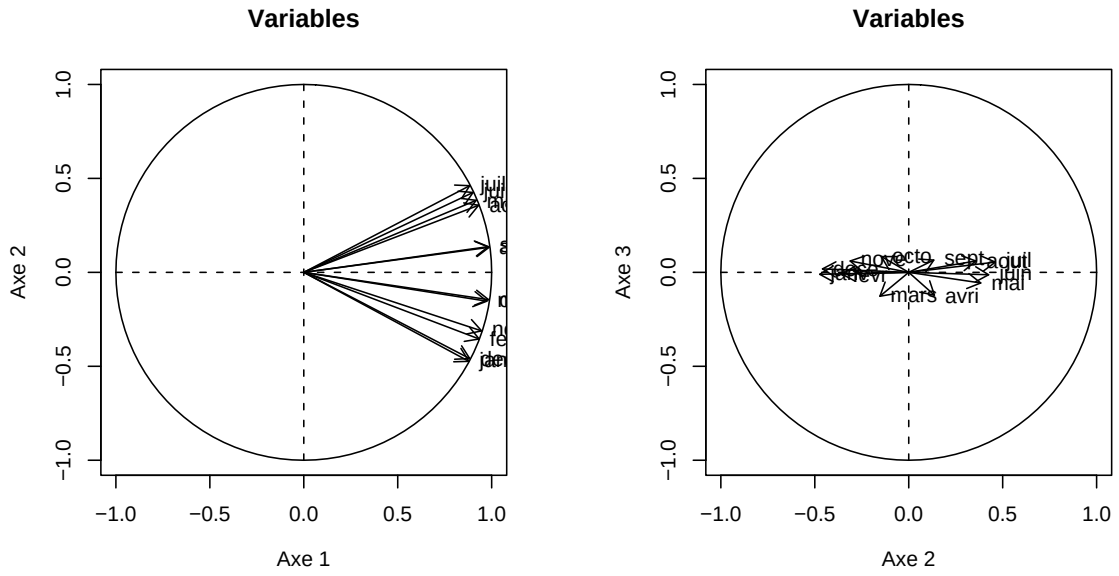


FIG. 3.5 – *Températures : Premier et deuxième plan des variables.*

5.3 Représentation simultanée ou "biplot"

À partir de la décomposition en valeurs singulières de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$, on remarque que chaque valeur

$$x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$$

s'exprime comme produit scalaire usuel des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j .$$

Pour $q = 2$, la quantité $\widehat{z}_i^j$ en est une approximation limitée aux deux premiers termes.

Cette remarque permet d'interpréter deux autres représentations graphiques en ACP projetant *simultanément* individus et variables.

i. la représentation *isométrique ligne* utilise les matrices $\mathbf{C}$ et $\mathbf{V}$; elle permet d'interpréter les distances entre individus ainsi que les produits scalaires entre un individu et une variable qui sont, dans le premier plan principal, des approximations des valeurs observées $X^j(\omega_i)$;

ii. la représentation *isométrique colonne* utilise les matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$; elle permet d'interpréter les angles entre vecteurs variables (corrélations) et les produits scalaires comme précédemment.

Remarques

i. Dans le cas fréquent où $\mathbf{M} = \mathbf{I}_p$ et où les variables sont réduites, le point représentant X^j, en superposition dans l'espace des individus se confond avec un pseudo individu supplémentaire qui prendrait la valeur 1 (écart-type) pour la variable j et 0 pour les autres.

ii. En pratique, ces différents types de représentations (simultanées ou non) ne diffèrent que par un changement d'échelle sur les axes ; elles sont très voisines et suscitent souvent les mêmes interprétations.

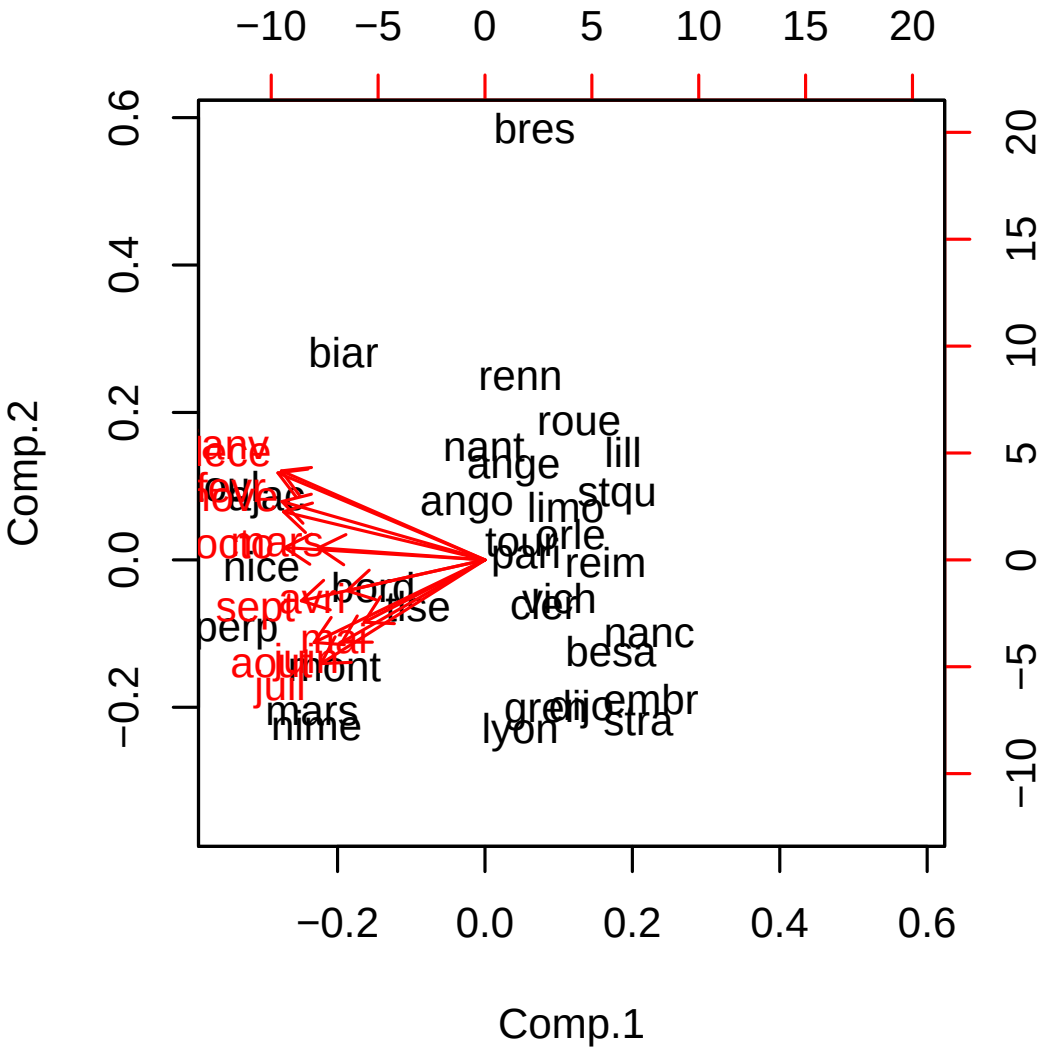


FIG. 3.6 – *Températures : Représentation simultanée du premier plan.*

6 Choix de dimension

La qualité des estimations auxquelles conduit l'ACP dépend, de façon évidente, du choix de q, c'est-à-dire du nombre de composantes retenues pour reconstituer les données, ou encore de la dimension du sous-espace de représentation.

De nombreux critères de choix pour q ont été proposés dans la littérature. Nous présentons ici ceux, les plus courants, basés sur une heuristique et un reposant sur une quantification de la stabilité du sous-espace de représentation. D'autres critères, non explicités, s'inspirent des pratiques statistiques décisionnelles ; sous l'hypothèse que l'erreur admet une distribution *gaussienne*, on peut exhiber les lois asymptotiques des valeurs propres et donc construire des tests de nullité ou d'égalité de ces dernières. Malheureusement, outre la nécessaire hypothèse de normalité, ceci conduit à une procédure de tests emboîtés dont le niveau global est incontrôlable. Leur utilisation reste donc heuristique.

6.1 Part d'inertie

La "qualité globale" des représentations est mesurée par la *part d'inertie expliquée* :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

La valeur de q est choisie de sorte que cette part d'inertie expliquée r_q soit supérieure à une valeur seuil fixée a priori par l'utilisateur. C'est souvent le seul critère employé.

6.2 Règle de Kaiser

On considère que, si tous les éléments de Y sont indépendants, les composantes principales sont toutes de variances égales (égales à 1 dans le cas de l'ACP réduite). On ne conserve alors que les valeurs propres supérieures à leur moyenne car seules jugées plus "informatives" que les variables initiales ; dans le cas d'une ACP réduite, ne sont donc retenues que celles plus grandes que 1. Ce critère, utilisé implicitement par SAS/ASSIST, a tendance à surestimer le nombre de composantes pertinentes.

6.3 Éboulis des valeurs propres

C'est le graphique (figure 6.3) présentant la décroissance des valeurs propres. Le principe consiste à rechercher, s'il existe, un "coude" (changement de signe dans la suite des différences d'ordre 2) dans le graphe et de ne conserver que les valeurs propres jusqu'à ce coude. Intuitivement, plus l'écart $(\lambda_q - \lambda_{q+1})$ est significativement grand, par exemple supérieur à $(\lambda_{q-1} - \lambda_q)$, et plus on peut être assuré de la stabilité de $\widehat{E_q}$.

6.4 Diagramme en boîte des variables principales

Un graphique (figure 6.4) présentant, en parallèle, les boîtes-à-moustaches des variables principales illustre bien leurs qualités : stabilité lorsqu'une grande boîte est associée à de petites moustaches, instabilité en présence d'une petite boîte, de grandes moustaches et de points isolés. Intuitivement, on conserve les premières "grandes boîtes". Les points isolés ou "outliers" désignent les points à forte contribution, ou potentiellement influents, dans une direction principale. Ils nécessitent une étude clinique : une autre analyse dans laquelle ils sont déclarés supplémentaires (poids nuls) afin d'évaluer leur impact sur l'orientation des axes.

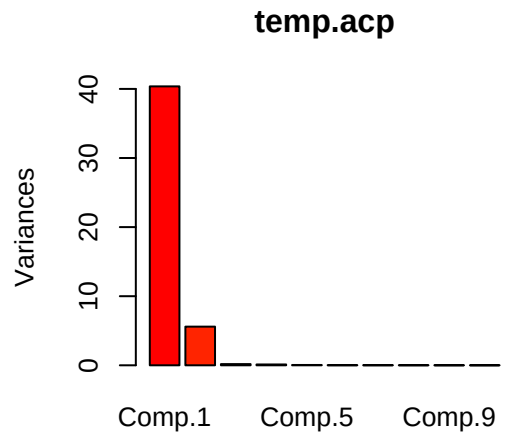


FIG. 3.7 – *Températures : éboulis des valeurs propres.*

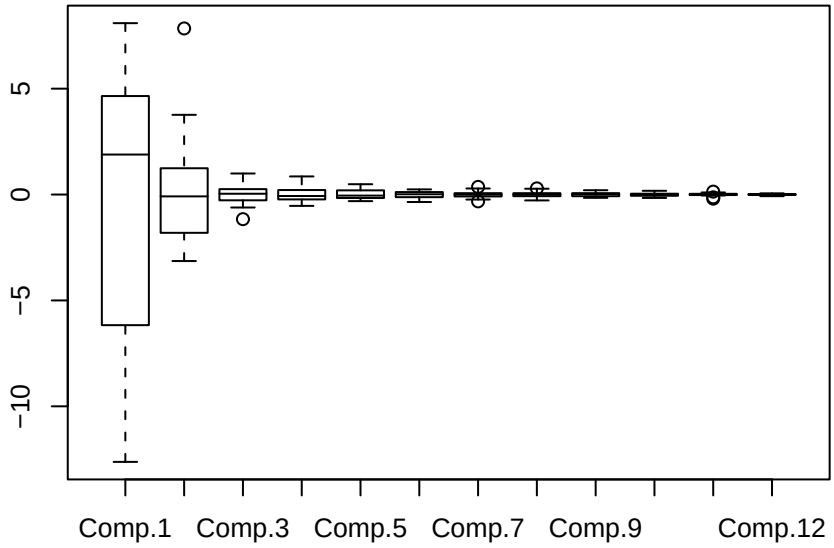


FIG. 3.8 – *Températures : composantes en boîtes.*

7 Interprétation

Les macros SAS décrites en exemple, de même que la plupart des logiciels, proposent, ou autorisent, l'édition des différents indicateurs (contributions, qualités, corrélations) et graphiques définis dans les paragraphes précédents.

- Les *contributions* permettent d'identifier les individus très influents pouvant déterminer à eux seuls l'orientation de certains axes ; ces points sont vérifiés, caractérisés, puis éventuellement considérés comme *supplémentaires* dans une autre analyse.
- Il faut choisir le nombre de composantes à retenir, c'est-à-dire la dimension des espaces de représentation.
- Les axes factoriels sont interprétés par rapport aux variables initiales bien représentées.
- Les graphiques des individus sont interprétés, en tenant compte des qualités de représentation, en termes de regroupement ou dispersions par rapport aux axes factoriels et projections des variables initiales.

Les quelques graphiques présentés suffisent, dans la plupart des cas, à l'interprétation d'une ACP classique et évitent la sortie volumineuse, lorsque n est grand, des tableaux usuels d'aide à l'interprétation. On échappe ainsi à une critique fréquente, et souvent justifiée, des anglo-saxons vis-à-vis de la pratique française de "l'analyse des données" qui, paradoxalement, cherche à "résumer au mieux l'information" mais produit plus de chiffres en sortie qu'il n'y en a en entrée !

Remarque. — L'ACP est une technique *linéaire* optimisant un critère *quadratique* ; elle ne tient donc pas compte d'éventuelles liaisons non linéaires et présente une forte sensibilité aux valeurs extrêmes.

8 Données d'expression

Les exemples illustratifs précédents ont l'avantage d'être simples et d'interprétation triviale. La réalité des données d'expression est tout autre et ce, en particulier, en raison du nombre de gènes en présence, c'est-à-dire en faite du nombre de variables d'expression observées p sur un nombre en général beaucoup plus réduit n d'individus. C'est le cas des données sur le cancer du pancréas pour lesquels 871 gènes sont observés pour seulement 65 tissus ou lignées cellulaires. L'incitation est évidemment forte à considérer les gènes comme des individus ce qui n'est pas sans conséquence.

8.1 Exploration élémentaire

Il n'est pas question de tracer 871 histogrammes. En revanche il est possible de représenter simultanément ces distributions par des diagrammes en boîtes même si celles-ci, les boîtes, deviennent très squelettiques compte tenu de leur nombre. La figure 8.1 affiche ces distributions qui se caractérisent par une certaine dérive des moyennes qui apparaissent pour le moins peu homogènes d'un gène à l'autre.

8.2 Analyse en composantes principales

Diverses options peuvent être mises en œuvre correspondant à plusieurs questions : quelles sont les variables (tissus, gènes) ? Quel centrage ? Faut-il réduire les variables ? Quelle représentation simple ou biplot faut-il privilégier ?

Dans R comme dans Splus, deux fonctions (`prcomp, princomp`), trois si l'on considère la librairie `multidim` de Carlier et Croquette, sont disponibles pour calculer des analyses en composantes principales. Deux extraient les valeurs propres et vecteurs propres de la matrice des

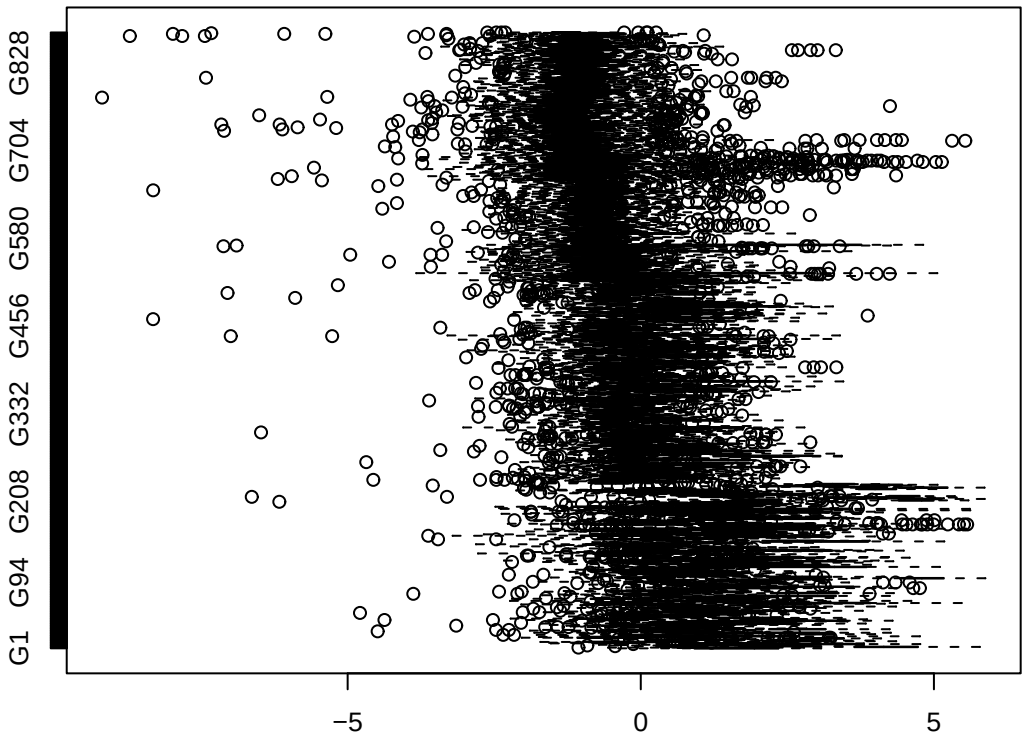


FIG. 3.9 – *Pancréas : diagrammes boîte affichant globalement les distributions des 871 gènes.*

covariances ou des corrélations tandis qu'une (`prcomp`) calcule directement la décomposition en valeurs singulières (SVD) de la matrice $\mathbf{X}$ des données centrées. L'ACP est donc centrée par défaut mais pas réduite par défaut sauf pour la fonction `acp`. Ces considérations techniques ont des conséquences importantes sur le soin à apporter dans le choix des options et sur le choix de la fonction : `prcomp` accepte un nombre p de colonnes supérieur au nombre n de lignes de la matrice ce qui n'est pas le cas de (`princomp`) . Tous les programmes sont détaillés dans les séances de travaux pratiques.

Quelques essais systématiques fournissent les graphiques de la figure 8.2 avec les différentes options de centrage et réduction. De toute façon, l'ACP centre par défaut les colonnes. Il apparaît important de centrer les gènes sinon (fig. 8.2 a) un effet taille, masque l'essentiel des données : certains gènes s'expriment de toute façon plus pour l'ensemble des tissus et emportent la définition du premier axe. Un centrage préalable des gènes ou lignes de la matrice (fig. 8.2 b) remédie à ce problème. Une réduction des "variables" tissus (fig. 8.2 c) n'apporte rien de mieux alors qu'une réduction des "variables" gènes (fig. 8.2 d) augmente sérieusement la confusion. C'est logique dans la mesure où cela revient à négliger la variance de l'expression des gènes qui affichent donc tous la même capacité d'expression. Cette réduction a un sens lorsqu'un sous-ensemble réduit de gènes sont considérés : ceux dont on sait *a priori* qu'ils sont différentiellement exprimés. La structure de corrélation fournit alors des informations sur les possibles co-régulations ou inhibitions.

Nous retiendrons plus particulièrement le graphique de la figure 8.2 croisant les tissus en ligne et les gènes en colonne qui sont donc centrés. Elle semble la plus explicite et informative compte tenu du problème posé. Ce graphe est strictement équivalent à celui de la figure 8.2 b qui nécessite un centrage préalable des lignes. Pour cette ACP, la décroissance des valeurs propres ou, plus précisément, les distributions des composantes principales (fig. 8.2) suggère de considérer 4 composantes nécessitant ainsi des graphiques et des interprétations plus élaborées.

À ce niveau de l'analyse on voit combien le nombre de gènes est susceptible de compliquer lecture et interprétation des graphiques. Diverses approches sont envisageables afin de se limiter aux plus "pertinents". Plusieurs répétitions permettent de tester d'un gène s'il est significativement différentiellement exprimé mais la répétition de ce test à chaque gène rend le procédé peut fiable. Une réflexion approfondie doit accompagner cette démarche pour le dépistage des "faux positifs" qui, sur le seul fait du hasard, rend 5% de tests significatifs en moyenne. C'est l'objet d'un autre cours.

Dans le cadre de l'analyse en composante principale, il est cohérent de rechercher quels sont les gènes contribuant le plus à la définition des axes retenus. Une adaptation de l'expression de la contribution d'un individu (section 5.1) :

$$\gamma_i = \frac{w_i \sum_{k=1}^{4} (c_i^k)^2}{\sum_{k=1}^{4} \lambda_k},$$

permet de rechercher les, par exemples, 5% des gènes contribuant le plus à la définition de l'espace propre à 4 dimensions jugé pertinent. La figure 8.2 ne représentant que ces gènes sélectionnés dans le premier plan factoriel devient beaucoup plus lisible. Naturellement, une ACP calculée sur ces seuls gènes les plus influents varie peu de celle calculée sur l'ensemble. La lecture de ce graphique permet, en particulier, de repérer les gènes contribuant le plus à une sorte de double discrimination :

- entre les échantillons biologiques issues des lignées cellulaires et les échantillons prélevés sur les patients,
- et, pour ces derniers échantillons, une distinction entre pancréas sains et pièces tumorales.

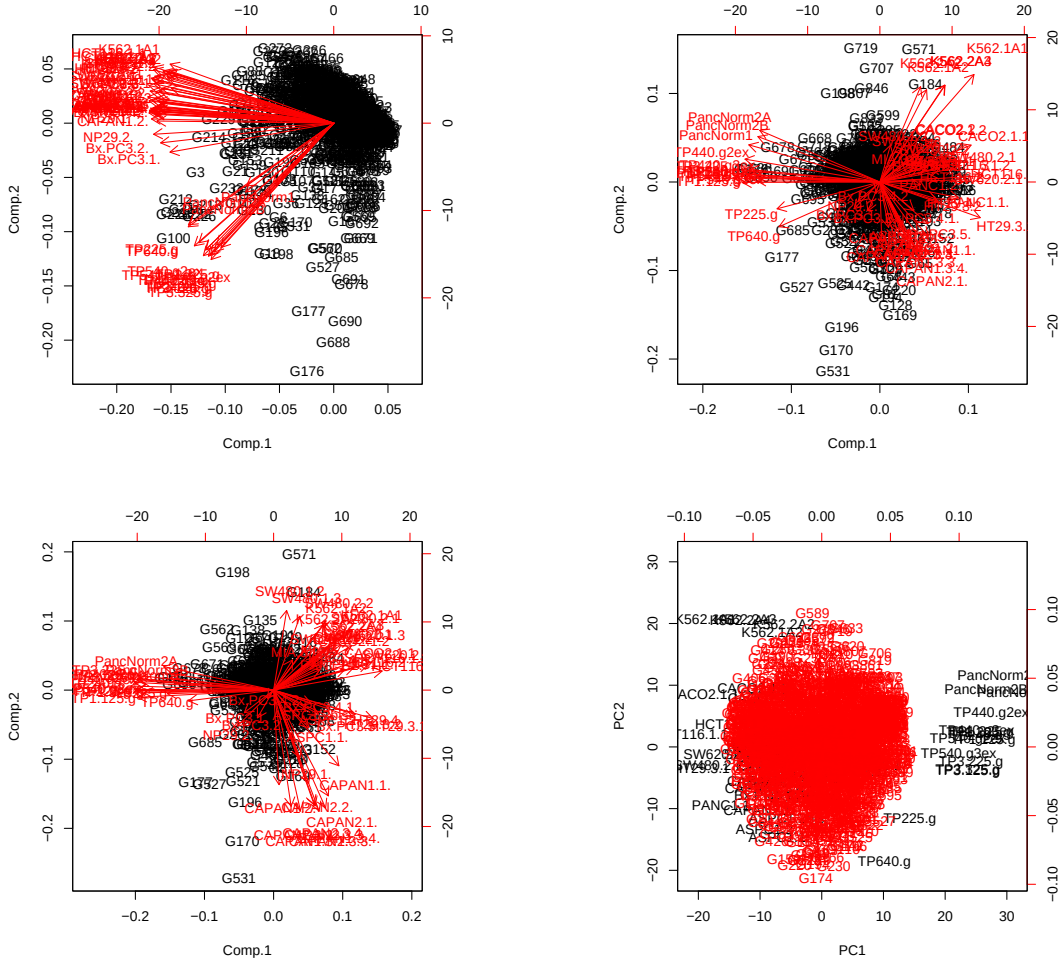


FIG. 3.10 – *Pancréas : différentes options de l'ACP. a : gènes X tissus centrés. b : gènes centrés X tissus centrés. c : gènes centrés X tissus centrés réduits. d : tissus X gènes centrés réduits.*

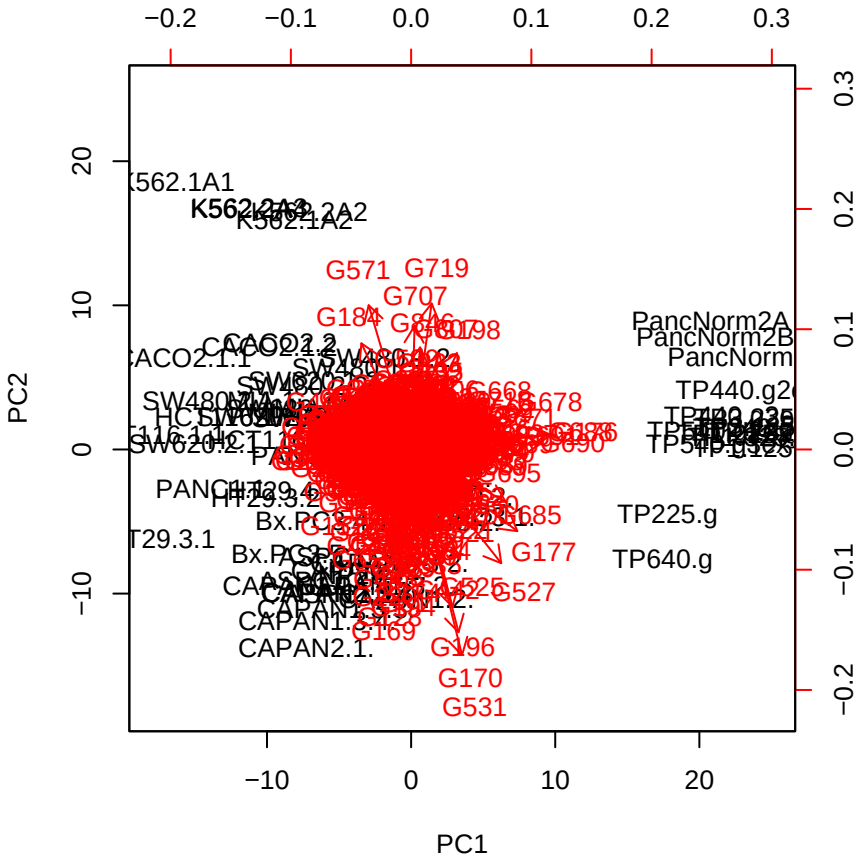


FIG. 3.11 – *Pancréas : ACP du tableau mesurant les expressions de 871 gènes (variables centrées) sur 65 échantillons biologiques.*

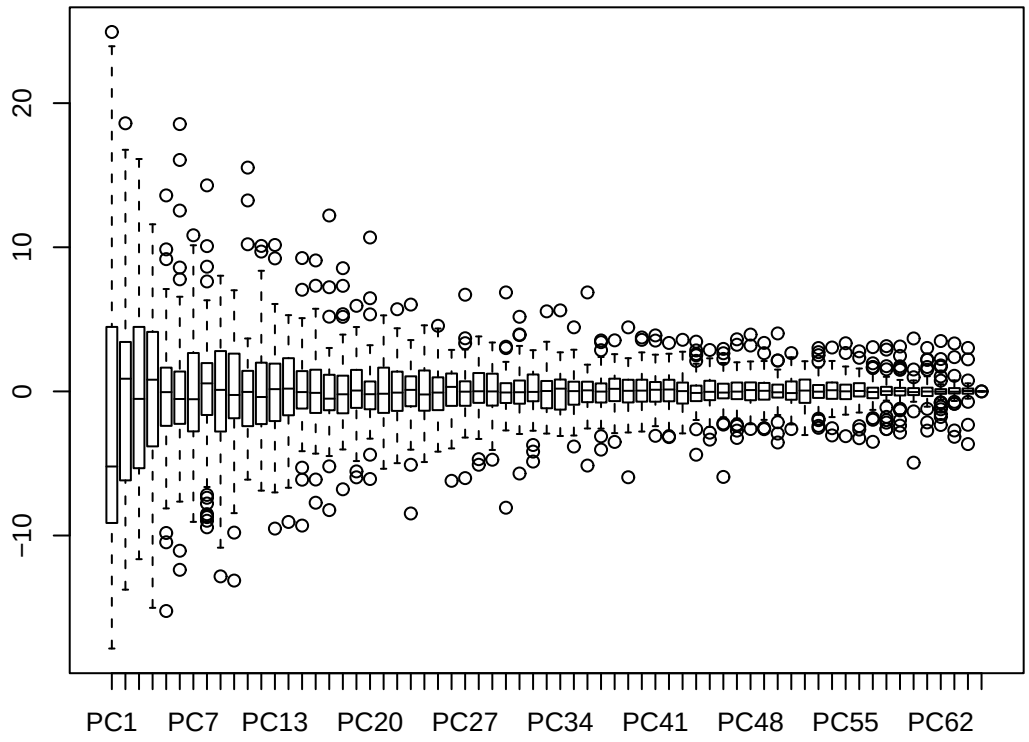


FIG. 3.12 – *Pancréas : Distribution des composantes principales dans l'ordre décroissant de leur variance. Quatre composantes ou dimensions semblent pertinentes à retenir sur les 65 initiales (pas 871) qui est le rang de la matrice diagonalisée.*

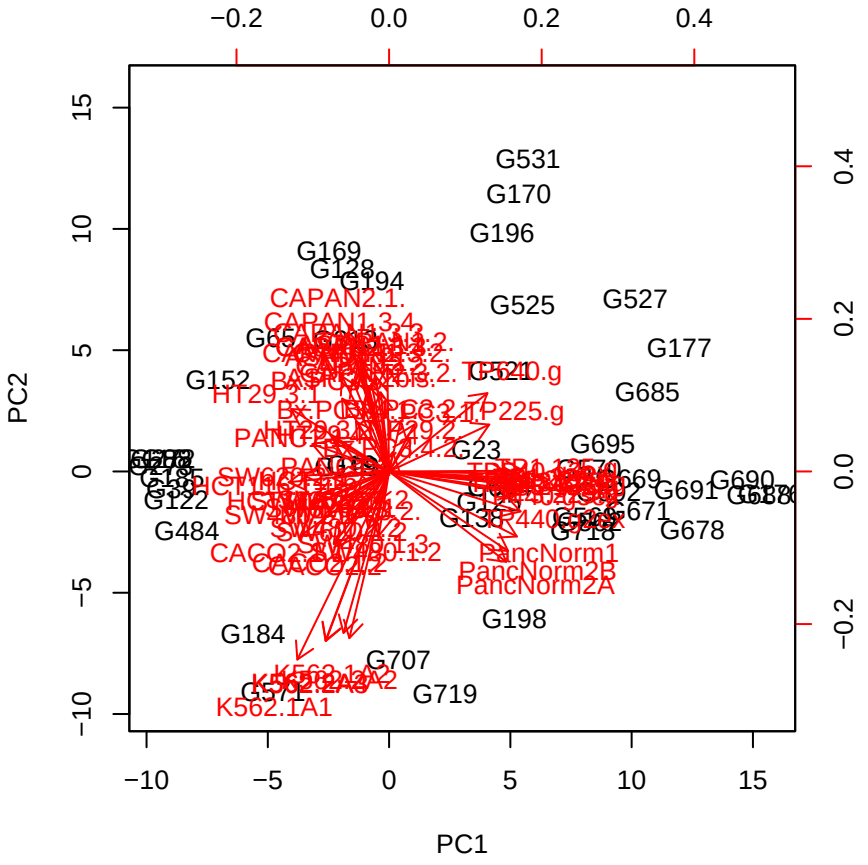


FIG. 3.13 – *Pancréas : Représentation dans le premier plan des 46 gènes contribuant le plus à la définition des 4 premiers axes.*

Bien évidemment, cette dernière distinction est la plus intéressante et ouvre des pistes de recherches ciblées sur ces gènes. Néanmoins, l'introduction des lignées cellulaire semble apporter une contribuation intéressante sous la forme d'une "calibration" des anbalyses.

D'autres approches sont possibles qu'il serait intéressant de comparer pour apporter ainsi plus de "confiance" dans l'analyse des résultats. Des tests liés à un modèle d'analyse de variance plus ou moins sophistiqué (à effet aléatoire) ou la recherche d'un modèle susceptible de discriminer au mieux certains facteurs. Le choix des gènes s'apparente, dans ce dernier cas, à un choix de variables en régression ; les forêts aléatoires de Breiman (2001) ou d'autres méthodes issues de la théorie de l'apprentissage (cf. Besse, 2003 pour une introduction) semblent apporter des réponses intéressantes. Cela sort du strict cadre exploratoire de ce cours.

9 Exemple : nutrition chez la souris

Nous donnons pour cet exemple le graphique des premières valeurs propres (figure 3.14) qui conduit à considérer trois dimensions représentant environ les deux tiers de l'inertie globale.

Les figures 3.15 et 3.16 donnent la représentation des souris et celle des gènes, d'abord

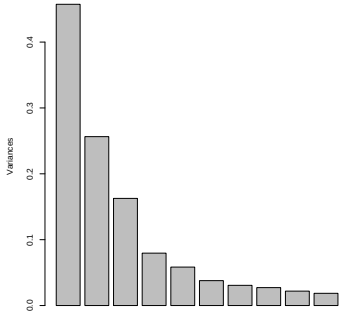


FIG. 3.14 – *Souris : éboulis des dix premières valeurs propres de l'ACP.*

dans le premier plan principal, ensuite dans celui correspondant aux dimensions 1 et 3. Dans le cadre de cette ACP, il est cohérent de rechercher quels sont les 25% des gènes contribuant le plus à la définition de l'espace propre à trois dimensions jugé pertinent. Avec cette sélection, la représentation des variables ainsi restreinte à 30 gènes est plus facilement lisible sur les figures 3.15 et 3.16. Toutefois, dans le cas d'une puce pangénomique, avec potentiellement plusieurs milliers de gènes, une telle représentation ne serait pas exploitable.

Le premier plan (Fig. 3.15) doit être interprété globalement puisque sa première bissectrice sépare exactement les souris `WT` des souris `PPAR`. Les gènes à coordonnées négatives sur l'axe 1 et positives sur l'axe 2 sont sensiblement plus exprimés chez les souris `WT`, en particulier `CYP3A11`, `CYP4A10`, `CYP4A14`, `THIOL`, `PMDCI`, `GSTpi2`, `L.FABP` et `FAS`. À l'inverse, les gènes à forte coordonnée négative sur l'axe 2 s'expriment davantage chez les souris `PPAR`, par exemple, `S14`, `PAL` et `CAR1`. Ceci est en partie connu des biologistes (Aoyama *et al.*, 1998).

Le phénomène le plus marquant concernant l'axe 3 (Fig. 3.16) est l'opposition, chez les souris WT, entre les régimes `dha` (1), dont les coordonnées sont toutes positives, et `efad` (2), dont les coordonnées sont toutes négatives. Les gènes les plus exprimés dans le premier cas (régime `dha` chez les souris WT) sont `CYP3A11`, `CYP4A10`, `CYP4A14`, `CYP2c29` et `CAR1` ; dans le second cas (régime `efad` chez les mêmes souris), il s'agit des gènes `FAS`, `S14`, `Lpin` et `Lpin1`. Parmi ces régulations, on note une opposition entre les `CYP4A`, connus pour être impliqués dans le catabolisme des acides gras, et les gènes `FAS` et `S14` impliqués eux dans la synthèse des lipides. Par ailleurs, la régulation de `CYP3A11` par le DHA a déjà été décrite dans Berger *et al.* (2002).

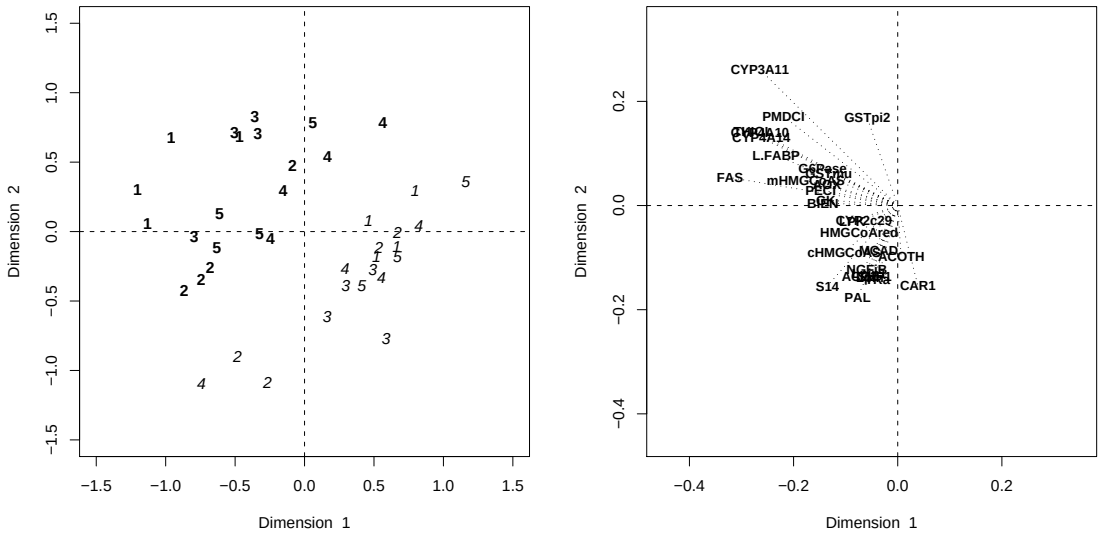


FIG. 3.15 – *Représentations de l'ACP sur le premier plan principal. À gauche : individus-souris identifiés par leur génotype (*`WT` *en gras,* `PPAR` *en italique) et leur régime (1-*`dha`*, 2-*`efad`*, 3-*`lin` *, 4-*`ref`*, 5-*`tsol`*). À droite : 30 variables-gènes qui contribuent le plus aux trois premiers axes.*

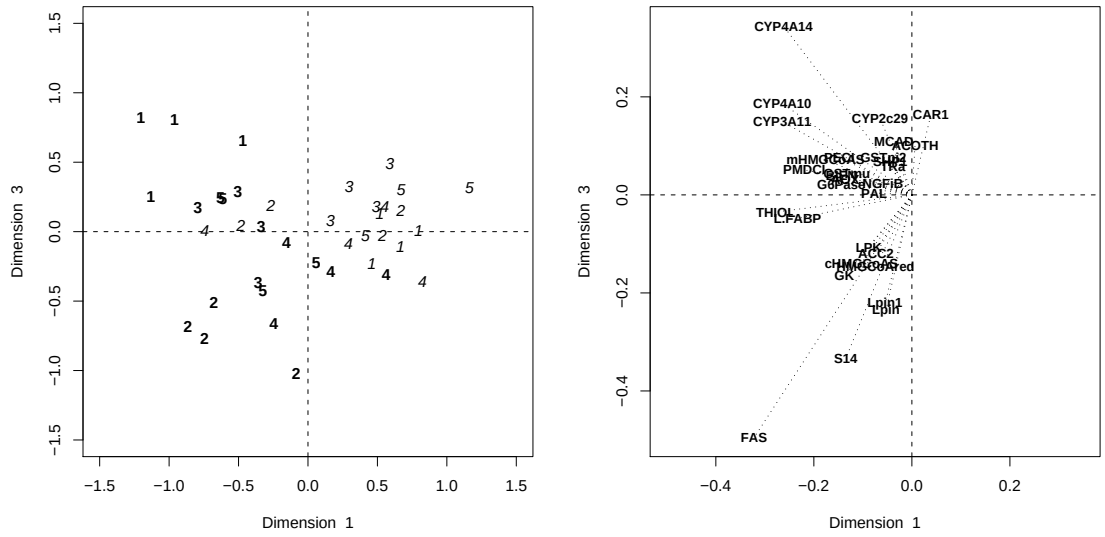


FIG. 3.16 – *Représentations de l'ACP sur le plan composé des axes principaux 1 et 3 avec les mêmes conventions que pour la figure 3.15.*

Chapitre 4

Analyse Factorielle Discriminante

1 Introduction

1.1 Données

Les données sont constituées de

- p variables *quantitatives* $X^1, \ldots, X^p$ jouant le rôle de variables explicatives comme dans le modèle linéaire,
- une variable *qualitative* T, à m modalités $\{\mathcal{T}_1, \ldots, \mathcal{T}_m\}$, jouant le rôle de variable à expliquer.

La situation est analogue à celle de la régression linéaire multiple mais, comme la variable à expliquer est qualitative, on aboutit à une méthode très différente. Les variables sont observées sur l'ensemble Ω des n individus affectés des poids $w_i > 0$, $(\sum_{i=1}^n w_i = 1)$, et l'on pose

$$\mathbf{D} = \text{ diag}(w_i \,; i = 1, \ldots, n).$$

La variable T engendre une partition $\{\Omega_\ell \,; \ell = 1, \ldots, m\}$ de l'ensemble Ω des individus dont chaque élément est d'effectif n_ℓ.

On note $\mathbf{T}$ $(n \times m)$ la matrice des indicatrices des modalités de la variable T ; son terme général est

$$t_i^\ell = t^\ell(\omega_i) = \begin{cases} 1 & \text{si } T(\omega_i) = \mathcal{T}_\ell \\ 0 & \text{sinon} \end{cases} .$$

En posant

$$\overline{w_\ell} = \sum_{i \in \Omega_\ell} w_i,$$

il vient

$$\overline{\mathbf{D}} = \mathbf{T}'\mathbf{D}\mathbf{T} = \text{ diag}(\overline{w_1}, \ldots, \overline{w_m}).$$

1.2 Objectifs

Deux techniques cohabitent sous la même appellation d'analyse discriminante :

descriptive : cette méthode recherche, parmi toutes les ACP possibles sur les variables X^j, celle dont les représentations graphiques des individus *discriminent* "au mieux" les m classes engendrées par la variable T (e.g. recherche de facteurs de risque en statistique médicale) ;

décisionnelle : connaissant, pour un individu donné, les valeurs des Y^j mais pas la modalité de T, cette méthode consiste à affecter cet individu à une modalité (e.g. reconnaissance de formes). Cette méthode est décrite dans la partie *modélisation* de ce cours.

Remarque. — Lorsque le nombre et les caractéristiques des classes sont connues, il s'agit d'une *discrimination* ; sinon, on parle de *classification* ou encore, avec des hypothèses sur les distributions, de *reconnaissance de mélanges*.

1.3 Notations

On note $\mathbf{X}$ la matrice $(n \times p)$ des données quantitatives, $\mathbf{G}$ la matrice $(m \times p)$ des barycentres des classes :

$$\mathbf{G} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{X} = \begin{bmatrix} \mathbf{g}_1{}' \\ \vdots \\ \mathbf{g}_m{}' \end{bmatrix} \text{ où } \mathbf{g}_\ell = \frac{1}{\overline{w_\ell}} \sum_{i \in \Omega_\ell} w_i \mathbf{x}_i,$$

et $\mathbf{X}_e$ la matrice $(n \times p)$ dont la ligne i est le barycentre $\mathbf{g}_\ell$ de la classe Ω_ℓ à laquelle appartient l'individu i :

$$\mathbf{X}_e = \mathbf{T}\mathbf{G} = \mathbf{P}\mathbf{G} ;$$

$\mathbf{P} = \mathbf{T}\overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}$ est la matrice de projection $\mathbf{D}$-orthogonale sur le sous-espace engendré par les indicatrices de T ; c'est encore l'espérance conditionnelle sachant T.

Deux matrices "centrées" sont définies de sorte que $\overline{\mathbf{X}}$ se décompose en

$$\overline{\mathbf{X}} = \overline{\mathbf{X}}_r + \overline{\mathbf{X}}_e$$

avec

$$\overline{\mathbf{X}}_r = \mathbf{X} - \mathbf{X}_e \text{ et } \overline{\mathbf{X}}_e = \mathbf{X}_e - \mathbf{1}_n \overline{\mathbf{x}}'.$$

On note également $\overline{\mathbf{G}}$ la matrice centrée des barycentres :

$$\overline{\mathbf{G}} = \mathbf{G} - \mathbf{1}_m \overline{\mathbf{x}}'.$$

On appelle alors variance intraclasse (within) ou résiduelle :

$$\mathbf{S}_r = \overline{\mathbf{X}}_r{}'\mathbf{D}\overline{\mathbf{X}}_r = \sum_{\ell=1}^{m} \sum_{i \in \Omega_\ell} w_i (\mathbf{x}_i - \mathbf{g}_\ell)(\mathbf{x}_i - \mathbf{g}_\ell)',$$

et variance interclasse (between) ou expliquée :

$$\mathbf{S}_e = \overline{\mathbf{G}}'\overline{\mathbf{D}\mathbf{G}} = \overline{X}_e'\mathbf{D}\overline{X}_e = \sum_{\ell=1}^{m} \overline{w_\ell} (\mathbf{g}_\ell - \overline{\mathbf{x}})(\mathbf{g}_\ell - \overline{\mathbf{x}})'.$$

PROPOSITION 4.1. — *La matrice des covariances se décompose en*

$$\mathbf{S} = \mathbf{S}_e + \mathbf{S}_r.$$

2 Définition

2.1 Modèle

Dans l'espace des individus, le principe consiste à projeter les individus dans une direction permettant de mettre en évidence les groupes. À cette fin, Il faut privilégier la variance interclasse au détriment de la variance intraclasse considérée comme due au bruit.

En ACP, pour chaque effet $\mathbf{z}_i$ à estimer, on ne dispose que d'une observation $\mathbf{x}_i$; dans le cas de l'AFD on considère que les éléments d'une même classe Ω_ℓ sont les observations répétées n_ℓ fois du même effet $\mathbf{z}_\ell$ pondéré par $\overline{w_\ell} = \sum_{i\in\Omega_\ell} w_i$. Le modèle devient donc :

$$\begin{array}{l} \{\mathbf{x}_i\,;i=1,\ldots,n\},\ n \text{ vecteurs indépendants de } E,\\ \forall \ell, \forall i \in \Omega_\ell, \mathbf{x}_i = \mathbf{z}_\ell + \boldsymbol{\varepsilon}_i \text{ avec } \left\{ \begin{array}{l} E(\boldsymbol{\varepsilon}_i) = 0,\ \text{var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Gamma},\\ \boldsymbol{\Gamma} \text{ régulière et inconnue,} \end{array} \right.\\ \exists A_q,\ \text{sous-espace affine de de dimension } q \text{ de } E \text{ tel que}\\ \forall \ell, \mathbf{z}_\ell \in A_q,\ (q < \min(p, m-1)). \end{array} \tag{4.1}$$

Remarque. — Soit $\overline{\mathbf{z}} = \sum_{\ell=1}^{m} \overline{w_\ell}\mathbf{z}_\ell$. Le modèle entraîne que $\overline{\mathbf{z}} \in A_q$. Soit E_q le sous-espace de dimension q de E tel que $A_q = \overline{\mathbf{z}} + E_q$. Les paramètres à estimer sont E_q et $\{\mathbf{z}_\ell\,;\ell=1,\ldots,m\}$; $\overline{w_\ell}$ est un paramètre de nuisance qui ne sera pas considéré.

2.2 Estimation

L'estimation par les moindres carrés s'écrit ainsi :

$$\min_{E_q,\mathbf{z}_\ell} \left\{ \sum_{\ell=1}^{m} \sum_{i\in\Omega_\ell} w_i \left\|\mathbf{x}_i - \mathbf{z}_\ell\right\|^2_{\mathbf{M}}\ ;\ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}.$$

Comme on a

$$\sum_{\ell=1}^{m} \sum_{i\in\Omega_\ell} w_i \left\|\mathbf{x}_i - \mathbf{z}_\ell\right\|^2_{\mathbf{M}} = \sum_{\ell=1}^{m} \sum_{i\in\Omega_\ell} w_i \left\|\mathbf{x}_i - \mathbf{g}_\ell\right\|^2_{\mathbf{M}} + \sum_{\ell=1}^{m} \overline{w_\ell} \left\|\mathbf{g}_\ell - \mathbf{z}_\ell\right\|^2_{\mathbf{M}},$$

on est conduit à résoudre :

$$\min_{E_q,z_\ell} \left\{ \sum_{\ell=1}^{m} \overline{w_\ell} \left\|\mathbf{g}_\ell - \mathbf{z}_\ell\right\|^2_{\mathbf{M}}\ ;\ \dim(E_q) = q, \mathbf{z}_\ell - \overline{\mathbf{z}} \in E_q \right\}.$$

La covariance $\sigma^2\boldsymbol{\Gamma}$ du modèle (4.1) étant inconnue, il faut l'estimée. Ce modèle stipule que l'ensemble des observations d'une même classe Ω_l suit une loi (inconnue) de moyenne $\mathbf{z}_e ll$ et de variance $\boldsymbol{\Gamma}$. Dans ce cas particulier, la matrice de covariances intraclasse ou matrice des covariances résiduelles empiriques $\mathbf{S}_r$ fournit donc une estimation "optimale" de la métrique de référence :

$$\mathbf{M} = \widehat{\boldsymbol{\Gamma}}^{-1} = \mathbf{S}_r^{-1}$$

.

PROPOSITION 4.2. — *L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle 4.1 est obtenue par l'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$. C'est l'Analyse Factorielle Discriminante (AFD) de $(\mathbf{X}|\mathbf{T}, \mathbf{D})$.*

3 Réalisation de l'AFD

Les expressions matricielles définissant les représentations graphiques et les aides à l'interprétation découlent de celles de l'ACP.

3.1 Matrice à diagonaliser

L'ACP de $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ conduit à l'analyse spectrale de la matrice positive $\mathbf{S}_r^{-1}$-symétrique :

$$\overline{\mathbf{G}}'\overline{\mathbf{D}}\,\overline{\mathbf{G}}\mathbf{S}_r^{-1} = \mathbf{S}_e\mathbf{S}_r^{-1}.$$

Comme $\mathbf{S}_r^{-1}$ est régulière, cette matrice est de même rang que $\mathbf{S}_e$ et donc de même rang que $\mathbf{G}$ qui est de dimension $(m \times p)$. Les données étant centrées lors de l'analyse, le rang de la matrice à diagonaliser est

$$h = \text{rang}(\mathbf{S}_e\mathbf{S}_r^{-1}) \leq \inf(m-1, p),$$

qui vaut en général $m-1$ c'est-à-dire le nombre de classes moins un.

On note $\lambda_1 \geq \cdots \geq \lambda_h > 0$ les valeurs propres de $\mathbf{S}_e\mathbf{S}_r^{-1}$ et $\mathbf{v}^1, \ldots, \mathbf{v}^h$ les vecteurs propres$\mathbf{S}_r^{-1}$-orthonormés associés. On pose

$$\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_h) \text{ et } \mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^h].$$

Les vecteurs $\mathbf{v}^k$ sont appelés *vecteurs discriminants* et les sous-espaces vectoriels de dimension 1 qu'ils engendrent dans $\mathbb{R}^p$ les *axes discriminants*.

3.2 Représentation des individus

L'espace des individus est ($\mathbb{R}^p$, b. c., $\mathbf{S}_r^{-1}$). Une représentation simultanée des individus $\mathbf{x}_i$ et des barycentres $\mathbf{g}_\ell$ des classes par rapport aux mêmes axes discriminants est obtenue dans cet espace au moyen des coordonnées :

$$\begin{aligned} \mathbf{C} &= \overline{\mathbf{X}}\mathbf{S}_r^{-1}\mathbf{V} \text{ pour les individus et} \\ \overline{\mathbf{C}} &= \overline{\mathbf{G}}\mathbf{S}_r^{-1}\mathbf{V} = \overline{\mathbf{D}}^{-1}\mathbf{T}'\mathbf{D}\mathbf{C} \text{ pour les barycentres.} \end{aligned}$$

Les individus initiaux sont projetés comme des individus supplémentaires dans le système des axes discriminants. Comme en ACP, on peut calculer des cosinus carrés pour préciser la qualité de représentation de chaque individu.

Il est utile de différencier graphiquement la classe de chaque individu afin de pouvoir apprécier visuellement la qualité de la discrimination.

3.3 Représentation des variables

L'espace des variables est ($\mathbb{R}^m$, b. c., $\overline{\mathbf{D}}$). Chaque variable X^j est représenté par un vecteur dont les coordonnées dans le système des axes factoriels est une ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.

3.4 Interprétations

Les interprétations usuelles : la norme est un écart-type, un cosinus d'angle est un coefficient de corrélation, doivent être faites en termes d'écarts-types et de corrélations *expliquées* par la partition.

La représentation des variables est utilisée pour interprétée les axes en fonction des variables initiales conjointement avec la matrice des corrélations expliquées variables×facteurs : $\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$. La matrice $\mathbf{\Sigma}_e^{-1}$ étant la matrice diagonale des écarts-types expliqués σ_e^j c'est-à-dire des racines carrées des éléments diagonaux de la matrice $\mathbf{S}_e$.

Le point pratique essentiel est de savoir si la représentation des individus-barycentres et des individus initiaux permet de faire une bonne discrimination entre les classes définies par la variable

T. Si ce n'est pas le cas, l'AFD ne sert à rien, les X^j n'expliquent pas T. Dans le cas favorable, le graphique des individus permet d'interpréter la discrimination en fonction des axes et, celui des variables, les axes en fonction des variables initiales. La synthèse des deux permet l'interprétation de T selon les X^j.

4 Variantes de l'AFD

4.1 Individus de mêmes poids

L'AFD peut être définie de différentes façon. Dans la littérature anglo-saxonne, et donc dans la version standard d'AFD du logiciel SAS (procédure `candisc`), ce sont les estimations sans biais des matrices de variances "intra" (within) et "inter" (between) qui sont considérées dans le cas d'individus de mêmes poids $1/n$.

Dans ce cas particulier,

$$\mathbf{D} = \frac{1}{n}\mathbf{I}_n \text{ et } \overline{\mathbf{D}} = \frac{1}{n}\text{diag}(n_1, \ldots, n_m) \text{ où } n_\ell = \text{card}(\Omega_\ell)$$

et les matrices de covariances empiriques ont alors pour termes généraux :

$$\begin{aligned}
(\mathbf{S})_j^k &= \frac{1}{n}\sum_{i=1}^{n}(x_i^j - \overline{x}^j)(x_i^k - \overline{x}^k),\\
(\mathbf{S}_e)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m} n_\ell(g_\ell^j - \overline{x}^j)(g_\ell^k - \overline{x}^k),\\
(\mathbf{S}_r)_j^k &= \frac{1}{n}\sum_{\ell=1}^{m}\sum_{i\in\Omega_\ell}(x_i^j - g_\ell^j)(x_i^k - g_\ell^k).
\end{aligned}$$

Du point de vue de le Statistique inférentielle, on sait que les quantités calculées ci-dessus ont respectivement $(n-1)$, $(m-1)$ et $(n-m)$ degrés de liberté. En conséquence, ce point de vue est obtenu en remplaçant dans les calculs

$$\begin{aligned}
\mathbf{S} \quad &\text{par} \quad \mathbf{S}^* = \frac{n}{n-1}\mathbf{S},\\
\mathbf{S}_e \quad &\text{par} \quad \mathbf{S}_e^* = \mathbf{B} = \frac{n}{m-1}\mathbf{S}_e,\\
\mathbf{S}_r \quad &\text{par} \quad \mathbf{S}_r^* = \mathbf{W} = \frac{n}{n-m}\mathbf{S}_r.
\end{aligned}$$

Les résultats numériques de l'AFD se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e^*\mathbf{S}_r^{*-1}$	$=$	$\frac{n-m}{m-1}\mathbf{S}_e\mathbf{S}_r^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}^*$	$=$	$\frac{n-m}{m-1}\mathbf{\Lambda}$,
– vecteurs propres :	$\mathbf{V}^*$	$=$	$\sqrt{\frac{n}{n-m}}\mathbf{V}$,
– représentation des barycentres :	$\overline{\mathbf{C}}^*$	$=$	$\sqrt{\frac{n-m}{n}}\overline{\mathbf{C}}$,
– représentation des variables :	$\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\sqrt{\frac{n}{m-1}}\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{*-1}\mathbf{V}^*\mathbf{\Lambda}^{*1/2}$	$=$	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Ainsi, les représentations graphiques sont identiques à un facteur d'échelle près tandis que les parts de variance expliquée et les corrélations variables-facteurs sont inchangées.

4.2 Métrique de Mahalanobis

L'AFD est souvent introduite dans la littérature francophone comme un cas particulier d'Analyse Canonique entre un ensemble de p variables quantitatives et un ensemble de m variables indicatrices des modalités de T. La proposition suivante établit les relations entre les deux approches :

PROPOSITION 4.3. — *l'ACP de* $(\mathbf{G}, \mathbf{S}_r^{-1}, \overline{\mathbf{D}})$ *conduit aux mêmes vecteurs principaux que l'ACP de* $(\mathbf{G}, \mathbf{S}^{-1}, \overline{\mathbf{D}})$. *Cette dernière est l'ACP des barycentres des classes lorsque l'espace des individus est muni de la métrique dite de Mahalanobis* $\mathbf{M} = \mathbf{S}^{-1}$ *et l'espace des variables de la métrique des poids des classes* $\overline{\mathbf{D}}$.

Les résultats numériques de l'AFD se trouvent alors modifiés de la façon suivante :

– matrice à diagonaliser :	$\mathbf{S}_e\mathbf{S}^{-1}$,
– valeurs propres :	$\mathbf{\Lambda}(\mathbf{I}+\mathbf{\Lambda})^{-1}$,
– vecteurs propres :	$\mathbf{V}(\mathbf{I}+\mathbf{\Lambda})^{1/2}$,
– représentation des barycentres :	$\overline{\mathbf{C}}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}$,
– représentation des variables :	$\mathbf{V}\mathbf{\Lambda}^{1/2}$,
– corrélations variables-facteurs :	$\mathbf{\Sigma}_e^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

Les représentations graphiques des individus (voir ci-dessus) ne diffèrent alors que d'une homothétie et conduisent à des interprétations identiques, les corrélations variables-facteurs ainsi que les représentations des variables sont inchangées.

5 Exemples

Ce chapitre est illustré par une comparaison des sorties graphiques issues d'une ACP et d'une AFD. Les données décrivent trois classes d'insectes sur lesquels ont été réalisées 6 mesures anatomiques. On cherche à savoir si ces mesures permettent de retrouver la typologie de ces insectes. Ce jeu de données est très "scolaire" mais il montre bien le rôle joué par la métrique en AFD qui a tendance à rendre les classes plus sphériques autour de leur barycentre.

Cette technique n'est pas très adaptée aux problèmes liés aux données d'expression. En effet, le nombre de paramètre discriminants y très important et conduit le plus souvent à un problème d'indétermination. Plus précisément, avec le nombre de variables/gènes présents, il est toujours possible de trouver un ou des axes discriminants différents types d'échantillons biologiques. Le problème est en fait mal posé (plus d'inconnues que d'équations). Une sélection drastique du nombre de gènes préalable à l'AFD doit donc être réalisée ; elle a été ici conduite à la l'aide de la procédure `discrim` de SAS qui recherche avec un algorithme de type `backward` les variables les plus discriminantes. Cela conduit aux résultats de la figure 5.

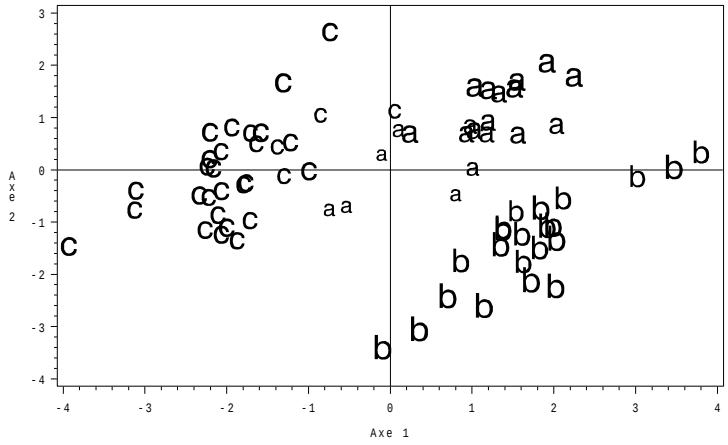


FIG. 4.1 – *Insectes : premier plan factoriel de l'ACP.*

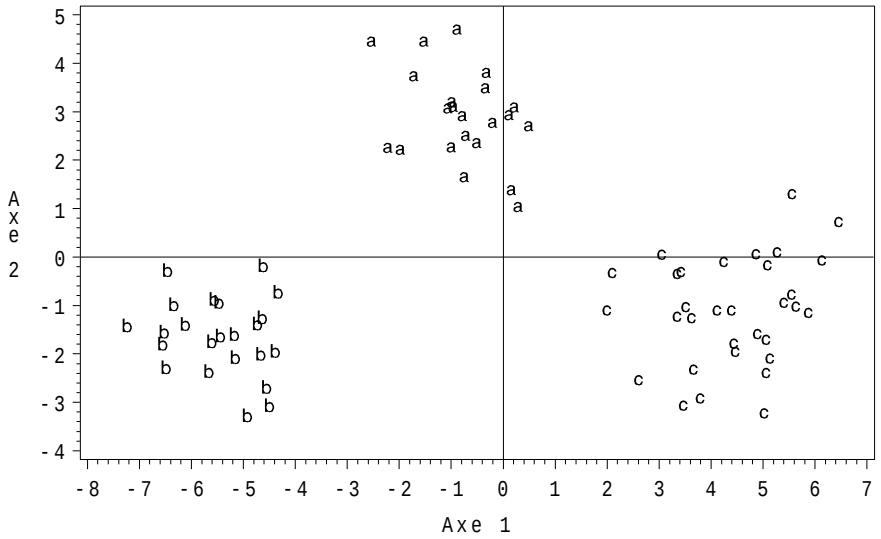


FIG. 4.2 – *Insectes : premier plan factoriel de l'AFD.*

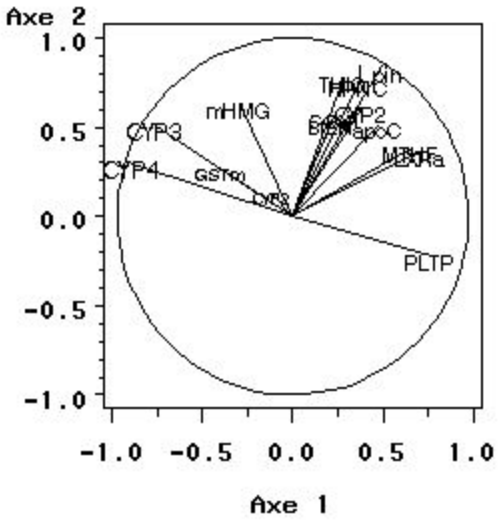


FIG. 4.3 – *Souris : premier plan factoriel de l'AFD. Représentation des gènes participant le plus à la discrimination des régimes des souris sauvages.*

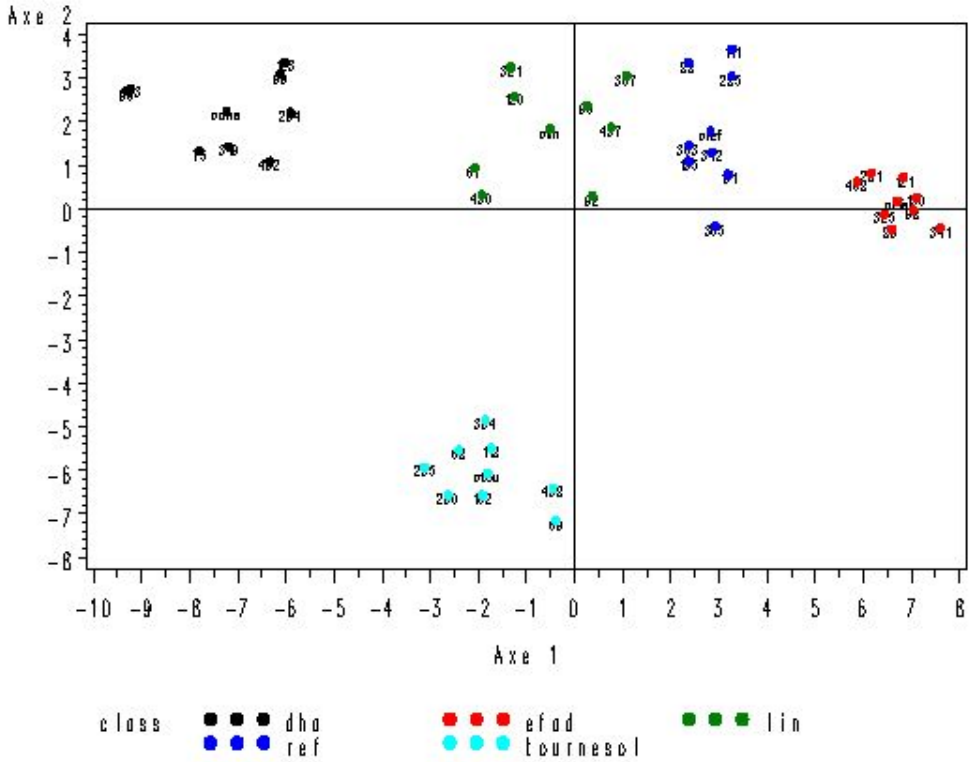


FIG. 4.4 – *Souris : premier plan factoriel de l'AFD. Représentation des souris sauvages en fonction de leur classe de régime.*

Chapitre 5

Positionnement multidimensionnel

1 Introduction

Considérons n individus. Contrairement aux chapitres précédents, on ne connaît pas les observations de p variables sur ces n individus mais dans certains cas les $n(n-1)/2$ valeurs d'un indice (de distance, dissimilarité ou dissemblance) observées ou construites pour chacun des couples d'individus. Ces informations sont contenues dans une matrice $(n \times n)$ $\boldsymbol{\mathcal{D}}$. L'objectif du *positionnement multidimensionnel* (multidimensional scaling, ou MDS, ou ACP d'un tableau de distances) est de construire, à partir de cette matrice, une représentation euclidienne des individus dans un espace de dimension réduite q qui approche au "mieux" les indices observés. Autrement dit, visuellement le graphique obtenu représente en dimension (en général) 2 la meilleure approximation des distances observées entre les individus pouvant être des gènes ou des échantillons biologiques.

Le principal intérêt de cette technique est donc de pouvoir observer graphiquement le même ensemble de données à travers différentes "optiques" et même d'en comparer les représentations ; chaque optique est définie par la façon dont on mesure des distances ou dissimilarités entre les objets.

Citons trois exemples typiques dans le cas spécifique de gènes décrits par leurs transcrits :

- chaque gène est un vecteur dans un espace vectoriel muni de la distance euclidienne classique (racine de la somme des carrés des écarts). Le MDS ou ACP du tableau des distances qui en découle est équivalent à l'ACP dans laquelle les gènes sont les individus (les lignes).
- On mesure la dissimilarité entre deux gènes X^j et X^k par $1-\text{cor}(X^j, X^k)$ faisant intervenir la corrélation linéaire de Pearson ou celle robuste sur les rangs de Spearman. Les gènes co-régulés (fortement positivement corrélés) sont très proches, les gènes associés dans un mécanisme d'inhibition (fortement négativement corrélés) seront aussi proches.
- On mesure la distance entre deux gènes par $\sqrt{1-\text{cor}(X^j, X^k)^2}$. Elle vérifie, dans ce cas, les propriétés qui en font une distance euclidienne. Co-régulés ou inhibés, les gènes corrélés positivement ou négativement sont proches dans les représentations graphiques.

Exemple élémentaire : Considérons un tableau contenant les distances kilométriques par route (Source : IGN) entre 47 grandes villes en France et dans les pays limitrophes. Toutes ces valeurs sont rangées dans le triangle inférieur d'une matrice carrée avec des 0 sur la diagonale. La structure du réseau routier, le relief, font que cette matrice de distances n'est pas euclidienne qui, dans ce cas, correspondrait à la distance à "vol d'oiseau". Mais, comme le montre le graphique issu d'un positionnement multidimensionnel, l'approximation euclidienne en est très proche.

Le MDS étant encore une technique factorielle, comme en ACP il est nécessaire de déterminer le nombre de dimensions fixant la taille de l'espace de représentation. Le graphique représentant

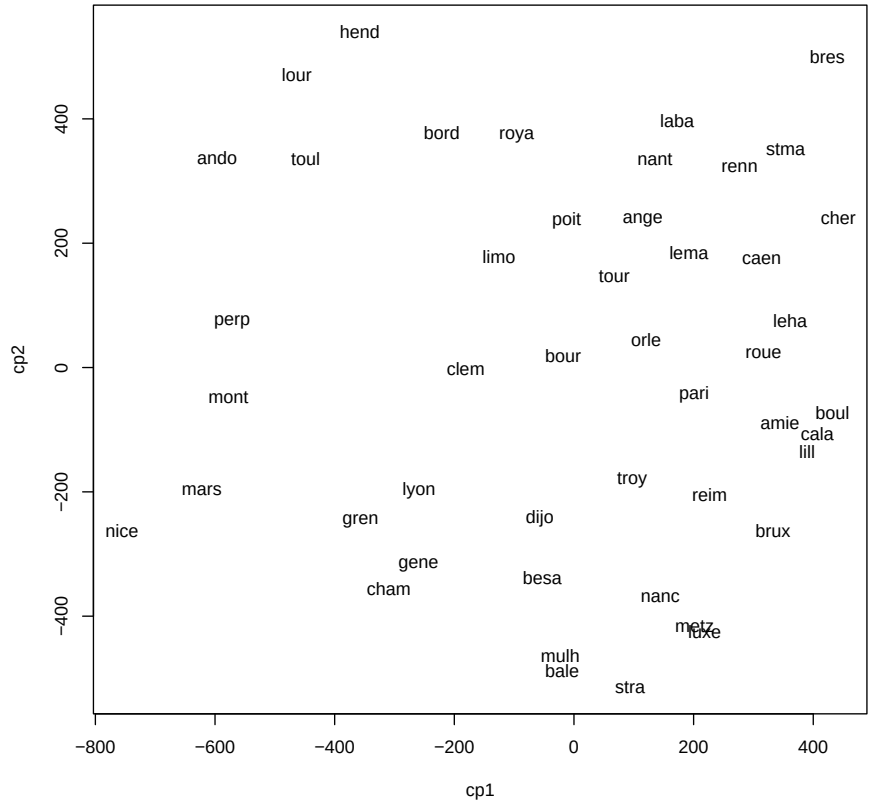


FIG. 5.1 – *Villes : Positionnement de 47 villes à partir de la matrice de leurs distances kilométriques.*

la décroissance des valeurs propres aide à ce choix.

Les preuves et développements théoriques sont omis dans cet exposé succinct, ils sont à chercher dans la bibliographie. Voir par exemple Mardia et col. (1979).

2 Distance, similarités

Rappelons quelques propriétés et définitions élémentaires mais basiques à propos de la notion de distance.

2.1 Définitions

DÉFINITION 5.1. —

- *Une matrice* $(n \times n)$ $\boldsymbol{\mathcal{D}}$ *est appelée matrice de distance si elle est symétrique et si :*

$$d_j^j = 0 \text{ et } \forall (j,k), j \neq k, d_j^k \geq 0.$$

- *Une matrice* $(n \times n)$ $\boldsymbol{\mathcal{C}}$ *est appelée matrice de similarité si elle est symétrique et si*

$$\forall (j,k), c_j^k \leq c_j^j.$$

Une matrice de similarité se transforme en matrice de distance par :

$$d_j^k = (c_j^j + c_k^k - 2c_j^k)^{-1/2}.$$

DÉFINITION 5.2. — *Une matrice de distance est dite euclidienne s'il existe une configuration de vecteurs* $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ *dans un espace vectoriel euclidien* E *de sorte que*

$$d_j^{k\,2} = \langle \mathbf{x_j} - \mathbf{x_k}, \mathbf{x_j} - \mathbf{x_k} \rangle \ .$$

On note $\mathbf{A}$ la matrice issue de $\boldsymbol{\mathcal{D}}$ de terme général $d_j^k = -d_j^{k\,2}/2$ et $\mathbf{H}$ la matrice de centrage :

$$\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}'\mathbf{D},$$

qui est la matrice de projection sur le sous-espace $\mathbf{D}$-orthogonal au vecteur $\mathbf{1}$ dans l'espace euclidien F des variables muni de la métrique des poids.

PROPOSITION 5.3. —

- *Soit* $\boldsymbol{\mathcal{D}}$ *une matrice de distance et* $\mathbf{B}$ *la matrice obtenue par double centrage de la matrice* $\mathbf{A}$ *issue de* $\boldsymbol{\mathcal{D}}$ *:*

$$\mathbf{B} = \mathbf{HAH}',$$

  alors $\boldsymbol{\mathcal{D}}$ *est une matrice euclidienne si et seulement si* $\mathbf{B}$ *est positive (toutes ses valeurs propres sont positives ou nulles).*
- *Si la matrice de similarité* $\boldsymbol{\mathcal{C}}$ *est positive alors la matrice de distance* $\boldsymbol{\mathcal{D}}$ *déduite est euclidienne.*

2.2 Distances entre variables

L'un des intérêts pratiques du positionnement multidimensionnel est d'aider à comprendre, visualiser, les structures de liaison dans un grand ensemble de variables. On obtient ainsi des indications pour guider le choix d'un sous-ensemble de variables, par exemple les plus liées à une variable à expliquer. Cette approche nécessite la définition d'indices de similarité entre variables. Beaucoup sont proposés dans la littérature et concrètement utilisés pour les données d'expression. Les gènes étant considérés comme des variables, on s'intéresse alors à différents critères basés sur la corrélation linéaire usuelle de Pearson ou robuste (non paramétrique de Spearman).

On note X et Y deux variables statistiques dont les observations sur les mêmes n individus sont rangées dans les vecteurs *centrés* $\mathbf{x}$ et $\mathbf{y}$ de l'espace euclidien F muni de la métrique des poids $\mathbf{D}$. On vérifie facilement :

$$\begin{aligned} \operatorname{cov}(X,Y) &= \mathbf{x}'\mathbf{D}\mathbf{y} \\ \sigma_X &= \|\mathbf{x}\|_{\mathbf{D}} \\ \operatorname{cor}(X,Y) &= \frac{\mathbf{x}'\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}}\,\|\mathbf{y}\|_{\mathbf{D}}}. \end{aligned}$$

La valeur absolue ou le carré du coefficient de corrélation définissent des indices de similarité entre deux variables quantitatives. Il est facile d'en déduire des distances. Le carré du coefficient de corrélation linéaire a la particularité d'induire une distance euclidienne :

$$d^2(X,Y) = 2(1 - \operatorname{cor}^2(X,Y)).$$

PROPOSITION 5.4. — *La distance entre variables quantitatives $d^2(X,Y)$ est encore le carré de la distance $\|\mathbf{P}_{\mathbf{x}} - \mathbf{P}_{\mathbf{y}}\|_{\mathbf{D}}$ entre les projecteurs $\mathbf{D}$-orthogonaux sur les directions engendrées par les vecteurs $\mathbf{x}$ et $\mathbf{y}$.*

Des indices de dissimilarité peuvent également être définis pour un couple de variables qualitatives (à partir de l'indice de Tschuprow) ou pour une variable quantitative et une variable qualitative (à parti du rapport de corrélation). Ils ont moins d'intérêt pour des données d'expression et sont laissés de côté.

3 Recherche d'une configuration de points

Le positionnement multidimensionnel est la recherche d'une configuration de points dans un espace euclidien qui admette $\boldsymbol{\mathcal{D}}$ comme matrice de distances si celle-ci est euclidienne ou, dans le cas contraire, qui en soit la meilleure approximation à un rang q fixé (en général 2) au sens d'une norme sur les matrices. Nous ne nous intéressons dans ce chapitre qu'à la version "métrique" du MDS, une autre approche "non métrique" construite sur les rangs est développée dans la bibliographie.

Ainsi posé, le problème admet une infinité de solutions. En effet, la distance entre deux vecteurs $\mathbf{x}_i$ et $\mathbf{x}_k$ d'une configuration est invariante par toute transformation affine $\mathbf{z}_i = \mathbf{F}\mathbf{x}_i + \mathbf{b}$ dans laquelle $\mathbf{F}$ est une matrice orthogonale quelconque et $\mathbf{b}$ un vecteur de $\mathbb{R}^p$. Une solution n'est donc connue qu'à une rotation et une translation près.

3.1 Propriétés

La solution est donnée par les résultats (Mardia et col.79) ci-dessous :

PROPOSITION 5.5. — *Soit $\mathcal{D}$ une matrice de distance et $\mathbf{B} = \mathbf{HAH}$ la matrice centrée en lignes et colonnes associée.*

- *Si $\mathcal{D}$ est la matrice de distance euclidienne d'une configuration $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ alors $\mathbf{B}$ est la matrice de terme général*

$$b_j^k = (\mathbf{x}_j - \overline{\mathbf{x}})'(\mathbf{x}_k - \overline{\mathbf{x}})$$

qui se met sous la forme

$$\mathbf{B} = (\mathbf{HX})(\mathbf{HX})'.$$

Elle est donc positive et appelée matrice des produits scalaires de la configuration centrée.

- *Réciproquement, si $\mathbf{B}$ est positive de rang p, une configuration de vecteurs admettant $\mathbf{B}$ pour matrice des produits scalaires est obtenue en considérant sa décomposition spectrale $\mathbf{B} = \mathbf{U\Delta U}'$. Ce sont les lignes de la matrice centrée $\mathbf{X} = \mathbf{U\Delta}^{1/2}$ qui fournissent les coordonnées des vecteurs de la représentation euclidienne.*

3.2 Explicitation du MDS

Pour résumé, dans le cas d'une matrice $\mathcal{D}$ euclidienne supposée de rang q, le MDS est obtenu en exécutant les étapes suivantes :

i. construction de la matrice $\mathbf{A}$ de terme général $-1/2{d_j^k}^2$,

ii. calcul de la matrice des produits scalaires par double centrage $\mathbf{B} = \mathbf{HAH}'$,

iii. diagonalisation de $\mathbf{B} = \mathbf{U\Delta U}'$;

iv. les coordonnées d'une configuration, appelées *coordonnées principales*, sont les lignes de la matrice $\mathbf{X} = \mathbf{U\Delta}^{1/2}$.

Dans le cas euclidien, ACP et MDS sont directement connectés.

PROPOSITION 5.6. — *Soit $\mathbf{Y}$ la matrice des données habituelles en ACP. L'ACP de ($\mathbf{Y}$, $\mathbf{M}$, $1/n\mathbf{I}$) fournit les mêmes représentations graphiques que le positionnement calculé à partir de la matrice de distances de terme général $\|\mathbf{y}_i - \mathbf{y}_j\|_{\mathbf{M}}$. Si $\mathbf{C}$ désigne la matrice des composantes principales, alors les coordonnées principales sont $\sqrt{n}\mathbf{C}$.*

L'intérêt du MDS apparaît évidemment lorsque les observations $\mathbf{Y}$ sont inconnues ou encore si l'on cherche la meilleure représentation euclidienne de distances non-euclidiennes entre les individus ; c'est l'objet du théorème suivant. En ce sens, le MDS "généralise" l'ACP et permet, par exemple, de considérer une distance de type robuste à base de valeurs absolues mais la représentation des variables pose alors quelques problèmes car le "biplot" n'est plus linéaire.

PROPOSITION 5.7. — *Si $\mathcal{D}$ est une matrice de distance, pas nécessairement euclidienne, $\mathbf{B}$ la matrice de produit scalaire associée, alors, pour une dimension q fixée, la configuration issue du MDS a une matrice de distance $\widehat{\mathcal{D}}$ qui rend $\sum_{j,k=1}^{n}(\{d_j^k\}^2 - \widehat{d_j^k}^2)$ minimum et, c'est équivalent, une matrice de produit scalaire $\widehat{\mathbf{B}}$ qui minimise $\left\|\mathbf{B} - \widehat{\mathbf{B}}\right\|^2$.*

4 Application au choix de variables

La sélection d'un sous-ensemble de variables pour la mise en œuvre de techniques factorielles (Jolliffe 2002) n'est pas aussi claire que dans le cadre de la recherche d'un modèle linéaire parcimonieux. Le problème vient souvent de la confusion de deux objectifs :

- supprimer des variables très liées, donc redondantes, et dont la multiplicité vient renforcer artificiellement l'influence de certains phénomènes,
- supprimer des variables afin de simplifier l'interprétation des axes tout en conservant au mieux les représentations graphiques.

Le premier objectif modifie donc les représentations en visant à être plus proche de la "réalité" ou au moins d'une réalité moins triviale tandis que, par principe, le deuxième objectif recherche le sous-ensemble restreint de variables susceptibles d'engendrer le même sous-espace de représentation.

Il n'existe pas de solution miracle, néanmoins les outils présentés dans ce chapitre : indices de similarité entre variable et positionnement multidimensionnel, peuvent aider à ces choix surtout lorsque l'analyse d'un grand nombre de variables nécessite de segmenter l'analyse en sous-groupes. Les algorithmes de classification (hiérarchique ou centres mobiles) appliqués sur les mêmes tableaux de distance apportent un éclairage complémentaire.

5 Données d'expression

Une analyse en composantes principales (cf. chapitre 3) fournit un premier aperçu de la représentation de gènes relativement aux échantillons biologiques par l'intermédiaire d'un biplot. Le but ici est de s'intéresser aux éventuelles co-régulations ou inhibitions entre gènes. Le cas échéant, ceux-ci apparaîtront corrélés positivement ou négativement.

Deux options sont possibles :

- utiliser une dissimilarité d'un moins la corrélation rend proches deux gènes co-régulés et éloigne deux gènes dont on peut considérer que l'un inhibe l'autre.
- utiliser un moins la corrélation au carré rapproche deux gènes liés qu'il y ait co-régulation ou inhibition.

En cas de problème de robustesse (valeurs atypiques) encore présent après transformation en logarithme, remplacer la corrélation linéaire de Pearson par celle sur les rangs de Spearman peut s'avérer utile.

Dans l'exemple des données d'obésité, plusieurs options sont possibles pour la représentation des gènes. La première utilise l'ACP dans laquelle les gènes sont les variables. La figue 5 montre une sorte d'effet taille. Les expressions de tous les gènes sont corrélées positivement avec une direction particulière sans soute associée à la "taille" des cellules des sujets.

Cette représentation n'est donc pas "optimale", influencée par un artefact relativement fort. Une autre approche est préférable. Le double centrage, en lignes et colonnes, implicitement contenu dans le MDS, élimine cet artefact.

Comme en ACP, un graphique représentant la décroissance des valeurs propres aide au choix de la dimension. Dans le cas de la matrice calculée avec les carrés des corrélations, deux dimensions, au moins dans une première approche, s'avèrent suffisantes (cf. figure 5-b). A la vue de ces graphique (figure 5 et 5), seul le biologiste peut juger de la pertinence des résultats ; il choisira, en connaissance de cause, le graphique le plus explicite.

D'autres représentations sont possibles sous forme d'arbres. C'est l'objet du chapitre suivant.

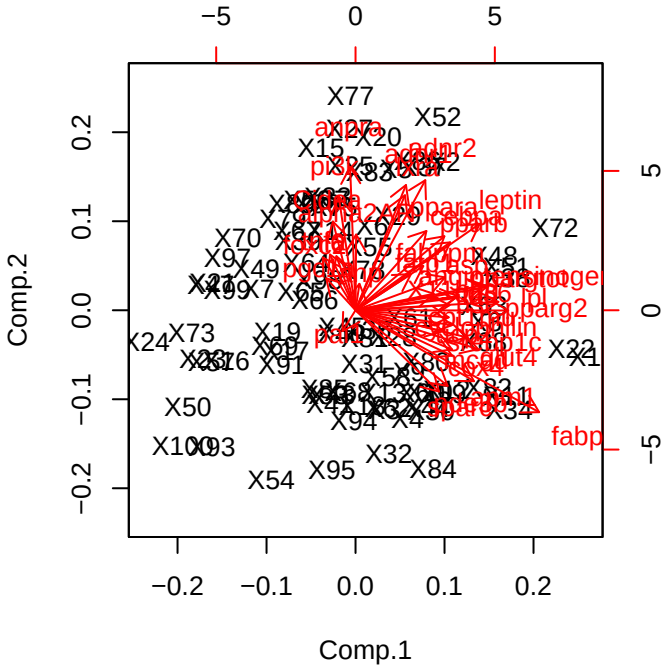


FIG. 5.2 – *Obésité : Représentation du premier plan de l'analyse en composantes principales faisant apparaître un certain "effet taille".*

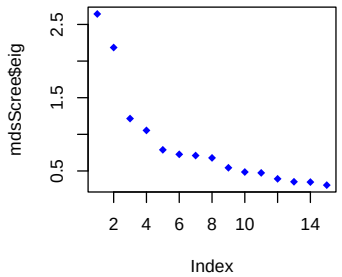


FIG. 5.3 – *Obésité : Décroissance des valeurs propres et choix de la dimension de l'espace dans lequel rechercher une représentation euclidienne des gènes la plus conforme à leurs distances au sens de la corrélation au carré.*

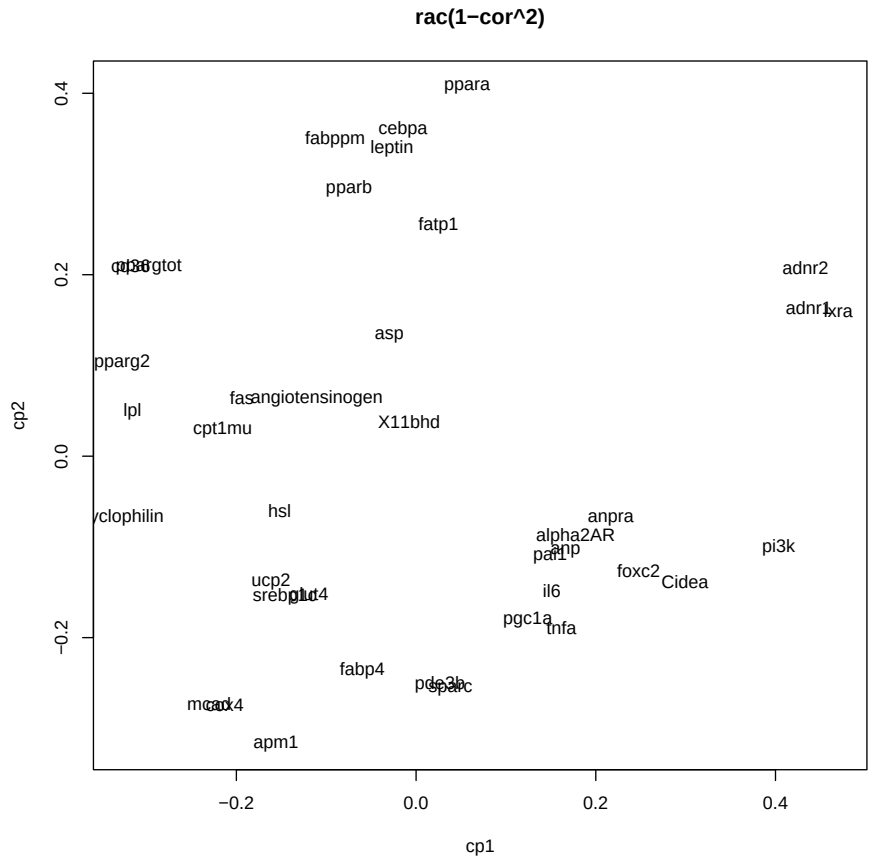


FIG. 5.4 – *Obésité : représentation des gènes en fonction de leur proximité au sens de la corrélation au carré.*

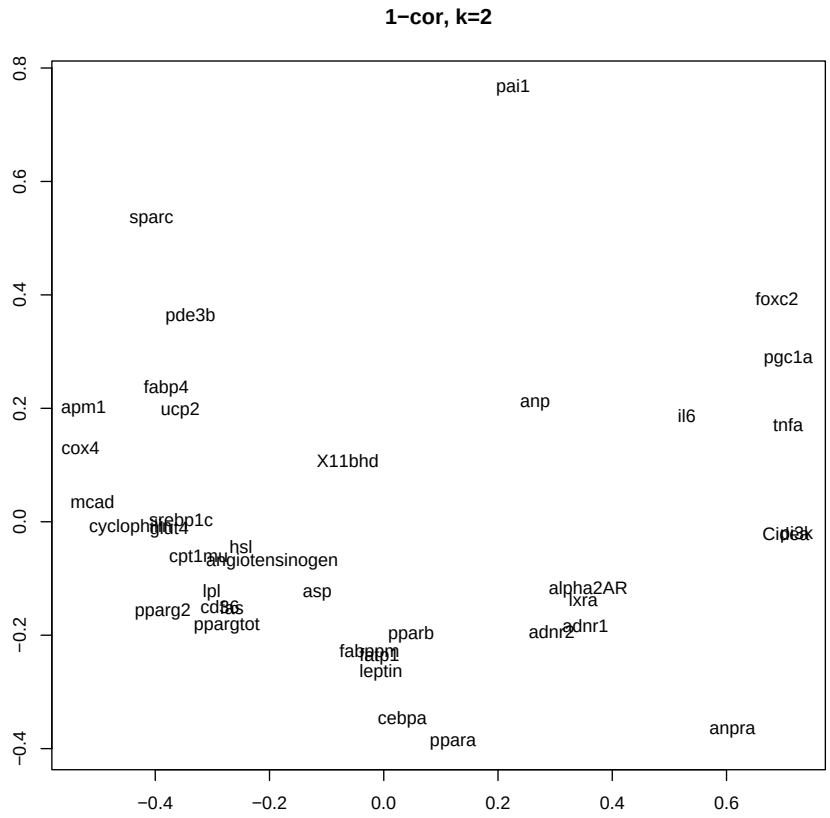


FIG. 5.5 – *Obésité : représentation des gènes en fonction de leur proximité au sens de la corrélation.*

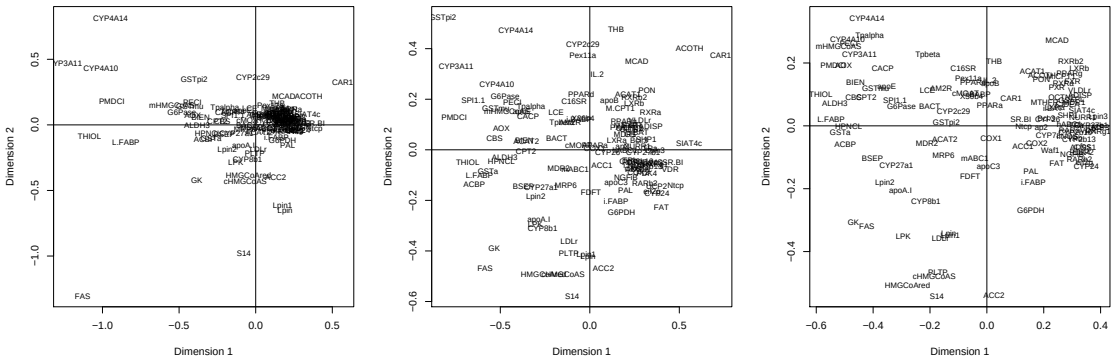


FIG. 5.6 – *Souris : positionnement multidimensionnel des gènes sur les axes 1 et 2 selon 3 distances différentes : distance euclidienne (d_1 à gauche), corrélation (d_3 au centre), corrélation carrée (d_2 à droite).*

6 Exemple : nutrition chez la souris

Appliqué à ces données, le positionnement multidimensionnel permet de considérer différentes façon de prendre en compte des distances inter-gènes :

- distance euclidienne, $d_1(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$, positive ou nulle ;
- distance associée à la corrélation carrée, $d_2(X,Y) = \sqrt{1 - \text{cor}(X,Y)^2}$, comprise entre 0 et 1 ;
- distance associée à la corrélation, $d_3(X,Y) = 1 - \text{cor}(X,Y)$, comprise entre 0 et 2.

Remarquons tout d'abord que dans les trois cas, plus la valeur est petite, plus les gènes dont on mesure l'éloignement sont proches. Ensuite, pour d_2 et d_3, une valeur proche de 1 caractérise deux gènes non corrélés, ce qui n'est pas nécessairement le cas de la distance euclidienne. Enfin, il est important de noter qu'une corrélation forte et négative entre deux gènes conduit à deux résultats opposés selon d_2 (valeur proche de 0) et d_3 (valeur proche de 2).

La figure 5.6 illustre les trois possibilités avec le positionnement multidimensionnel des gènes. L'analyse conjointe de ces trois graphiques conduit à de nombreuses interprétations sur le plan biologique. Sans rentrer dans les détails, nous noterons que ces trois graphiques tendent à séparer deux groupes de gènes qui interviennent dans deux fonctions biologiques opposées : les `CYP4A`, `PMDCI`, `PECI`, `AOX`, `BIEN`, `THIOL`, `CPT2`, `mHMGCoAS`, `Tpalpha` et `Tpbeta` sont impliqués dans le catabolisme des lipides et la cétogénèse alors que les gènes `FAS`, `S14`, `ACC2`, `cHMGCoAS`, `HMGCoAred` et, plus indirectement, `GK` et `LPK` sont impliqués dans la synthèse de lipides au niveau hépatique. On observera qu'aucun des trois graphiques de la figure 5.6, analysé individuellement, ne conduit à la totalité de cette interprétation mais que c'est bien l'analyse conjointe de ces représentations qui permet d'affiner la connaissance du biologiste sur ces données. Succintement, notons également que d'autres gènes tendent à participer à ces groupes. Par exemple, le gène `Lpin1` est proche des gènes impliqués dans la lipogénèse. Bien que sa fonction soit actuellement inconnue, Peterfy *et al.* (2001) ont observé que la lignée de souris déficiente pour `Lpin1` présente des altérations du métabolisme des lipides.

Les gènes dont la position sur le graphique sera le plus modifié en passant de la distance d_2 à la distance d_3 seront ceux présentant des corrélations négatives et importantes avec de nombreux

autres gènes. Un cas typique dans notre exemple est celui de CAR1 dont l'ACP (ainsi, que la matrice des corrélations) a montré qu'il était négativement corrélés avec des gènes tels que GSTpi2, CYP3A11, FAS... La position relative des couples de gènes ainsi obtenus change de façon importante entre les deux graphiques. On observera en particulier le couple CAR1−GSTpi2 totalement opposé sur l'axe 1 selon d_3 et relativement proche selon d_2 (tandis qu'il présente une opposition moins marquée selon d_1). La surexpression du gène CAR1 et la sous-expression du gène GSTpi2 chez les souris déficientes en récepteur PPARα n'a pas été décrite et constitue l'un des résultats originaux de ce travail. L'étude d'un lien potentiel entre ces deux modifications d'expression nécessitera la mise en œuvre d'expériences complémentaires.

D'une manière générale, on peut retenir que l'utilisation de la distance euclidienne tend à rapprocher des gènes dont les expressions sont proches. En revanche, les deux autres indicateurs considèrent que deux gènes sont proches si leur expression varie dans le même sens selon les conditions expérimentales. La corrélation (d_3) distingue les gènes corrélés négativement, ce que ne permet pas la corrélation carrée (d_2) qui doit donc être utilisée en connaissance de cause.

Notons que la distance d_1 est plus courante en statistique alors que d_3 l'est davantage dans les études relatives aux biopuces. Autant que possible une comparaison des trois distances est recommandée. On se référera à Draghici (2003, chapitre 11) pour une discussion plus détaillée sur le sujet.

Chapitre 6

Classification

1 Introduction

1.1 Les données

Comme dans le cas du chapitre précédent (MDS), les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés, pour simplifier, du même poids :

- un tableau de distances (ou dissimilarités, ou mesures de dissemblance), $n \times n$, entre les individus pris deux à deux ;
- les observations de p variables quantitatives sur ces n individus ;
- les observations, toujours sur ces n individus, de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

D'une façon ou d'une autre, il s'agit, dans chaque cas, de se ramener au tableau des distances deux à deux entre les individus (c'est-à-dire au premier cas). Le choix d'une matrice de produit scalaire permet de prendre en compte simplement un ensemble de variables quantitatives tandis que le troisième cas nécessite plus de développements ; il n'est pas présenté ici car de peu d'intérêt pour des données d'expression.

1.2 Les objectifs

L'objectif d'une méthode de classification déborde le cadre strictement exploratoire. C'est la recherche d'une *typologie*, ou *segmentation*, c'est-à-dire d'une partition, ou répartition des individus en *classes*, ou catégories. Ceci est fait en optimisant un *critère* visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement (en anglais *classification*) pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage. Nous sommes dans une situation d'apprentissage *non-supervisé*, ou en anglais de *clustering*[1].

1.3 Les méthodes

Un calcul élémentaire de combinatoire montre que le nombre de partitions possibles d'un ensemble de n éléments croît plus qu'exponentiellement avec n. Ainsi, pour $n = 20$, il est de l'ordre de 10^{13}. Il n'est donc pas question de chercher à optimiser le critère sur toutes les partitions

[1] Faire attention aux faux amis français / anglais : discrimination / classification (supervisée) et classification / clustering (non-supervisée)

possibles. Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une "bonne" partition qui correspond en général à un optimum local. Même si le besoin de classer des objets est très ancien, seule la généralisation des outils informatiques en a permis l'automatisation dans les années 1970. Celeux et col. (1989) décrivent en détail ces algorithmes.

Différents choix sont laissés à l'initiative de l'utilisateur :

- une mesure d'éloignement (dissemblance, dissimilarité ou distance) entre individus ;
- le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir de la trace d'une matrice de variances-covariances ; soit les variances et covariances interclasses (la trace correspond alors à l'inertie de la partition), soit les variances et covariances intraclasse ;
- la méthode : la classification ascendante hiérarchique et celle par réallocation dynamique sont les plus utilisées, seules ou combinées ;
- le nombre de classes : c'est un point délicat.

Enfin, différents outils recherchent une interprétation, ou des caractérisations, des classes obtenues.

On notera que les principes algorithmiques de ces méthodes sont relativement élémentaires.

Classification ascendante hiérarchique, ou CAH

Il s'agit de regrouper itérativement les individus, en commençant par le bas (les deux plus proches) et en construisant progressivement un arbre, ou *dendrogramme*, regroupant finalement tous les individus en une seule classe, à la racine (cf. figure 2 qui reprend les données élémentaires du chapitre précédent). Ceci suppose de savoir calculer, à chaque étape ou regroupement, la distance entre un individu et un groupe ainsi que celle entre deux groupes. Ceci nécessite donc, pour l'utilisateur de cette méthode, de faire un choix supplémentaire : comment définir la distance entre deux groupes connaissant celles de tous les couples d'individus entre ces deux groupes. Différents choix, appelés *saut* en français et *linkage* en anglais, sont détaillés plus loin. Le nombre de classes est déterminé *a posteriori*, à la vue du dendrogramme ou d'un graphique représentant la décroissance de la hauteur de chaque saut, ou écart de distance, opéré à chaque regroupement.

Classification par réallocation dynamique

Dans ce cas, le nombre de classes, k, est fixé *a priori*. Ayant initialisé k centres de classes par tirage aléatoire, tous les individus sont affectés à la classe dont le centre est le plus proche au sens de la distance choisie (en principe, euclidienne pour cette méthode). Dans une deuxième étape, l'algorithme calcule des barycentres de ces classes qui deviennent les nouveaux centres. Le procédé (affectation de chaque individu à un centre, détermination des centres) est itéré jusqu'à convergence vers un minimum (local) ou un nombre d'itérations maximum fixé.

Classification mixte

La CAH nécessite impérativement la construction d'un tableau de distances $n \times n$ et son stockage en mémoire ; le nombre maximum d'individus traités peut s'en trouver limité. Ce n'est pas le cas dans l'algorithme de réallocation, d'où l'intérêt possible d'une approche mixte pour, à la fois, classer de grands volumes de données et sélectionner le nombre de classes par CAH.

Dans le cas plus spécifique de données d'expression, et comme pour le chapitre précédent (MDS), le choix principal est celui de la distance (ou dissimilarité) utilisée. S'ajoute en plus le choix du critère de saut en CAH et celui du nombre de classes (*a priori* avec la réallocation dynamique, ou *a posteriori* avec la CAH). La plupart des logiciels dédiés à ces données proposent une

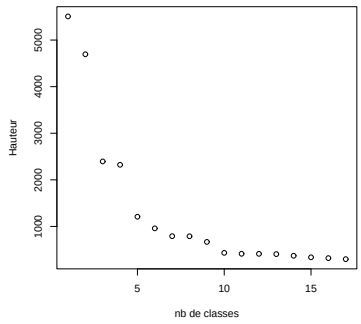


FIG. 6.1 – *Villes : Décroissance de la variance interclasses à chaque regroupement dans le cas du saut de Ward.*

double CAH des lignes (gènes) et des colonnes (échantillons biologiques) dans une représentation graphique habilement colorée.

2 Illustration

En guise de première illustration sur les méthodes de classification, nous reprenons l'étude des mêmes données que dans le chapitre précédent : un tableau contenant les distances kilométriques par route (Source : IGN) entre 47 grandes villes en France et dans les pays limitrophes. Toutes ces valeurs sont rangées dans le triangle inférieur d'une matrice carrée avec des 0 sur la diagonale. Il s'agit donc de regrouper au mieux ces villes, en tenant compte de leurs proximités relatives au sens de cette distance routière.

À l'issue de l'exécution, la classification ascendante hiérarchique fournit les deux graphiques précisés ci-dessous.

- Un graphique d'aide au choix du nombre de classes (cf. figure 2). Il représente à rebours, en fonction du nombre de classes, la décroissance de la distance interclasses. La présence d'une rupture importante dans cette décroissance aide au choix du nombre de classes comme dans le cas du choix de dimension en ACP, avec l'éboulis des valeurs propres. Dans ce cas, il faut lire le graphe de droite à gauche et s'arrêter avant le premier saut jugé significatif. Avec l'indice de Ward, cela revient à couper l'arbre avant une perte, jugée trop importante, de la variance interclasses. Dans le cas des villes repérées par leurs distances kilométriques, le choix de 5 classes semble raisonnable.
- Le *dendrogramme* (cf. figure 2) est une représentation graphique, sous forme d'arbre binaire, des agrégations successives jusqu'à la réunion en une seule classe de tous les individus. La hauteur d'une branche est proportionnelle à l'indice de dissemblance ou distance entre les deux objets regroupés. Dans le cas du saut de Ward, c'est la perte de variance interclasses.

Une fois un nombre de classes sélectionné à l'aide du premier graphique, une coupure de l'arbre fournit, dans chaque sous-arbre, la répartition des individus en classes. Ces classes peuvent ensuite être représentées dans les axes d'une analyse factorielle, en général une ACP ou un MDS (figure 2).

Signalons qu'il est courant, dans la pratique, de mettre en œuvre, à l'issue d'une CAH, une méthode de réallocation dynamique avec pour nombre de classes celui choisi par CAH et pour

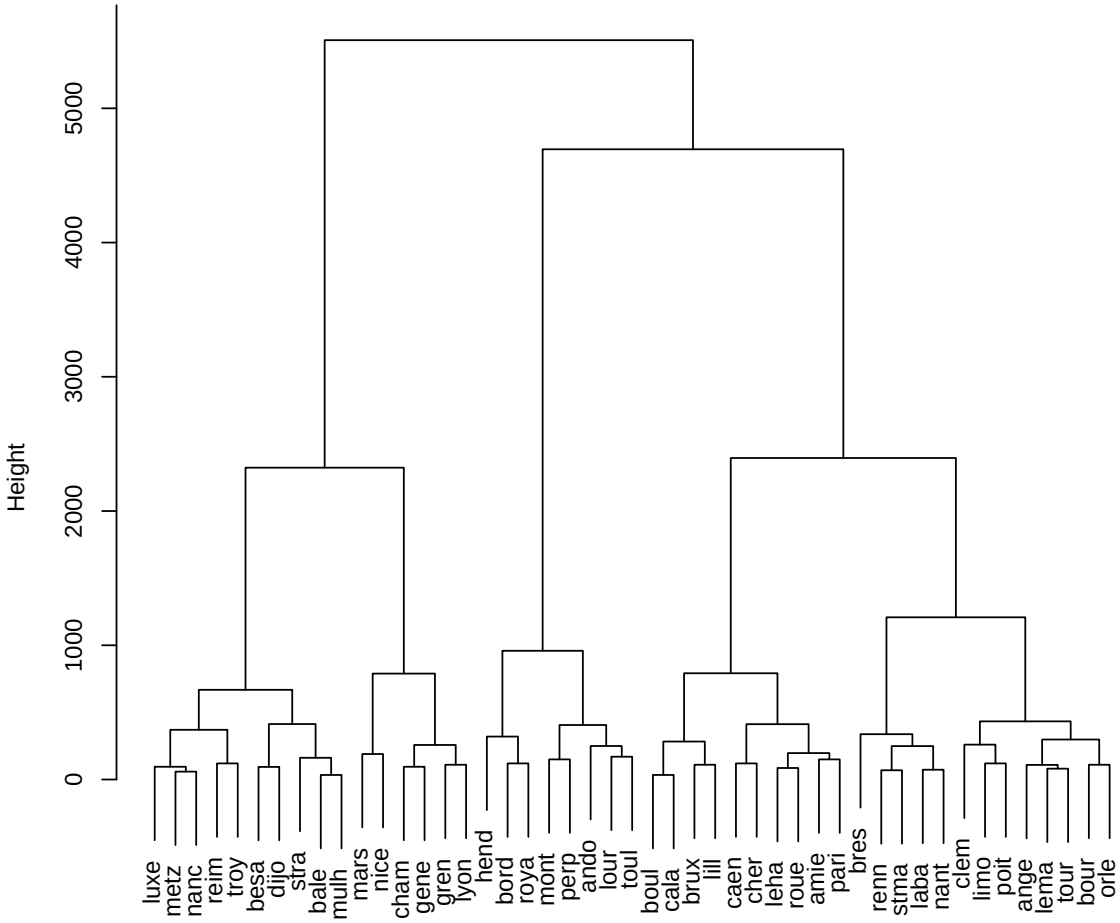


FIG. 6.2 – *Villes : Exemple d'un dendrogramme issu de la classification des données par CAH et saut de Ward.*

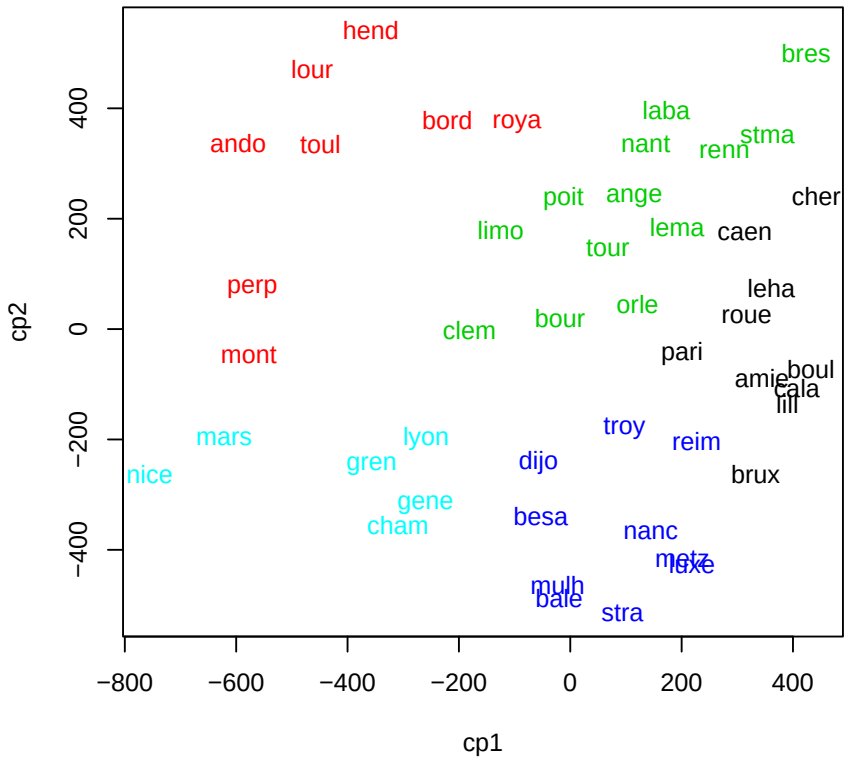


FIG. 6.3 – *Villes : Représentation des classes (couleurs) obtenues par CAH dans les coordonnées du MDS.*

centres initiaux les barycentres des classes obtenues : on stabilise ainsi les classes.

Notons également que l'exemple présenté ici est relativement simple et bien structuré. Modifier le critère de saut ne change pas grand chose dans ce cas. Mais, attention, il est facile de vérifier expérimentalement qu'une classification ascendante est un objet très sensible. En effet, il suffit de modifier une distance dans le tableau, par exemple de réduire sensiblement la distance de Grenoble à Brest, pour que la classification (nombre de classes, organisation) devienne très sensible au choix du critère de saut. En revanche, la structure des données fait que la représentation factorielle de l'ACP du tableau de distance (MDS) est très robuste à ce type d'"erreur de mesure".

3 Mesures d'éloignement

Notons $\Omega = \{i = 1, \ldots, n\}$ l'ensemble des individus. Cette section se propose de définir sur $\Omega \times \Omega$ différentes mesures d'éloignement entre deux individus. Les hypothèses et propriétés étant de plus en plus fortes.

3.1 Indice de ressemblance, ou similarité

C'est une mesure de proximité définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ et vérifiant :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie ;} \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \ \text{ressemblance d'un individu avec lui-même ;} \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \text{la ressemblance est majorée par } S. \end{aligned}$$

Un indice de ressemblance normé s^* est facilement défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega \, ;$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$.

3.2 Indice de dissemblance, ou dissimilarité

Une dissimilarité est une application d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant :

$$\begin{aligned} d(i,j) &= d(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie ;} \\ d(i,i) &= 0, \forall i \in \Omega \ : \ \text{nullité de la dissemblance d'un individu avec lui-même.} \end{aligned}$$

Les notions de similarité et dissimilarité se correspondent de façon élémentaire. Si s est un indice de ressemblance, alors

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

est un indice de dissemblance. De façon réciproque, si d est un indice de dissemblance avec $D = \sup_{(i,j) \in \Omega \times \Omega} d(i,j)$, alors $s(i,j) = D - d(i,j)$ est un indice de ressemblance. Comme s^*, un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$. Du fait de cette correspondance immédiate, seule la notion de dissemblance, ou dissimilarité, normée est considérée par la suite.

3.3 Indice de distance

Un indice de distance est, par définition, un indice de dissemblance qui vérifie de plus la propriété :

$$d(i,j) = 0 \Longrightarrow i = j.$$

Cette propriété évite des incohérences pouvant apparaître entre dissemblances, par exemple :

$$\exists\, k \in \Omega : d(i,k) \neq d(j,k), \quad \text{avec pourtant } i \neq j \text{ et } d(i,j) = 0.$$

3.4 Distance

Une distance sur Ω est, par définition, un indice de distance vérifiant en plus la propriété d'*inégalité triangulaire*. Autrement dit, une distance d est une application de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant :

$$\begin{aligned} d(i,j) &= d(j,i), \ \forall (i,j) \in \Omega \times \Omega\ ; \\ d(i,i) &= 0 \Longleftrightarrow i = j\ ; \\ d(i,j) &\leq d(i,k) + d(j,k), \ \forall (i,j,k) \in \Omega^3. \end{aligned}$$

Si Ω est fini, la distance peut être normée.

3.5 Distance euclidienne

Dans le cas où Ω est un espace vectoriel muni d'un produit scalaire, donc d'une norme, la distance définie à partir de cette norme est appelée distance euclidienne :

$$d(i,j) = < i-j, i-j >^{1/2} = \|i-j\|.$$

La condition pour qu'une matrice donnée de distances entre éléments d'un espace vectoriel soit issue d'une distance euclidienne est explicitée dans le chapitre précédent. Toute distance n'est pas nécessairement euclidienne ; voir, par exemple, celle construite sur la valeur absolue.

3.6 Utilisation pratique

Concrètement, il peut arriver que les données à traiter soient directement sous la forme d'une matrice d'un indice de ressemblance ou de dissemblance. Il est alors facile de la transformer en une matrice de dissemblances normées avant d'aborder une classification.

Nous précisons ci-dessous les autres cas.

Données quantitatives

Lorsque les p variables sont toutes quantitatives, il est nécessaire de définir une matrice $\mathbf{M}$ de produit scalaire sur l'espace $\mathbb{R}^P$. Le choix $\mathbf{M} = \mathbf{I}_p$, matrice identité, est un choix élémentaire et courant ; mais il est vivement conseillé de *réduire* les variables de variances hétérogènes, comme en ACP, ce qui revient à considérer, comme matrice de produit scalaire, la matrice diagonale composée des inverses des écarts-types :

$$\mathbf{M} = \Sigma^{-1} = \text{diag}\,(\frac{1}{\sigma_1} \cdots \frac{1}{\sigma_p}).$$

La métrique dite de Mahalanobis (inverse de la matrice des variances-covariances) peut aussi être utilisée pour atténuer la structure de corrélation.

Données qualitatives

Dans le cas très particulier où toutes les variables sont binaires (présence ou absence de caractéristiques), de nombreux indices de ressemblances ont été proposés dans la littérature. Ils ne sont pas détaillés dans le cadre d'un cours spécifique aux données d'expression.

3.7 Bilan

Une fois ces préliminaires accomplis, nous nous retrouvons donc avec
- soit un tableau de mesures quantitatives $n \times p$, associé à une matrice de produit scalaire $p \times p$ (en général $\mathbf{I}_p$) définissant une métrique euclidienne,
- soit directement un tableau $n \times n$ de dissemblances ou de distances entre individus.

Attention, si n est grand, la deuxième solution peut se heurter rapidement à des problèmes de stockage en mémoire pour l'exécution des algorithmes.

4 Classification ascendante hiérarchique

4.1 Principe

L'initialisation de cet algorithme consiste, s'il n'est déjà donné, à calculer un tableau de distances (ou de dissemblances) entre les individus à classer. L'algorithme démarre alors de la partition triviale des n singletons (chaque individu constitue une classe) et cherche, à chaque étape, à constituer des classes par agrégation des deux éléments les plus proches de la partition de l'étape précédente. L'algorithme s'arrête avec l'obtention d'une seule classe. Les regroupements successifs sont représentés sous la forme d'un arbre binaire ou *dendrogramme*.

4.2 Distance, ou dissemblance, entre deux classes

À chaque étape de l'algorithme, il est nécessaire de mettre à jour le tableau des distances (ou des dissemblances). Après chaque regroupement, de deux individus, de deux classes ou d'un individu à une classe, les distances entre ce nouvel objet et les autres sont calculées et viennent remplacer, dans la matrice, les distances des objets qui viennent d'être agrégés. Différentes approches sont possibles à ce niveau, donnant lieu à différentes CAH.

Notons A et B deux classes, ou éléments, d'une partition donnée, w_A et w_B leurs pondérations, et $d_{i,j}$ la distance entre deux individus quelconques i et j.

Le problème est de définir $d(A, B)$, distance entre deux éléments d'une partition de Ω.

Cas d'une dissemblance

Les stratégies ci-dessous s'accomodent d'un simple indice de dissemblance défini entre les individus. Elles s'appliquent également à des indices plus structurés (distance) mais n'en utilisent pas toutesles propriétés.

$$
\begin{aligned}
d(A,B) &= \min_{i\in A, j\in B}(d_{ij}) \quad \text{(saut minimum, } \textit{single linkage}\text{)},\\
d(A,B) &= \sup_{i\in A, j\in B}(d_{ij}) \quad \text{(saut maximum ou diamètre, } \textit{complete linkage}\text{)},\\
d(A,B) &= \frac{1}{\text{card}(A)\text{card}(B)} \sum_{i\in A, j\in B} d_{ij} \quad \text{(saut moyen, } \textit{group average linkage}\text{)}.
\end{aligned}
$$

Cas d'une distance euclidienne

Les stratégies suivantes nécessitent la connaissance de représentations euclidiennes des individus : matrice $n \times p$ des individus afin, au minimum, de pouvoir définir les barycentres notés g_A et g_B des classes.

$$\begin{aligned} d(A,B) &= d(g_A, g_B) \quad \text{(distance des barycentres, centroïd)}, \\ d(A,B) &= \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{(saut de Ward)}. \end{aligned}$$

Important

Le saut de Ward joue un rôle particulier et est la stratégie la plus courante ; c'est même l'option par défaut (SAS) dans le cas d'une distance euclidienne entre individus. En effet, ce critère induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse.

4.3 Algorithme

ALGORITHME 6.1 *:*

classification ascendante hiérarchique

- `Initialisation` *Les classes initiales sont les singletons. Calculer la matrice de leurs distances deux à deux.*
- `Itérer` *les deux étapes suivantes jusqu'à l'agrégation en une seule classe :*
  - *i. regrouper les deux classes les plus proches au sens de la "distance" entre classes choisie,*
  - *ii. mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa "distance" avec chacune des autres classes.*

4.4 Graphes

Les graphes obtenus à l'issue d'une CAH ont été présentés et illustrés dans le paragraphe 2. Il s'agit du graphique d'aide au choix du nombre de classes et du dendrogramme.

5 Agrégation autour de centres mobiles

5.1 Principes

Différents types d'algorithmes ont été définis autour du même principe de *réallocation dynamique* des individus à des centres de classes, eux-mêmes recalculés à chaque itération. Ces algorithmes requièrent une représentation vectorielle des individus dans $\mathbb{R}^p$ muni d'une métrique, généralement euclidienne. Une adaptation de cet algorithme, PAM (pour *Partitioning — clustering — of the data into k clusters Around Medoïds* ; Kaufman & Rousseeuw, 1990), en est une version robuste, également adaptée à une matrice de dissimilarités. Ce dernier algorithme est en revanche limité au niveau du nombre d'observations (200).

Il est important de noter que, contrairement à la méthode hiérarchique précédente, le nombre de classes k doit être déterminé *a priori*.

Ces méthodes sont itératives : après une initialisation des centres consistant, le plus souvent, à tirer aléatoirement k individus, l'algorithme répète deux opérations jusqu'à la convergence d'un critère :

i. Chaque individu est affecté à la *classe* dont le centre est le plus proche.

ii. Calcul des k *centres* des classes ainsi constituées.

5.2 Principale méthode

Il s'agit de la méthode (*kmeans*) proposée dans Forgy (1965).

ALGORITHME 6.2 *:*

- `Initialisation` *Tirer au hasard, ou sélectionner pour des raisons extérieures à la méthode, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux.*
- `Itérer` *les deux étapes suivantes, jusqu'à ce que le critère de variance interclasses ne croisse plus de manière significative, c'est-à-dire jusqu'à la stabilisation des classes.*
  - *i. Allouer chaque individu au centre (c'est-à-dire à la classe) le plus proche au sens de la métrique euclidienne choisie ; on obtient ainsi, à chaque étape, une classification en k classes, ou moins si, finalement, une des classes devient vide.*
  - *ii. Calculer le centre de gravité de chaque classe : il devient le nouveau noyau ; si une classe s'est vidée, on peut éventuellement retirer aléatoirement un noyau complémentaire.*

5.3 Propriétés

Convergence Le critère (la variance interclasses) est majoré par la variance totale. Il est simple de montrer qu'il ne peut que croître à chaque étape de l'algorithme, ce qui en assure la convergence. Il est équivalent de maximiser la variance interclasses ou de minimiser la variance intraclasse. Cette dernière est alors décroissante et minorée par 0. Concrètement, une dizaine d'itérations suffit généralement pour atteindre la convergence.

Optimum local La solution obtenue est un optimum local, c'est-à-dire que la répartition en classes dépend du choix initial des noyaux. Plusieurs exécutions de l'algorithme permettent de s'assurer de la présence de *formes fortes*, c'est-à-dire de classes, ou partie de classes, présentes de manière stable dans la majorité des partitions obtenues.

5.4 Variantes

Algorithme kmeans

Il s'agit d'une modification de l'algorithme précédent, proposée par Mac Queen (1967). Les noyaux des classes, ici les barycentres des classes concernées, sont recalculés à chaque allocation d'un individu à une classe. L'algorithme est ainsi plus efficace, mais il dépend de l'odre des individus dans le fichier.

Nuées dynamiques

La variante proposée par Diday (1971) consiste à remplacer chaque centre de classe par un noyau constitué d'éléments représentatifs de cette classe. Cela permet de corriger l'influence d'éventuelles valeurs extrêmes sur le calcul du barycentre.

Partitionning Around Medoïds

Cet algorithme, proposé par Kaufman & Rousseeuw (1990), permet de classifier des données de façon plus robuste, c'est-à-dire moins sensible à des valeurs atypiques. Il permet également de

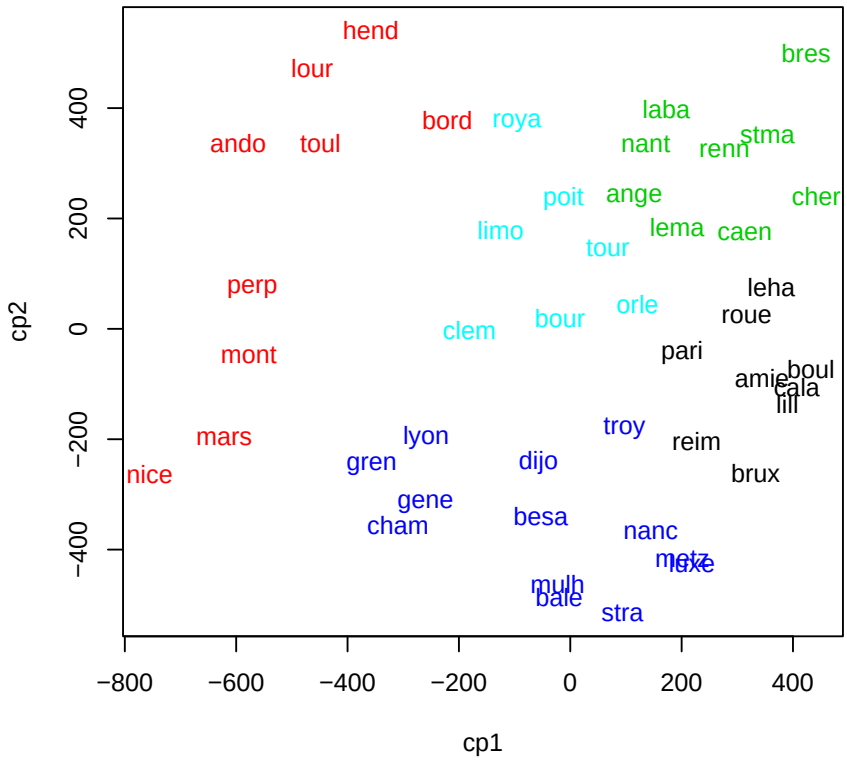


FIG. 6.4 – *Villes : Représentation des classes (couleurs) obtenues par PAM dans les coordonnées du MDS.*

traiter des matrices de dissimilarités. Les résultats sont fournis dans la figure 5.4, pour lequels le nombre de classe est fixé *a priori* à 5, comme le suggère la CAH, mais pour lesquels les classes obtenues sont sensiblement différentes.

5.5 Combinaison

Chaque méthode précédente peut être plus ou moins adaptée à la situation rencontrée. La classification hiérarchique, qui construit nécessairement la matrice des distances, n'accepte qu'un nombre limité d'individus ; de son côté, la réallocation dynamique nécessite de fixer *a priori* le nombre de classes. La stratégie suivante, adaptée aux grands ensembles de données, permet de contourner ces difficultés.

i. Exécuter une méthode de réallocation dynamique en demandant un grand nombre de classes, de l'ordre de 10% de n.

ii. Sur les barycentres des classes précédentes, exécuter une classification hiérarchique puis déterminer un nombre "optimal" k de classes.

iii. Exécuter une méthode de réallocation dynamique sur tout l'ensemble en fixant à k le nombre de classes. Pour initialiser l'algorithme, il est habituel de choisir pour noyaux les barycentres (calculés en pondérant par les effectifs de classes) des classes de l'étape précédente.

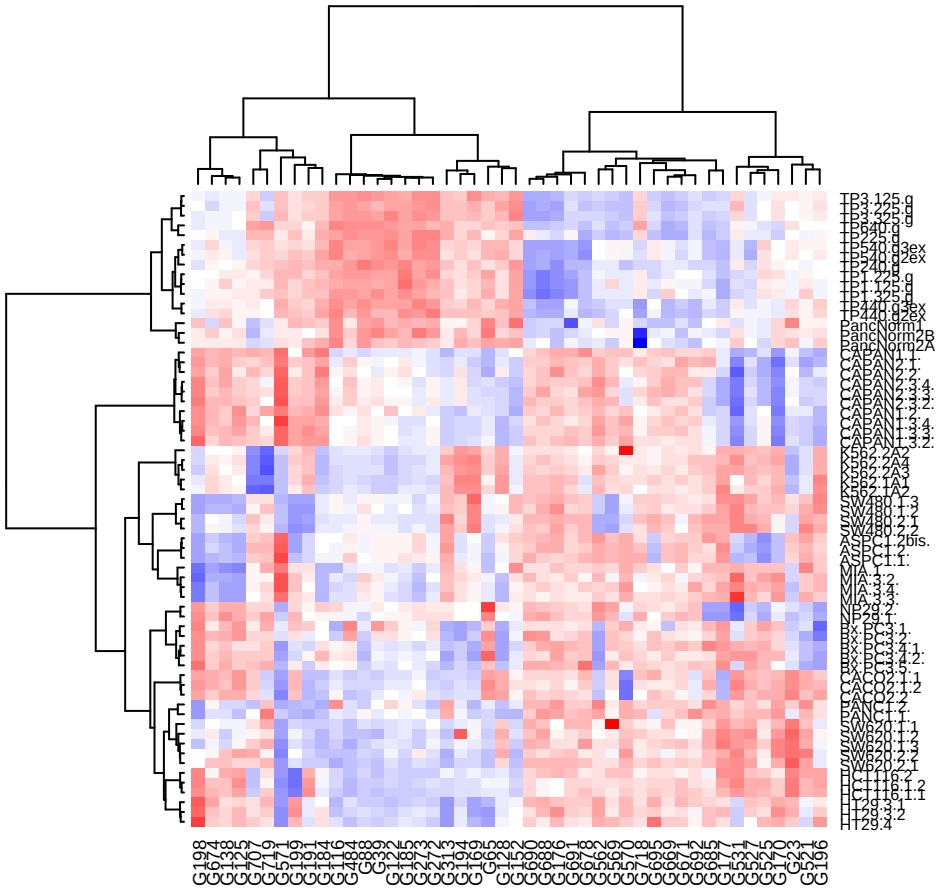


FIG. 6.5 – *Pancréas : Double classification ascendante hiérarchique des gènes et échantillons biologiques avec saut de Ward. La représentation utilise des fausses couleurs pour visualiser les proximités.*

6 Données d'expression

Pour ce type de données, les biologistes apprécient particulièrement de construire une double classification hiérarchique opérant à la fois sur les lignes et sur les colonnes. Une représentation en fausses couleurs fournit une lecture susceptible de prendre en compte les "distances" respectives des lignes (gènes) d'une part et des colonnes (échantillons biologiques) d'autre part, et de se faire ainsi une idée des gènes pouvant influencer la hiérarchie obtenue pour les échantillons. Néanmoins, cette lecture, même en se limitant à une sélection des gènes proposés par l'analyse en composantes principales (chapitre 3), n'est pas très aisée (figure 6).

Le choix de la distance est prépondérant dans les résultats d'une classification. Les figure 6 et 6 fournissent les dendrogrammes de la CAH dans le cas d'une dissimilarité calculée à partir de la corrélation et dans celui d'une distance basée sur la corrélation au carré. Comme pour le MDS (chapitre précédent), c'est au biologiste de choisir la ou les représentations aidant au mieux sa compréhension des régulations et/ou inhibitions entre gènes.

Comme pour les données considérant les distances entre villes, il serait facile de coupler pour

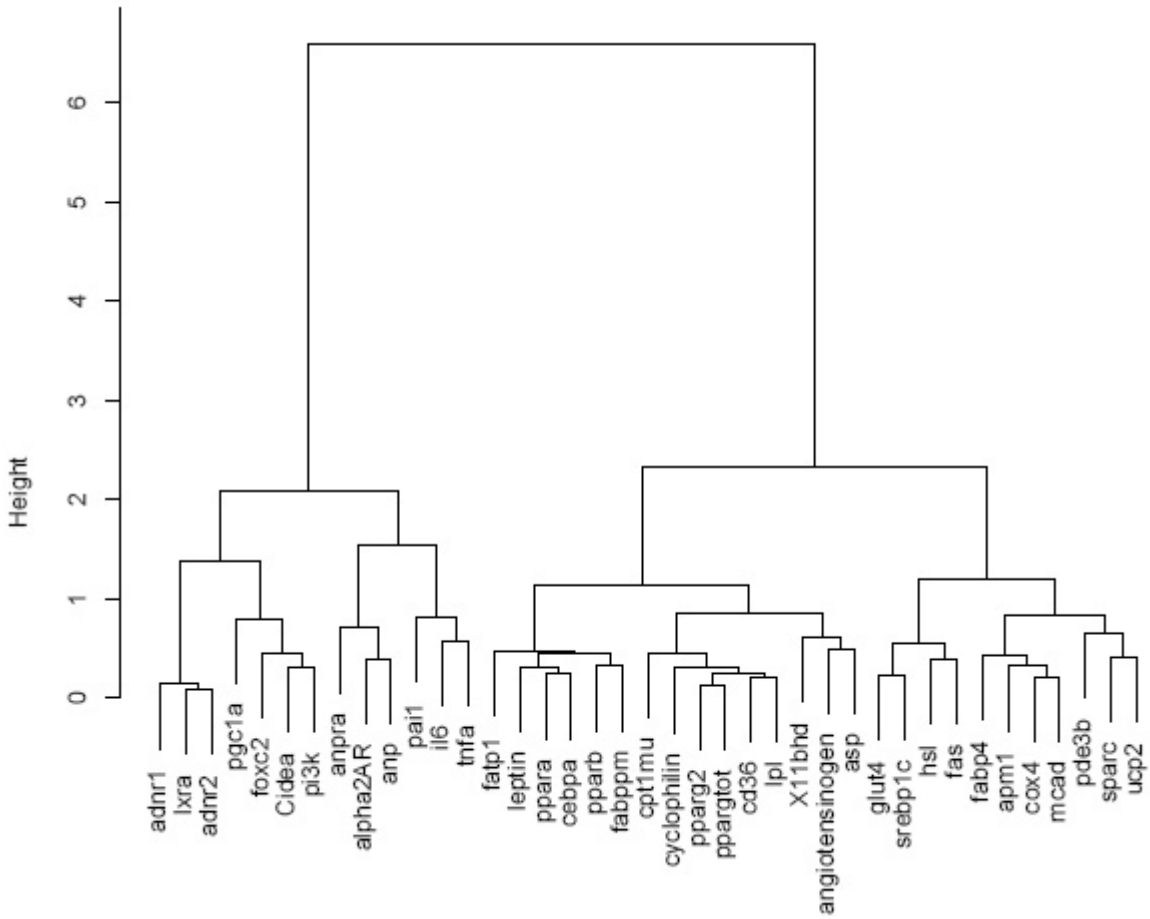


FIG. 6.6 – *Obésité : Classification ascendante hiérarchique des gènes avec saut de Ward considérant la corrélation.*

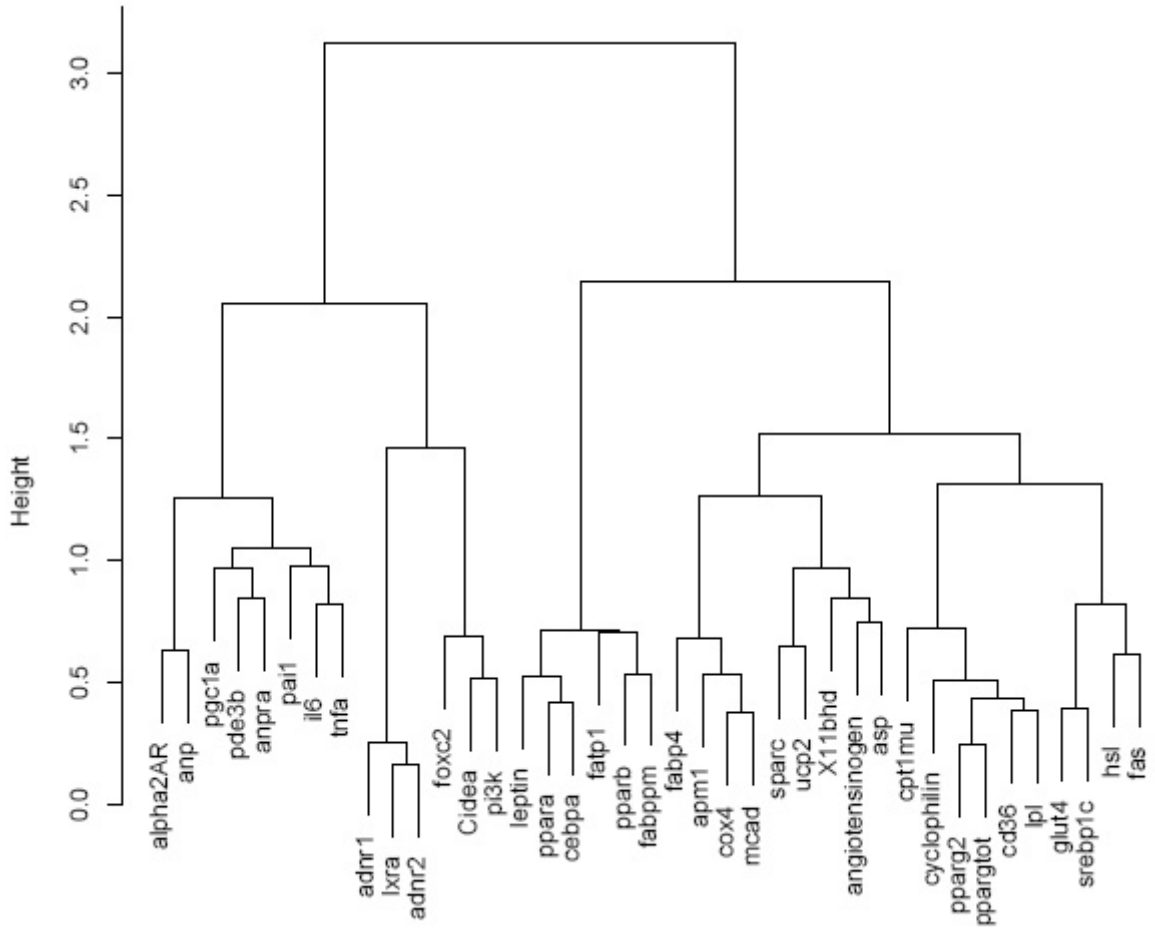


FIG. 6.7 – *Obésité : Classification ascendante hiérarchique des gènes avec saut de Ward considérant le carré de la corrélation.*

les données d'expression une représentation des classes par des couleurs dans le graphe du MDS, ou encore de celui d'une ACP. Nous laissons au lecteur l'appréciation sur le nombre de combinaisons d'options possibles (centrage, réduction, distance, critère de saut, projection, classification) qui sont offertes par l'ensemble de ces outils.

7 Exemple : nutrition chez la souris

Pour ce type de données, les biologistes apprécient particulièrement de construire une double classification hiérarchique opérant à la fois sur les lignes et sur les colonnes (gènes et échantillons). Une représentation en fausses couleurs fournit une lecture susceptible de prendre en compte les "distances" respectives des lignes (gènes) d'une part et des colonnes (échantillons biologiques) d'autre part, et de se faire ainsi une idée des gènes pouvant influencer la hiérarchie obtenue pour les échantillons. Néanmoins, cette lecture, même en se limitant à une sélection des gènes proposés par l'analyse en composantes principales (chapitre 3), n'est pas très aisée (figure 6).

Le choix de la distance est évidemment important. La plus fréquemment rencontrée pour l'étude du transcriptome est du type de d_3, basée sur la corrélation. Il nous semble pertinent d'utiliser les trois types de distances et d'en apprécier leur complémentarité quant à l'interprétation des résultats. Nous avons fait le choix de limiter cette comparaison des distances au MDS et nous nous contenterons ici de présenter une classification basée sur la distance euclidienne d_1. Le deuxième choix intervenant en classification concerne le critère d'agglomération, c'est-à-dire la façon dont est définie la distance entre deux groupes, et n'a pas d'interprétation biologique simple. Ce choix a plus une implication géométrique, sur la forme des classes obtenues. Nous avons utilisé le critère de Ward parce qu'il favorise la construction de classes relativement "sphériques" et qu'on peut lui associer des critères guidant la détermination du nombre de classes.

L'interprétation de la double classification (Fig. 6.8) présente des analogies avec celle de l'ACP sur le premier plan principal. Si l'on s'intéresse aux individus-souris, on peut constater que les deux génotypes sont différenciés en deux groupes, à l'exception de trois souris de type `PPAR` ayant suivi les régimes `efad` (pour deux d'entre elles) et `ref`. Ce sont ces trois mêmes individus que l'on retrouve projetés dans la partie négative du premier axe de l'ACP (Fig. 3.15). Pour les variables-gènes, on peut distinguer deux grandes classes correspondant, d'après les données, à deux niveaux d'expressions : à gauche, les gènes dont l'expression est relativement faible, à droite les gènes dont l'expression est globalement plus élevée. Dans cette seconde classe, un groupe attire particulièrement l'attention sur l'image : sur une bande verticale correspondant à 14 gènes, les couleurs sont nettement plus variables que sur le reste de l'image. Il s'agit des gènes

```
CYP4A10, CYP4A14, CYP3A11, L.FABP, THIOL, PMDCI, S14,
  Lpin1, Lpin, FAS, GSTmu, GSTpi2, CYP2c29, G6Pase
```

qui apparaissent tous parmi les gènes les plus corrélés aux deux premiers axes principaux de l'ACP (Fig. 3.15).

MDS et classification apparaissent donc comme des techniques complémentaires, mais elles ne sont pas sensibles de la même façon aux perturbations. La perturbation d'une donnée peut fortement influencer la structure d'un dendrogramme alors qu'en MDS, la prise en compte conjointe de toutes les distances deux à deux assure une certaine robustesse pour le calcul des coordonnées principales. Pour cette raison, il est utile de représenter les classes dans une projection sur des axes factoriels obtenus soit par MDS soit par ACP. L'éboulis des valeurs propres (Fig. 6.9) nous oriente vers une représentation du MDS en deux dimensions.

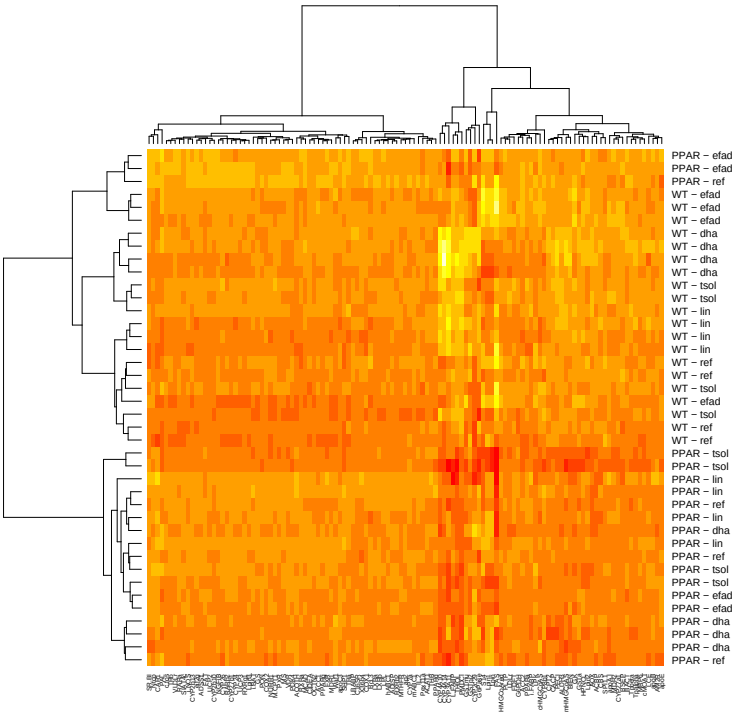


FIG. 6.8 – *Souris : double classification ascendante hiérarchique des individus-souris et des variables-gènes selon la méthode de Ward, avec la distance euclidienne.*

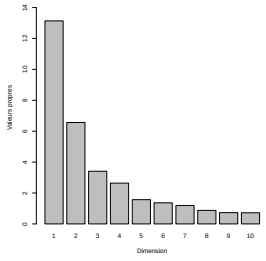


FIG. 6.9 – *Souris : éboulis des valeurs propres pour le MDS de la matrice de distance euclidienne intergènes.*

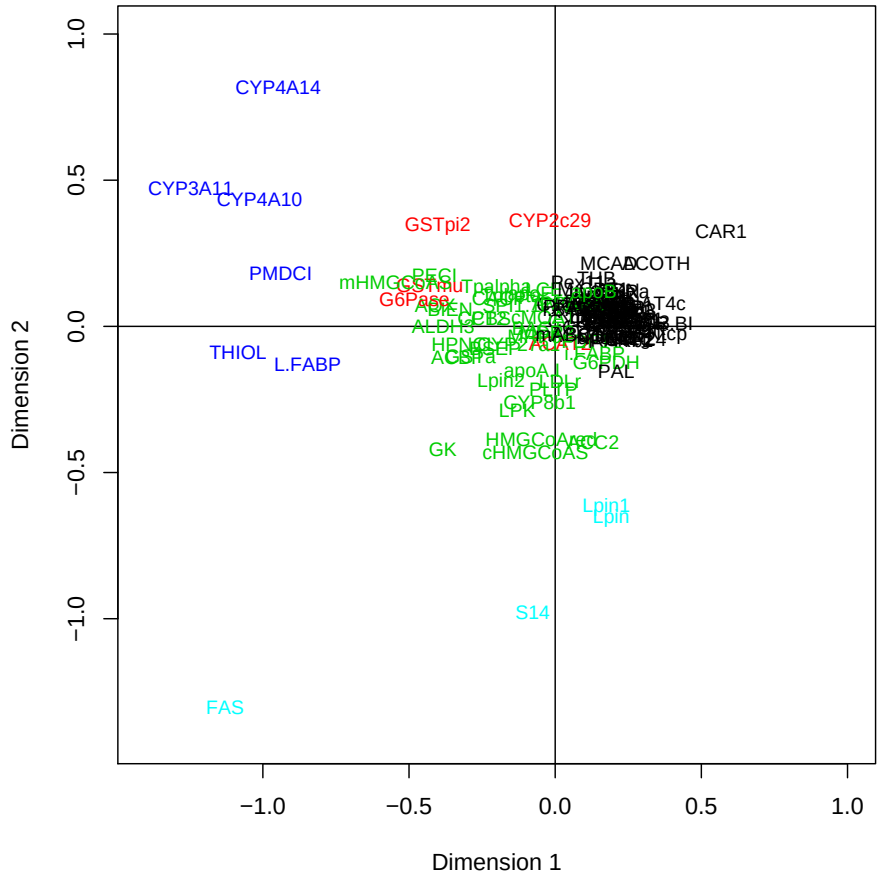


FIG. 6.10 – *Souris : représentation par positionnement multidimensionnel (distance euclidienne) des 5 groupes issues de la classification hiérarchique des gènes.*

La représentation de la figure 6.10 est analogue à celle déjà présentée (Fig. 5.6). Elle est complétée par un codage en couleurs des gènes, selon leur appartenance à une classe issu de la classification hiérarchique. Pour cela, nous avons coupé l'arbre afin d'en extraire 5 classes.

Brièvement, on peut noter que l'axe 1 met en évidence l'opposition précédemment évoquée entre CAR1 (surexprimé chez les souris `PPAR`) et un groupe de gènes (`CYP3A10`, `CYP4A10`, `CYP4A14`, `PMDCI`, `THIOL` et `L-FABP`) qui est surexprimé chez les souris `WT`. De manière similaire, l'axe 2 oppose les gènes induits par le régime `dha` (valeurs positives, gènes impliqués dans le catabolisme des lipides et dans le métabolisme des xénobiotiques) aux gènes induits par le régime `efad` (valeurs négatives, gènes principalement impliqués dans la synthèse de lipides). En remontant vers les feuilles de l'arbre de classification, on notera que le groupe des gènes représentés en vert est séparé en deux sous-groupes qui conservent une cohérence vis-à-vis des fonctions biologiques de catabolisme et de synthèse des lipides respectivement. Une observation des données individuelles révèle que ces régulations opérées par les régimes semblent plus marquées chez les souris `WT`. Baccini et col. (2005) montrent que d'autres techniques (forêts aléatoires par exemple) permettent de confirmer ces observations de manière plus objective.

Chapitre 7

Modèle linéaire et régression

1 Introduction

Ce chapitre fait suite aux précédents sur les analyses descriptives en adoptant un esprit différent puisqu'abordant la statistique inférentielle.

Un modèle linéaire est une expression qui relie une variable quantitative (la variable à expliquer) à des variables, quantitatives et/ou qualitatives (les variables explicatives).

Les analyses des modèles linéaires portent des noms différents selon la nature des variables explicatives utilisées dans le modèle. Le tableau suivant contient le nom des différentes analyses par nature des variables explicatives.

Variables explicatives	Nom de l'analyse
1 quantitative	régression simple
plusieurs quantitatives	régression multiple
plusieurs qualitatives	analyse de variance
1 ou plusieurs quantitatives et plusieurs qualitatives	analyse de covariance

Il existe cependant une théorie statistique englobant ces divers types de modèles : le modèle linéaire.

Notons que si non plus une, mais plusieurs variables quantitatives sont à expliquer conjointement, on se place dans le cadre de la régression multivariée, qui est fortement liée à l'analyse canonique. D'autre part, si la variable à expliquer est qualitative plutôt que quantitative, d'autre modèles sont à mettre en place comme la régression logistique ou la régression loglinéaire qui s'intègrent dans la famille du modèle linéaire général.

Dans la suite, nous aborderons en détail le modèle de régression simple, puis nous passerons en revue les autres modèles avec leurs spécificités en gardant en mémoire que les méthodes d'estimation des paramètres, les tests et les analyses diagnostics sont identiques.

2 Le modèle de régression simple

Le but d'une analyse de régression est d'étudier les relations qui existent entre des facteurs/variables mesurables à partir d'observations (données) prises sur ces facteurs. Des objectifs plus précis d'une telle analyse peuvent être :

- la prévision (ex : étant donné l'âge, fumeur/non fumeur, le poids, etc ..., combien d'années

un individu devrait-il survivre ?) ;
- la sélection de variables (ex : Parmi la température, l'ensoleillement, la pluie, l'altitude, le bruit ambiant, etc ..., quels facteurs ont une influence significative sur la croissance des pins des landes ?) ;
- la spécification de modèle (ex : Comment la durée de vie de transformateurs électriques varie-t-elle en fonction de leur grosseur ?) ;
- l'estimation de paramètres (ex : la luminosité en fonction de la distance des étoiles d'une certaine galaxie est de la forme $L = K_1 + K_2 d + \sigma\epsilon$, où K_1, K_2 et σ sont des paramètres inconnus à estimer à partir des observations).

Données pH. On veut étudier sur des carpes le pH (x) du milieu ambiant et le pH (y) de leur sang (données simulées) :

```
R ¿ x<- round(runif(30)*5+1,1)
```

```
R ¿ y<-2+3*x+rnorm(30,0,1)
```

Les données consistent en 30 unités statistiques (u.s.). Pour l'u.s. i, on a (x_i, y_i). Au vu de la

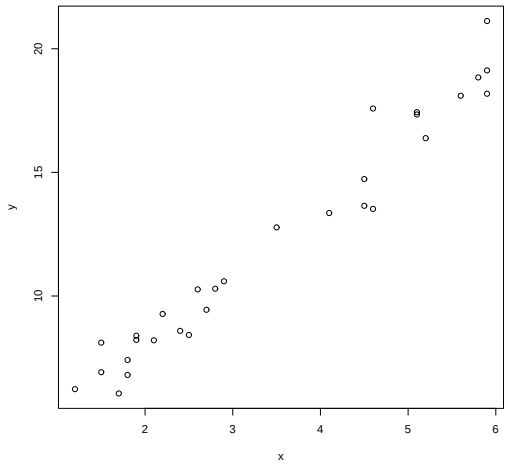


FIG. 7.1 – Données pH : le pH sanguin de 30 carpes *vs.* le pH ambiant.

figure 7.1, on pressent qu'il existe une relation linéaire entre x et y : $y \doteq \beta_1 + \beta_2 x$. On écrit donc le modèle de régression suivant, expliquant y par une combinaison linéaire de paramètres à estimer (β_1 et β_2) :

$$y_i = \beta_1 + \beta_2 x_i + e_i, \text{ pour } i = 1, ...30; \tag{7.1}$$

où e_i est un résidu que l'on espère le plus petit possible. La variable y est appelée variable endogène (variable réponse, variable dépendante) ; les variables x_i sont appelées variables exogènes (variables explicatives, facteurs, covariables, variables indépendantes).

Hypothèses : Les observations y_i sont des réalisations de 30 variables aléatoires indépendantes Y_i de moyenne $\beta_1 + \beta_2 x_i$ et de variance σ^2. De manière équivalente, les résidus e_i sont des réalisations de 30 variables aléatoires indépendantes E_i de moyenne 0 et de variance σ^2.

2.1 Ecriture et hypothèses du modèle

Nous observons n paires (y_1, x_1), ..., (y_n, x_n) et supposons que :

$$y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, ..., n;$$

où
$y_1, y_2, ..., y_i, ..., y_n$ sont les n observations de la variable endogène (ou variable à expliquer),
$x_1, ..., x_n$ sont les n observations de la variable exogène (ou variable explicative),
$e_1, ..., e_n$ sont les n termes d'erreur,
β_1 est le paramètre d'ordonnée à l'origine (la valeur moyenne de y lorsque x prend la valeur 0),
β_2 est le paramètre de pente (si x augmente d'une unité, alors y augmente de β_2 unités en moyenne).

Trois hypothèses sont essentielles à faire sur la distribution des termes d'erreur :

(i) les résidus sont de moyenne nulle (hypothèse de linéarité)

$$\mathrm{E}(e_i) = 0, \forall i = 1, ..., n$$

(ii) les e_i ont une variance identique (homoscédasticité)

$$\mathrm{Var}(e_i) = \sigma^2, \forall i = 1, ..., n$$

(iii) les e_i sont indépendants (donc non corrélés)

$$\mathrm{Cov}(e_i, e_j) = 0, \forall i \neq j$$

En supposant les valeurs x_i, $(i = 1, ..., n)$ comme étant non aléatoires (déterministes), les hypothèses ci-dessus impliquent que $\mathrm{E}(y_i) = \beta_1 + \beta_2 x_i$; $\mathrm{Var}(y_i) = \sigma^2$ et $\mathrm{Cov}(y_i, y_j) = 0$. On voit bien que quelle que soit la valeur de la variable explicative, seule l'espérance de y dépend de x, c'est à dire que la variance de y et la covariance entre deux observations de la variable à expliquer ne dépendent pas de la valeur de la variable explicative.

La droite de régression $(\beta_1 + \beta_2 x)$ représente la valeur attendue de y en fonction de la valeur de x. Les valeurs observées de y sont distribuées de façon aléatoire autour de cette droite. Les termes d'erreur sont les différences entre les valeurs observées de y et la droite. Comme la variance de ces termes d'erreur est constante en x, la distance moyenne des points à la droite est la même pour toute valeur de x. Finalement la non-corrélation entre les termes d'erreur signifie que la valeur d'un terme d'erreur n'est pas influencée par la valeur des autres termes d'erreur.

L'expression 2.1 se décompose de manière classique en :

$$\text{observation} = \text{modèle} + \text{résidu}.$$

2.2 Le modèle linéaire gaussien

Dans le modèle linéaire expliqué au paragraphe précédent, seuls les deux premiers moments des termes d'erreur (espérance et variance) sont supposés connus. Dans un modèle linéaire gaussien, on se donne une hypothèse supplémentaire : la distribution des résidus est supposée normale.

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

Cette hypothèse implique que les variables aléatoires y_i sont normalement distribuées.

2.3 Estimation des paramètres β_1 et β_2

Dans cette partie, la variance σ^2 est supposée connue. Les paramètres inconnus sont β_1 et β_2 et ils sont estimés à partir des données observées $(y_i, x_i,\ i = 1, ..., n)$. Deux grandes méthodes sont utilisées :

- la méthodes des moindres carrés, qui ne suppose connues que l'espérance et la variance de y ;
- la méthode du maximum de vraisemblance, qui suppose les résidus gaussiens.

On notera classiquement les estimations avec un "chapeau". Par exemple, $\hat{\beta}_1$ désigne l'estimation de β_1, c'est-à-dire une fonction de y (et de x). Rappelons ici que x est supposé connu, alors que y est une variable aléatoire (ou sa réalisation).

On appelle *ième* valeur ajustée ou prédite ou attendue la fonction de y suivante :

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

et le résidu correspondant vaut :

$$\hat{e}_i = y_i - \hat{y}_i.$$

Les moindres carrés ordinaires

L'idée est de trouver la droite qui explique la plus grande partie possible de la relation entre la variable endogène et la variable exogène, c'est à dire trouver la droite qui minimise la partie inexpliquée ou la partie due à la fluctuation aléatoire. On cherche donc la droite qui passe le plus près possible de tous les points ou, en d'autres termes, la droite qui minimise la distance des points à la droite (les termes d'erreurs).

La méthode des moindres carrés consiste à trouver les valeurs de $\hat{\beta}_1$ et $\hat{\beta}_2$ (estimateurs de β_1 et β_2) qui minimisent la somme des carrés des écarts entre les valeurs observées et les valeurs ajustées (somme des carrés des résidus) :

$$\min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} \hat{e}_i^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\hat{\beta}_1,\hat{\beta}_2} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$$

Puisque la fonction à minimiser a de bonnes propriétés (lisse, convexe), elle se minimise en prenant les dérivées de la somme par rapport à $\hat{\beta}_1$ et $\hat{\beta}_2$, en posant ces dérivées égales à zéro et en résolvant le système de deux équations à deux inconnues. On obtient :

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

et

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Prenons l'exemple des données pH. Avec le logiciel R, on peut ajuster le modèle (7.1) :

```
> reg1 <- lm(y ~ x)

Coefficients :
 (Intercept)           x
    5.572096   0.1951841

Degrees of freedom : 7 total; 5 residual
Residual standard error : 0.05754663
```

```
> plot(x,y,xlab="pH ambiant (x)",ylab="pH sanguin (y)")
> lines(x,reg1$fitted.values)
```

On vérifie graphiquement que l'ajustement est cohérent en comparant valeurs observées et valeurs ajustées (figure 7.2)

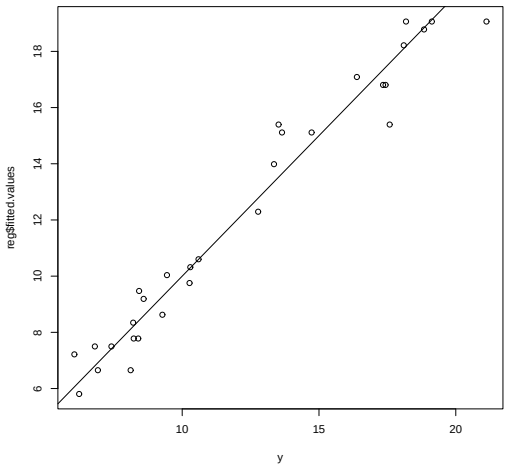


FIG. 7.2 – Données pH : droite de régression du modèle (7.1).

La méthode du maximum de vraisemblance

. Cette méthode nécessite l'ajout de l'hypothèse de normalité des résidus : $e_i \sim iidN(0, \sigma^2)$, ce qui implique que les y_i sont des variables aléatoires normales indépendantes :

$$y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2).$$

Cette méthode repose sur l'idée suivante : si les données de l'échantillon ont été observées, cela provient du fait que ces données sont les plus vraisemblables. Les estimateurs des paramètres inconnus du modèle sont donc calculés en maximisant une quantité (vraisemblance) qui mesure la probabilité d'observer l'échantillon. Dans le cadre de la régression linéaire simple, on cherche donc à maximiser la fonction de vraisemblance :

$$\begin{aligned} L(\beta_1, \beta_2, \sigma^2) &= \prod_{i=1}^{n} f(y_i; x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2} \\ &= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 \right\} \end{aligned}$$

Le logarithme de la vraisemblance, multiplié par (-2), s'écrit

$$l(\beta_1, \beta_2, \sigma^2) = -2 \ln L(\beta_1, \beta_2, \sigma^2) = n \ln(2\pi) + n \ln \sigma^2 + \sigma^{-2} \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$$

Maximiser la vraisemblance équivaut donc à faire des moindres carrés.

2.4 Propriétés des estimateurs

Pour le modèle de régression simple, on peut montrer que :

$\mathrm{E}(\hat{\beta}_1) = \beta_1$ et $\mathrm{E}(\hat{\beta}_2) = \beta_2$

On dit alors que les estimateurs des paramètres sont sans biais. Rappelons que l'espérance est calculée par rapport à la loi de y.

D'autre part,

$$\mathrm{Var}(\hat{\beta}_2) = \mathrm{Var}(\frac{S_{xy}}{S_{xx}}) = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\bar{y} - \hat{\beta}_2 \bar{x}) = \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathrm{Cov}(\bar{y} - \hat{\beta}_2 \bar{x}, \hat{\beta}_2) = -\frac{\bar{x}\sigma^2}{S_{xx}}$$

On remarque ici que les deux estimateurs peuvent être très fortement corrélés. Pour éviter cela, on peut opérer une reparamétrisation :

$$\beta_0 + \beta_2 (x_i - \bar{x})$$

avec $\beta_0 = \beta_1 + \beta_2 \bar{x}$. Les estimateurs de ces deux paramètres ne sont pas corrélés : $\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_2) = 0$, et la variance d'estimation de β_0 est plus faible que celle de β_1 : $\mathrm{Var}(\hat{\beta}_0) = \sigma^2/n < \mathrm{Var}(\hat{\beta}_1)$. Cette remarque souligne l'importance d'une bonne paramétrisation sur la précision des estimations. Des problèmes numériques sont aussi évités.

2.5 Estimation ponctuelle de σ^2

En général, on ne connait pas la valeur de σ^2, il faut alors l'estimer. Comme les résidus $\hat{e}_i = y_i - \hat{y}_i$ peuvent être vus comme des estimateurs des e_i, la variance d'échantillonnage des $\hat{e}_i$ est un estimateur raisonnable de $\sigma^2 = \mathrm{Var}(e_i)$. Un estimateur sans bais est donné par :

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\mathrm{SSE}}{n-2}$$

où $\mathrm{SSE} = \sum_{i=1}^{n} \hat{e}_i^2$ est la somme des carrés résiduels.

2.6 Tests d'hypothèse et intervalles de confiance

Si les y_i sont des variables aléatoires normales, puisque les $\hat{\beta}$ sont des combinaisons linéaires des y_i, alors ces estimateurs sont donc aussi des variables alétoires normales. Plus particulièrement :

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

$$\hat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{S_{xx}}).$$

On peut donc standardiser les estimateurs pour obtenir :

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_2 - \beta_2}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

Comme σ^2 n'est pas connue, nous remplaçons dans les expressions ci-dessus σ^2 par son estimation s^2. Ce faisant, on doit corriger la distribution (cf. Annexe D, plus le fait que $s^2 \sim \chi^2_{n-2}$) afin d'obtenir :

$$\frac{\hat{\beta}_1 - \beta_1}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

$$\frac{\hat{\beta}_2 - \beta_2}{s/\sqrt{S_{xx}}} \sim t_{n-2}$$

Les deux équations ci-dessus nous mènent aux intervalles de confiance à $(1-\alpha)100\%$ suivants pour β_1 et β_2 :

$$\left[\hat{\beta}_1 \pm t_{\alpha/2;n-2}\, s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right]$$

$$\left[\hat{\beta}_2 \pm t_{\alpha/2;n-2}\, s/\sqrt{S_{xx}}\right]$$

On peut aussi tester l'hypothèse de nullité d'un des paramètres. Par exemple, pour tester l'hypothèse nulle $H_0 : \beta_1 = 0$ *vs* l'hypothèse alternative $H_1 : \beta_1 \neq 0$, on utilise la statistique :

$$t_1 = \frac{\hat{\beta}_1}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

qui est distribuée selon une loi de Student à $n-2$ degrés de liberté.
On rejettera donc l'hypothèse nulle si $t_1 > t_{\alpha/2;n-2}$ au niveau α. Il est clair que ce test correspond à la présence (H_0) ou à l'absence (H_1) de 0 dans l'intervalle de confiance.

2.7 Vérification des hypothèses

Tous les résultats (estimation, tests, intervalle de confiance) du modèle linéaire reposent sur des hypothèses fondamentales faites sur la distribution des termes d'erreur. Les résidus du modèle sont donc des outils privilégiés pour vérifier ces hypothèses. Voici un plan de vérification qu'il serait bon de suivre après chaque analyse (les hypothèses à vérifier sont classées par ordre d'importance décroissante) :

i. Vérifier que les résidus sont centrés : $\mathrm{E}(\boldsymbol{e}) = 0$ (hypothèse de linéarité) ;
ii. Vérifier l'homoscédasticité (la variance des résidus doit être constante) ;
iii. Vérifier l'indépendance des observations (hypothèse de non corrélation) ;
iv. Vérifier la normalité des résidus.

Plusieurs versions de ces résidus existent :

Résidus ordinaires : le i *ème* résidu (ordinaire) est défini comme étant $e_i = y_i - \hat{y}_i$. Si l'hypothèse 1) est vraie, alors $\mathrm{E}(e_i) = 0$. Si les hypothèses 2) et 3) sont vraies, alors

$$\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2 \quad \text{etCov}(e_i, e_j) = -h_{ij}\sigma^2,$$

où $h_{ij} = 1/n + (x_i - \bar{x})(x_j - \bar{x})/S_{xx}$.

Enfin, si l'hypothèse iv) est vraie, $e_i \sim N(0, \sigma^2(1 - h_{ii}))$.

Avec le logiciel R, ces résidus sont obtenus avec la commande `monmodele$residuals`, où `monmodele` est le résultat de la fonction `lm` : `monmodele <- lm(y ~ x)`

Résidus standardisés : le i *ème* résidu standardisé est défini comme étant $r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$.

Avec le logiciel R, on obtient ces résidus par la commande `rstandard(monmodele)`.

Résidus studentisés : ce sont de légères modifications des précédents : $t_i = \frac{e_i}{s_i\sqrt{1-h_{ii}}}$, où s_i^2 est une estimations sans biais de $Var(\hat{e}i)$. On montre que les t_i suivent une loi de Student à $n - 3$ degrés de liberté.

Avec le logiciel R, on obtient ces résidus par la commande `rstudent(monmodele)`.

Vérification de la linéarité

Graphique des résidus vs valeurs ajustées : $\hat{y}_i$ vs $\hat{e}_i$ (figure 7.3)

Ce graphique permet surtout de cerner les problèmes avec l'hypothèse 1) de linéarité. Si l'hypothèse est raisonnable, ce graphique devrait montrer un nuage de points centrés horizontalement autour de 0. Le graphique devrait avoir une allure complètement aléatoire, c'est à dire qu'il ne devrait y avoir aucune tendance discernable (e_i croissant ou décroissant avec $\hat{y}_i$, graphique à l'allure quadratique, etc...). Ce graphique peut également cerner des problèmes avec les autres hypothèses, mais les graphiques basés sur les résidus studentisés sont plus appropriés dans ces cas.

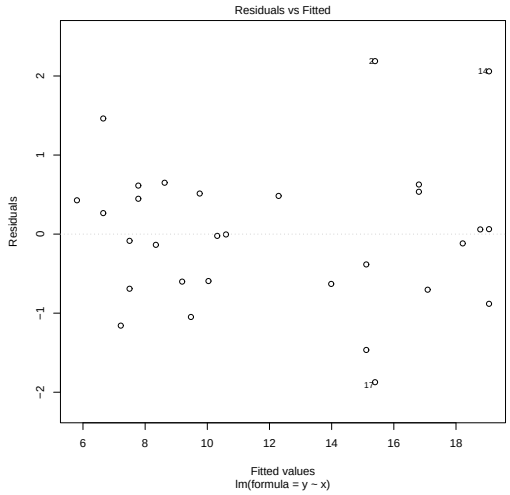


FIG. 7.3 – Données pH : plot des résidus vs valeurs ajustées, pour le modèle (7.1).

Graphique des résidus vs variable explicative : x_i vs $\hat{e}_i$ (figure 7.4)

Encore une fois, ce type de graphique permet de détecter des problèmes avec l'hypothèse de linéarité ; il devrait avoir l'air d'un nuage de points dispersés horizontalement de façon aléatoire autour de 0.

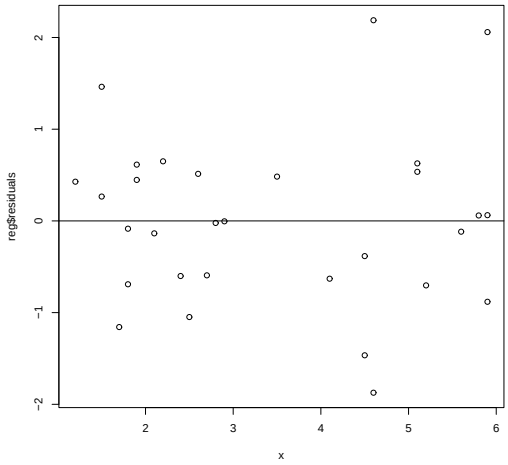


FIG. 7.4 – Données pH : plot des résidus vs variable explicative, pour le modèle (7.1).

Vérification de l'homoscédasticité

Cette hypothèse est importante. Une forte violation de cette dernière entraîne des conséquences désastreuses sur : les erreurs standards des paramètres, les risques des tests, les intervalles de confiance. La méthode la plus couramment utilisée est la vérification graphique. Elle consiste à représenter les résidus en fonction des valeurs ajustées, des valeurs observées ou des valeurs de x. On peut également utiliser les résidus studentisés pour vérifier l'hypothèse d'homoscédasticité. Un graphique ayant une apparence d'entonnoir indique que la variance ne semble pas constante (problème d'hétéroscédasticité). Si certains résidus ont des valeurs plus grandes que 2 en valeur absolue, ceci peut indiquer un manque de normalité ou la présence de données atypiques.

Vérification de l'indépendance

Graphique des résidus vs numéro d'observations : $\hat{e}_i$ vs i

Ce graphique sert à vérifier l'hypothèse de non corrélation des résidus. Si les résidus de grande (faible) valeur ont tendance à suivre des résidus de grande (faible) valeur, alors il y a un problème d'autocorrélation positive. Si les résidus de grande (faible) valeur ont tendance à suivre des résidus de faible (grande) valeur, alors il y a un problème d'autocorrélation négative.

Quand la régression est réalisée sur des données qui varient au cours du temps, les observations peuvent ne pas être indépendantes. Pour vérifier l'indépendance, un test est habituellement utilisé : le test de Durbin-Watson. Il est basé sur la statistique :

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Les e_i sont les résidus de la régression et n est le nombre d'observations. On peut montrer que $0 \leq d \leq 4$ et que $d \simeq 2 - 2\frac{\sum_{i=2}^{n}(e_i e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \simeq 2 - 2\rho_e$ où ρ_e est le coefficient d'autocorrélation

d'ordre 1 des résidus. Il est obtenu en calculant la corrélation entre la série des résidus et la même série décalée de 1. Si $\rho_e = 0$ soit $d \simeq 2$ alors les résidus sont non corrélés. Si par contre $\rho_e \neq 0$ ou encore $d \neq 2$ alors les résidus sont corrélés.

Vérification de la normalité

Cette étape n'est pas aussi importante qu'on le croit généralement. La normalité est une propriété qui permet aux estimateurs de converger rapidement. Le théorème central limite nous assure que pour des échantillons assez grands, les estimateurs que nous utilisons sont normalment distribués. La symétrie des distributions observées est un critère important qui assure une convergence rapide vers la loi normale. Les méthodes pour vérifier la normalité sont nombreuses, parmi celles-ci on peut citer les méthodes graphiques (QQplot, PPplot, histogrammes, boxplot, etc...) et les tests (Chi2, Shapiro-Wilk, Kolmogorov-Smirnov, ...).

Graphique des résidus studentisés vs quantiles de la loi normale : t_i vs u_i (figure 7.5) Ce graphique permet de détecter les problèmes avec le postulat de normalité. Il est parfois appelé QQplot normal ou droite de Henry, tout dépend de la forme utilisée pour les u_i. Dans le QQplot, il s'agit des quantiles de la loi normale standard. Dans le cas de la droite de Henry, il s'agit de l'espérance des statistiques d'orde de la loi normale standard. Dans les deux cas, si l'hypothèse de normalité est raisonnable, le graphique devrait avoir la forme d'une ligne droite de pente positive. Des graphiques à l'allure de courbe concave ou convexe indiquent une distribution non symétrique des résidus, alors qu'un graphique en forme "d'intégrale inversée couchée" indique que les résidus proviennent d'une distribution ayant des queues plus épaisses que celles de la loi normale.

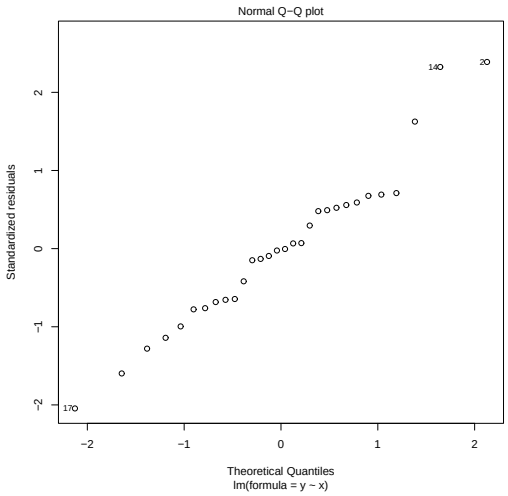


FIG. 7.5 – Données pH : QQplot du modèle (7.1).

Boxplot des résidus : Le Boxplot des résidus (ordinaires ou studentisés) sert à déterminer si ces derniers proviennent d'une distribution symétrique et si certains résidus sont de valeur extrême. Une distribution non symétrique est indiquée par une moustache plus longue que l'autre, ou une ligne médiane proche d'une extremité de la boîte. Un résidu extrême est indiqué par un point à l'extérieur des moustaches.

Détection et élimination de valeurs atypiques

Un examen critique des données est une étape importante en statistique. Il existe deux grands types de données généralement classées comme atypiques : les données qui ne sont pas habituelles, et les données qui violent une hypothèse de l'analyse statistique utilisée. Différentes attitudes devraient être adoptées suivant la nature du problème rencontré. Les données provenant d'erreurs grossières de mesures ou d'erreurs de frappe doivent être supprimées de l'analyse. Seul un jugement biologique permet de déclarer une valeur comme aberrante. Souvent, après un examen attentif des données on trouve des valeurs inhabituelles. Un expérimentateur prudent doit alors rechercher la (les) causE(s) de telles valeurs. Deux cas de figures se présentent alors : soit la cause est identifiée et il faut changer la donnée ou la méthode d'analyse ; soit la cause n'est pas identifiée et un test statistique peut être utilisée pour détecter une valeur atypique.

L'examen graphique des résidus est un bon outil (graphique des $\hat{e}$ en fonction de $\hat{y}$). Une autre technique consiste à calculer des indices pour chaque résidu. La plupart des indices calculés par les logiciels de statistique ont une signification inférentielle. Les trois les plus couramment usités sont : les résidus standardisés, les distances de Cook, les contributions.

Avec les résidus standardisés, il est donc possible de tester l'"aberrance" de chaque résidu en utilisant un test de Student. Attention toutefois aux tests multiples (voir plus loin).

Les contributions (leverage) et les mesures de Cook mesurent la contribution de chaque résidu à la variance résiduelle (non expliquée par le modèle). Sous les hypothèses usuelles (hypothèses du modèle), les distances de Cook suivent une loi de Fisher à p et $n-p$ degrés de liberté. Elles s'obtiennent par :

$$D_i = \frac{h_{ii}}{2(1-h_{ii})} r_i^2,$$

tenant compte ainsi de l'importance du résidu i et de l'influence h_{ii} de l'observation i sur la droite de régression (effet "levier"). Une méthode pour identifier les observations qui contribuent trop à la variance résiduelle consiste à réaliser un test de Fisher sur le résidu de Cook (i.e. de comparer sa valeur limite à un seuil donné d'une loi de Fisher à p et $n-p$ ddl).

Pour des données gaussiennes, les leverage devraient être voisines de $\frac{p}{n}$; p représente le nombre de paramètres indépendants estimés dans le modèle. Si pour un résidu, le leverage correspondant est supérieur à $\frac{2p}{n}$, la donnée peut être considérée comme suspecte.

Comment régler les problèmes ?

Manque de linéarité : Ceci est en général dû à une mauvaise spécification de la forme de la relation entre la variable endogène et la variable exogène. Le problème peut être réglé par une ou plusieurs options suivantes :

- transformer la variable exogène, ajouter au modèle des termes en x_i^2, x_i^3, ... ;
- ajouter au modèle de nouvelles variables exogènes ;
- transformer la variable endogène.

Hétéroscédasticité : La transformation de Box-Cox pourra souvent prescire une transformation de la variable endogène qui règlera ce problème (transformation stabilisatrice de la variance). Si la transformation de Box-Cox ne fonctionne pas, alors la régression pondérée ou l'utilisation d'une autre méthode statistique peut être utile.

Méthode de Box-Cox : Cette méthode suppose un modèle de régression général de la forme : $g(y_i;\lambda) = \beta_1 + \beta_2 x_i + e_i$ où λ est un paramètre inconnu et qu'il faut estimer à partir des données ; $g(y;\lambda) = \frac{y^\lambda - 1}{\lambda}$ si $\lambda \neq 0$ et $g(y;\lambda) = \ln \lambda$ si $\lambda = 0$. On estime λ en même temps que les autres

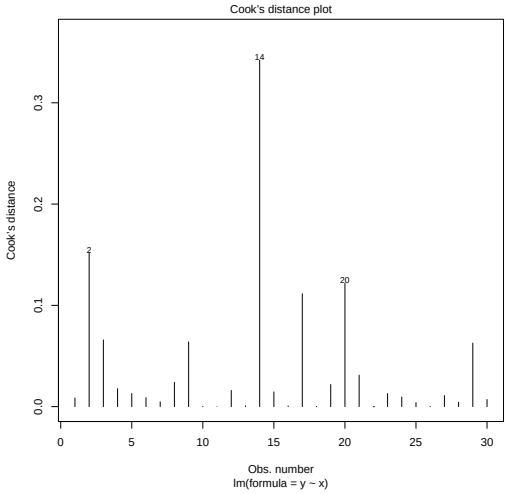


FIG. 7.6 – Distance de Cook pour le modèle (7.1) sur les données pH.

paramètres par la méthode du maximum de vraisemblance.

Auto-corrélation des résidus : ce problème est plus difficile à régler. Il est parfois possible de le régler en ajoutant une variable exogène qui explique pourquoi il y a autocorrélation (par exemple, si les premières mesures sont faites sur l'individu A, les mesures suivantes sur l'individu B, etc., alors peut-être ajouter une variable exogène dénotant l'individu sur lesquelles les mesures ont été faites pourra régler le problème). Mais en général, il faut avoir recours à un modèle plus complexe que le modèle linéaire simple (séries temporelles, modèle linéaire mixte).

Manque de normalité : Encore une fois, la transformation de Box-Cox règle souvent le problème. Parfois, le manque de normalité est tout simplement dû à quelques observations extrêmes. Dans ces cas, nous pourrons réglér le problème en traitant ces observations de façon appropriée. Une autre option consiste à utiliser des méthodes de régression robustes ou non paramétriques (non abordées dans ce cours).

3 Régression lineaire multiple

Quand on dispose de plusieurs variables explicatives, on peut mettre en oeuvre un modèle de régression linéaire multiple. Supposons que l'on dispose de n observations sur une variable continue y et p variables continues $x_1, ..., x_j, ...x_p$. On note y_i (resp. $x_{j,i}$) la *ième* observation de y (resp. x_j). Le modèle suivant :

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_j x_{j,i} + ... + \beta_p x_{p,i} + e_i$$

est un modèle de régression multiple, cherchant à expliquer la variable y par p variables explicatives $x_1, ...x_j, ...x_p$, à l'aide d'un échantillon de taille n.

Les résidus e_i sont supposés de moyenne nulle, de variance égale σ^2, et mutuellement indépendants, éventuellement gaussiens.

Les paramètres inconnus $\beta_1, ...\beta_j, ...\beta_p$ (et éventuellement σ^2) sont estimés par moindres

carrés (la somme des carrés des résidus la plus petite possible), ou par maximum de vraisemblance si les résidus sont supposés gaussiens, exactement comme dans le cas de la régression simple.

3.1 Multicolinéarité

Des problèmes d'estimation des paramètres et de variance peuvent arriver lorsque dans le modèle de régression, on utilise des variables explicatives corrélées entre elles. On parle alors de multicolinéarité et cela conduit à des estimations biaisées des paramètres avec des variances importantes. Pour diagnostiquer ces situations, une des méthodes est de faire la régression de chaque variable en fonction des autres variables explicatives et de mesurer les liaisons à l'aide du coefficient R_j^2 de chacune de ces régressions (où R_j est le coefficient de corrélation multiple obtenu en régressant la variable x_j sur les $(k-1)$ autres variables explicatives). On appelle tolérance, la valeur $1 - R_j^2$. Une tolérance qui est proche de 1 signifie une absence de multicolinéarité entre les variables explicatives. En revanche, si la tolérance tend vers 0, alors on détecte un problème de multicolinéarité entre les variables explicatives.

3.2 Critères de sélection de modèle

Pour obtenir un compromis satisfaisant entre un modèle trop simple (grands résidus) et un modèle faisant intervenir beaucoup de variables (donc très instable), on dispose de plusieurs critères qui ne donnent pas nécessairement le même résultat.

Coefficient de détermination et ses variantes

Pour mesurer la qualité d'un modèle de régression linéaire, ou pour comparer des modèles de régression linéaire entre eux, on définit le coefficient de détermination :

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_E}{SS_{Tot}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = corr^2(y, \hat{y})$$

On a que $0 \leq R^2 \leq 1$. Quand $R^2 = 0$, toute la variabilité est due à l'erreur aléatoire et le modèle n'explique absolument rien de la valeur de y_i. Quand $R^2 = 1$, tous les points tombent sur la droite de régression, c'est à dire que l'ajustement du modèle est parfait et que la valeur de y_i est une fonction exacte de x_i.

Le coefficient de détermination R^2 peut donc être interprété comme la proportion de la variabilité dans les y_i qui est expliquée par le modèle de régression.

Bien que facilement interprété et naturellement attrayant, le coefficient de détermination souffre de quelques problèmes qui font qu'il ne peut pas être utilisé pour comparer n'importe quels modèles de régression l'un avec l'autre. L'inconvénient principal est que dès que l'on ajoute un terme à un modèle de régression, le coefficient de détermination augmente. Afin de circonscrire à ce problème, nous pouvons utiliser le coefficient de détermination ajusté :

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

avec n le nombre d'observations et p le nombre de paramètres. Avec le R^2_{ajust}, l'ajout d'une variable explicative peut aussi résulter en une diminution de la statistique. La comparaison de modèles sur la base de ce critère revient à comparer deux modèles sur la base de leur estimé de la variance des termes d'erreur s^2. Le meilleur modèle sera celui ayant le R^2_{ajust} le plus grand.

C_p de Mallows

Une autre critère appelé, le coefficient C_p de Mallows peut être utilisé. Il est défini par :

$$C_p = \frac{SS_E}{\hat{\sigma}^2} - n + 2p$$

où SS_E est la somme des carrés résiduels du modèle et $\hat{\sigma}^2$ est l'estimation de la variance résiduelle sous le modèle complet. On choisira la modèle pour lequel le coefficient C_p est minimum.

Test de Fisher pour modèles emboités

Il se peut qu'on veuille tester si le modèle à p variables explicatives peut être réduit à q (q petit devant p) variables ; c'est à dire que l'on veut tester si un sous-modèle plus simple explique une partie suffisamment grande de la variabilité dans les y_i pour qu'il ne soit pas nécessaire d'utiliser le modèle le plus complexe (car trop de paramètres à estimer). Cela revient à tester l'hypothèse de nullité de $k(= p - q)$ paramètres du modèle :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \text{ avec } k \text{ petit devant } p$$

Sous l'hypothèse alternative, au moins un des paramètres $\beta_1, ..., \beta_k$ est non-nul.

Ce test peut être formulé comme la comparaison de deux modèles emboités, l'un à $p + 1$ paramètres et l'autre à $q+1$ paramètres. L'hypothèse H_0 peut être testée au moyen de la statistique :

$$F_{cal} = \frac{SS_{E_0} - SS_{E_1}}{SS_{E_1}} \frac{n - p - 1}{k} \sim F(k, n - p - 1)$$

où SS_{E_0} est la somme des carrés résiduelles du modèle réduit sous H_0 et SS_{E_1} est la somme des carrés résiduelles du modèle de référence (modèle complet à p variables explicatives).

On compare F_{cal} à la valeur limite de la statistique d'une loi de Fisher $F_\alpha(k, n - p - 1)$. Si $F_{cal} > F_\alpha(k, n - p - 1)$ alors on rejette H_0.

Remarque : Dans le cas où $k = 1$, on teste la nullité d'un seul paramètre du modèle. Etant donné la propriété selon laquelle une variable aléatoire distribuée selon une loi $F(1, m)$ est le carré d'une variable aléatoire de Student à m degré de liberté ; le test de Fisher ci-dessus et le test de Student donnent les mêmes conclusions.

Chapitre 8

Modèle linéaire : analyse de variance

L'analyse de variance est un cas particulier de la régression. La différence essentielle est la structure que possèdent les variables explicatives. L'objectif de l'analyse de variance est la recherche de relations entre une variable quantitative et des variables qualitatives (appelées aussi facteurs de variation). Quand un seul facteur de variation est utilisé pour expliquer les variations de Y, on réalise une analyse de la variance à un facteur (à une voie). Dans le cas général, plusieurs facteurs (p) sont présents pour expliquer les variations de Y, on parle alors d'analyse de variation à p facteurs.

1 ANOVA à un facteur

1.1 Un exemple

Données Ampoules. On considère maintenant plusieurs procédés de fabrication de lampes à ultra-violets : On numérote les u.s. (i, j), où i est le numéro du procédé de fabrication et j est le

TAB. 8.1 – Observations de durées de vie d'ampoules échantillonnées pour 6 procédés de fabrication.

F1	1602 1615 1624 1631
F2	1472 1477 1485 1493 1496 1504 1510
F3	1548 1555 1559 1563 1575
F4	1435 1438 1448 1449 1454 1458 1467 1475
F5	1493 1498 1509 1516 1521 1523
F6	1585 1592 1598 1604 1609 1612

numéro de la lampe à i fixé. On note y_{ij} la durée de vie de la j *ème* lampe fabriquée suivant le procédé i, et μ_i la durée de vie moyenne d'une lampe fabriquée suivant le procédé i.

Le modèle s'écrit :

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, ...6 \quad j = 1, ...n_i \tag{8.1}$$

où e_{ij} est un résidu tel que $e_{ij} \sim N(0, \sigma^2)$ et n_i le nombre d'observations pour le procédé i. Les résidus sont supposés être indépendantes. Le modèle peut également s'écrire comme celui d'une régression linéaire multiple :

$$y_{ij} = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_6 x_6 + e_{ij}$$

avec $x_i = 1$ si y_{ij} est une observation faite dans la classe i et $x_i = 0$ sinon. On a la relation suivante entre l'écriture des deux modèles d'analyse de variance : $\beta_i = \mu_i$.

1.2 Diverses paramétrisations

Analysons les données Ampoules.

Avec le logiciel R

```
> options(contrasts="contr.treatment")

> reg2 <-lm(dvie ~ proc)

> summary(reg2)

Call : lm(formula = dvie   proc)
Residuals :
    Min  1Q     Median   3Q  Max
    -19  -9  4.996e-15  9.5   22
   Coefficients :
                  Value  Std.Error   t value  Pr(>|t|)
 (Intercept)  1618.0000     6.2022  260.8773    0.0000
      procF2  -127.0000     7.7748  -16.3348    0.0000
      procF3   -58.0000     8.3211   -6.9703    0.0000
      procF4  -165.0000     7.5961  -21.7218    0.0000
      procF5  -108.0000     8.0069  -13.4883    0.0000
      procF6   -18.0000     8.0069   -2.2480    0.0321

Residual standard error : 12.4 on 30 degrees of freedom
Multiple R-Squared : 0.9644
F-statistic : 162.7 on 5 and 30 degrees of freedom, the p-value is 0
```

¿ reg2$fitted.values

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1618	1618	1618	1618	1491	1491	1491	1491	1491	1491	1491	1560	1560	1560	1560
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1560	1453	1453	1453	1453	1453	1453	1453	1453	1510	1510	1510	1510	1510	1510
31	32	33	34	35	36									
1600	1600	1600	1600	1600	1600									

Avec le logiciel SAS

```
   proc glm data=ampoules;
class proc;
model dvie=proc / solution p;
run;

                          The GLM Procedure

Dependent Variable : dvie
                              Sum of
 Source            DF       Squares  Mean Square  F Value  Pr > F
 Model              5  125144.7500   25028.9500   162.67  <.0001
 Error             30    4616.0000     153.8667
 Corrected Total   35  129760.7500
```

```
R-Square  Coeff Var  Root MSE  dvie Mean
0.964427   0.812021  12.40430   1527.583

Source  DF    Type I SS  Mean Square  F Value  Pr > F
proc     5  125144.7500   25028.9500   162.67  <.0001

Source  DF  Type III SS  Mean Square  F Value  Pr > F
proc     5  125144.7500   25028.9500   162.67  <.0001

                           Standard
Parameter       Estimate       Error  t Value  Pr > |t|
Intercept  1600.000000 B  5.06403440   315.95    <.0001
proc F1      18.000000 B  8.00694143     2.25    0.0321
proc F2    -109.000000 B  6.90111562   -15.79    <.0001
proc F3     -40.000000 B  7.51117686    -5.33    <.0001
proc F4    -147.000000 B  6.69908783   -21.94    <.0001
proc F5     -90.000000 B  7.16162613   -12.57    <.0001
proc F6       0.000000 B           .        .         .

NOTE : The X'X matrix has been found to be singular, and a generalized inverse
was used to solve the normal equations. Terms whose estimates are followed
by the letter 'B' are not uniquely estimable.

Observation      Observed     Predicted     Residual
          1  1602.0000000  1618.0000000  -16.0000000
          2  1615.0000000  1618.0000000   -3.0000000
          3  1624.0000000  1618.0000000    6.0000000
          4  1631.0000000  1618.0000000   13.0000000
          5  1472.0000000  1491.0000000  -19.0000000
          6  1477.0000000  1491.0000000  -14.0000000
          7  1485.0000000  1491.0000000   -6.0000000
          8  1493.0000000  1491.0000000    2.0000000
          9  1496.0000000  1491.0000000    5.0000000
         10  1504.0000000  1491.0000000   13.0000000
  etc ...
```

On observe que les valeurs ajustées (reg2$fitted.values pour R et colonne Predicted pour SAS) sont les mêmes avec R et SAS, et pourtant les estimations des β (Value pour R, Estimate pour SAS) sont différentes. Pourquoi ?

Explication

En fait, il existe plusieurs paramétrisations possibles, que nous allons décrire.

- *Paramétrisation du modèle* : les paramètres sont les μ_i pour $i = 1, ...p$.
- *Décomposition centrée :* on écrit $\mu_i = \mu + \alpha_i$ avec $\sum \alpha_i = 0$.
- *Décomposition SAS/R : une cellule de référence.* On écrit $\mu_i = \mu_p + a_i$ avec $a_p = 0$: dans SAS, le dernier niveau du facteur sert de référence ; soit $a_i = (\mu_i - \mu_p)$ le contraste entre le niveau i et le niveau p. Ou encore $\mu_i = \mu_1 + a_i$ avec $a_1 = 0$ dans R, le premier niveau du facteur sert de référence ; soit $a_i = (\mu_i - \mu_1)$.

Toutes ces paramétrisations sont équivalentes car on peut passer de l'une à l'autre par une bijection.

1.3 Vérification des hypothèses - Diagnostics

Comme dans le cadre de la régression, des vérifications sont à effectuer : normalité des résidus, homoscédasticité, valeurs aberrantes ...

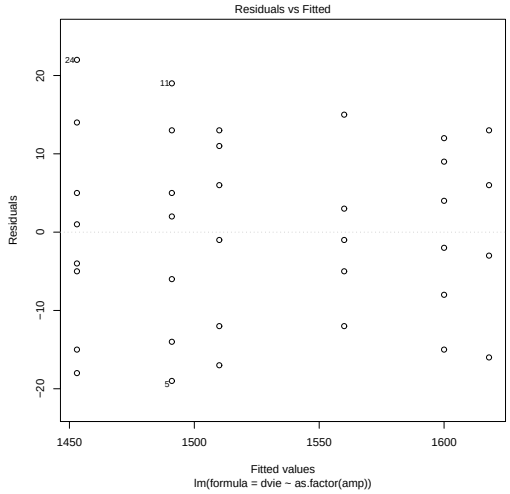


FIG. 8.1 – Données Ampoules : visualisation des résidus.

1.4 Estimation des paramètres

Comme pour la régression, les paramètres du modèle d'analyse de variance peuvent être estimés par la méthode des moindre carrés ou par la méthode du maximum de vraisemblance.

μ_i est estimé par la moyenne empirique :

$$\hat{\mu}_i = y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

Cette estimation est d'autant plus précise que le nombre d'observations pour la cellule i est grand :

$$Var(\hat{\mu}_i) = \frac{\sigma^2}{n_i}.$$

La variance résiduelle est estimée par :

$$\hat{\sigma}^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \frac{(y_{ij} - y_{i.}^2)}{n - p}.$$

1.5 Intervalle de confiance et tests d'hypothèses

Soit le modèle $y_{ij} = \mu_i + e_{ij}$ où les e_{ij} sont iid suivant une loi centrée de variance σ^2 qui sera supposée $N(0, \sigma^2)$ pour la construction des tests. Dans le cadre général du modèle gaussien, on a montré que les estimateurs des paramètres du modèle sont distribués selon une loi normale, donc :

$$\hat{\mu}_i \sim N(\mu_i, \sigma^2/n_i)$$

On peut en déduire un intervalle de confiance de μ_i de sécurité $1 - \alpha$:

$$\left[\hat{\mu}_i \pm t_{(n-I),(1-\alpha/2)} \sqrt{\frac{\hat{\sigma}^2}{n_i}} \right]$$

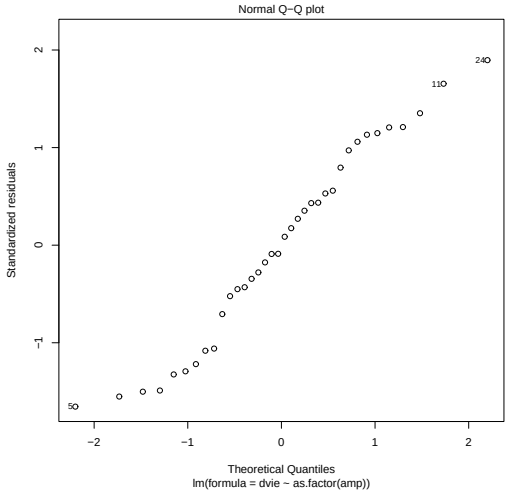


FIG. 8.2 – Données Ampoules : QQplot des résidus.

L'hypothèse $H_0 : \mu_1 = ... = \mu_I$ revient à dire que la moyenne est indépendante du niveau ou encore que le facteur n'a pas d'effet et l'hypothèse alternative H_1 est définie par $\exists(i,k)$ tel que $\mu_i \neq \mu_k$. Cette dernière hypothèse revient à reconnaître un effet ou une influence du facteur sur la variable Y. L'étude de cette hypothèse revient à comparer par un test de Fisher un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres et donc l'égalité des moyennes à celle de la dernière cellule ou à la moyenne générale. Les résultats nécessaires à la construction du test qui en découle sont résumés dans la table d'analyse de variance :

Source de variation	ddl	Somme des carrés	Variance	F
Modèle (inter)	p-1	SSB	MSB=SSB /(p-1)	MSB/MSW
Erreur (intra)	n-p	SSW	MSW=SSW/(n-p)	
Total	n-1	SST		

Avec SSB= $\sum_{i,j}(y_{i.} - y_{..})^2$; SSW= $\sum_{i,j}(y_{ij} - y_{i.})^2$; SST= $\sum_{i,j}(y_{ij} - y_{..})^2$; un point à la place d'un indice veut dire la moyenne sur l'indice considéré ($y_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$).

La statistique F peut s'interpréter comme le rapport de la variabilité inter-groupe sur la variabilité intra-groupe. En effet, le carré moyen du modèle mesure l'écart des moyennes des groupes à la moyenne générale (c'est une mesure de variabilité inter). Le carré moyen résiduel mesure l'écart de chaque unité statistique à la moyenne du groupe (c'est une mesure de la variabilité intra). Si le facteur a un effet sur la variable à expliquer, la variation INTER sera importante par rapport à la variation INTRA.

Dans le cas d'un facteur à 2 classes ($p = 2$), on retrouve un test équivalent au test de Student de comparaison des moyennes de deux échantillons indépendants.

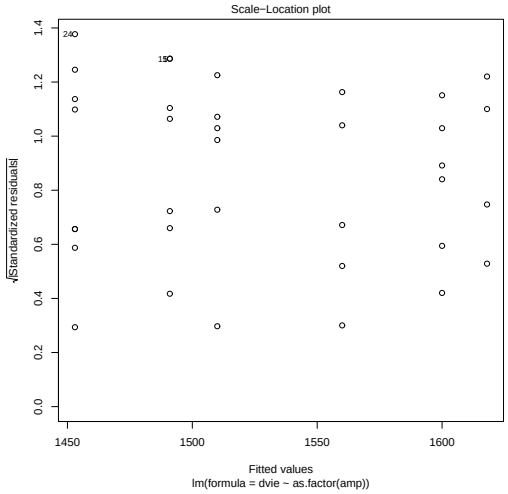


FIG. 8.3 – Données Ampoules : recherche de liaison moyenne-variance.

2 ANOVA à deux facteurs croisés

La considération de deux (ou plus) facteurs explicatifs dans un modèle d'analyse de variance engendre plusieurs complications dont en particulier la notion d'interaction entre variables explicatives. La présence d'interaction atteste du fait que les effets d'un des facteurs dépend des effets de l'autre facteur.

Les niveaux du premier facteur sont notés par un indice i variant de 1 à p, ceux du deuxième facteur par un indice j variant de 1 à q. Pour chaque combinaison, on observe un même nombre $n_{ij} = c > 1$ de répétitions, ce qui nous place dans le cas particulier d'un plan équilibré ou équirépété. Cecntroduit des simplifications importantes dans les estimations des paramètres ainsi que dans la décomposition des variances.

Le modèle général s'écrit :

$$y_{ijk} = \mu_{ij} + e_{ijk}$$

On suppose que les termes d'erreur e_{ijk} sont mutuellement indépendants et de même loi gaussienne. Le modèle d'analyse de variance à deux facteurs s'écrit également de la manière suivante :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

avec les contraintes : $\sum_{i=1}^{p} \alpha_i = \sum_{j=1}^{q} \beta_j = 0, \forall j, \sum_{i=1}^{p} \gamma_{ij} = 0$ et $\forall i, \sum_{j=1}^{q} \gamma_{ij} = 0$.

Lorsque les paramètres d'interaction γ_{ij} sont tous nuls, le modèle est dit $additif$, ce qui correspond à une situation très particulière. Ceci signifie que les écarts relatifs au premier facteur sont indépendants du niveau k du 2ème facteur et vice versa. Dans le cas équirépété, les tests des effets sont résumés dans la table d'analyse de variance suivante :

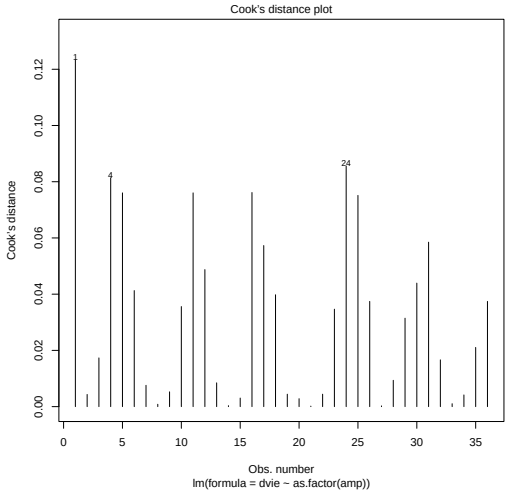


FIG. 8.4 – Données Ampoules : distance de Cook pour détection d'observations aberrantes.

Source de variation	ddl	Somme des carrés	Variance	F
1er facteur	p-1	SS1	MS1=SS1/(p-1)	MS1/MSE
2ème facteur	q-1	SS2	MS2=SS2/(q-1)	MS2/SSE
Interaction	(p-1)(q-1)	SSI	MSI=SSI/(p-1)(q-1)	MSI/MSE
Erreur	n-pq	SSE	MSE=SSE/(n-pq)=$\hat{\sigma}^2$	
Total	n-1	SST		

avec

$$\begin{aligned}
\text{SS1} &= qc\sum_i (y_{i..} - y_{...})^2;\\
\text{SS2} &= pc\sum_j (y_{.j.} - y_{...})^2;\\
\text{SSI} &= c\sum_{ij} (y_{ij.} - y_{i..} - y_{.j.} + y_{...})^2;\\
\text{SSE} &= \sum_{ijk} (y_{ijk} - y_{ij.})^2;\\
\text{SST} &= \sum_{ijk} (y_{ijk} - y_{...})^2.
\end{aligned}$$

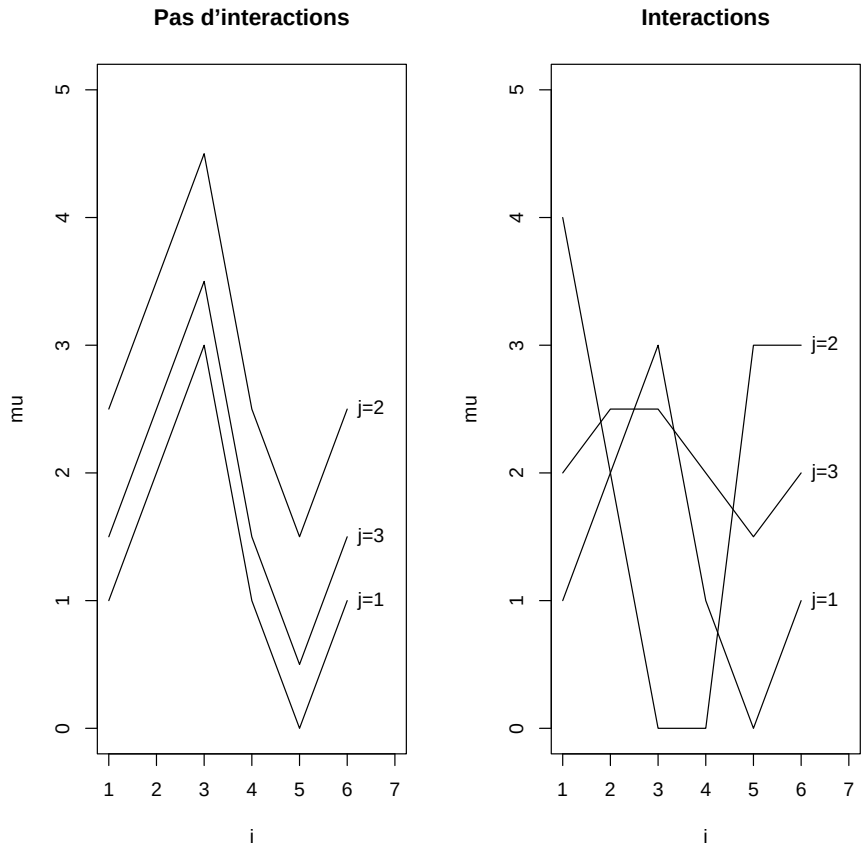
Ici aussi, plusieurs paramétrisations sont possibles et sont en correspondance bijective.

- *Paramétrisation du modèle :* les paramètres sont ici μ_{ij} pour $i = 1, ...p$ et $j = 1, ...q$.
- *Décomposition centrée :* On écrit $\mu_{ij} = \mu + \alpha_i^l + \alpha_j^c + \alpha_{ij}^x$ avec $\sum_i \alpha_i^l = \sum_j \alpha_j^l = \sum_j \alpha_{ij}^x = \sum_i \alpha_{ij}^x = 0$.
- *Décomposition SAS/R.* On écrit : $\mu_{ij} = \mu_{pq} + a_i^l + a_j^c + a_{ij}^x$ avec $a_p^l = a_q^c = a_{pj} = a_{iq} = 0$ (dans SAS), ou $\mu_{ij} = \mu_{11} + a_i^l + a_j^c + a_{ij}^x$ avec $a_1^l = a_1^c = a_{1j} = a_{i1} = 0$ (dans R).

L'estimation et l'inférence sur les paramètres, ainsi que les analyses post-modélisations sont les mêmes que dans le cadre de la régression.

On peut se faire une idée de la présence d'interactions en traçant le graphe des moyennes

empiriques $\bar{y}_{ij}$ en fonction de i, pour chaque j (figure 2). Si les droites sont parallèles, l'effet du premier facteur s'additionne à l'effet du deuxième, il n'y a donc pas d'interaction. Si par contre des droites se croisent, on peut suspecter la présence d'interactions.



3 Analyse de covariance

L'analyse de covariance se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Dans les cas les plus complexes, on peut avoir plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynomiale. Le principe général est toujours d'estimer des modèles $intra-groupes$ et de faire apparaître (tester) des effets différentiels $inter-groupes$ des paramètres des régressions. Ainsi, dans le cas simple où seulement une variable parmi les explicatives est quantitative, nous sommes amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

Prenons un cas simple, le modèle est explicité dans le cas élémentaire où une variable quantitative Y est expliquée par une variable qualitative T à q niveaux et une variable quantitative, appelée encore covariable, X. Pour chaque niveau j de T, on observe n_j valeurs $x_{1j}, ..., x_{n_j j}$ de X et n_j valeurs $y_{1j}, ..., y_{n_j j}$ de Y ; n est la taille de l'échantillon. Le modèle s'écrit :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$$

où les e_{ij} sont iid suivant une loi centrée de variance σ_e^2 qui sera supposée $N(0, \sigma_e^2)$ pour la construction des tests.

Différentes hypothèses peuvent alors être testées par un test de Fisher :

i. Test des interactions : les droites partagent la même pente ;

ii. Test de l'influence du facteur quantitatif ;

iii. Test de la significativité des différences des termes constants.

On commence par tester l'hypothèse (i), si le test n'est pas significatif, on regarde (ii) qui s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si (i) n'est pas significatif, on s'intéresse à (iii) pour juger de l'effet du facteur T.

4 Tests multiples

4.1 Rappels sur les risques de première et seconde espèce

Risque de première espèce : noté α. Le risque de première espèce est le risque de rejeter (avec la règle décision) l'hypothèse H_0 alors qu'en réalité cette hypothèse est vraie.

Risque de seconde espèce : noté β. Le risque de seconde espèce est le risque d'accepter (avec la règle de décision) l'hypothèse H_0 alors qu'en réalité cette hypothèse est fausse.

	Décision	
Réalité	H_0	H_1
H_0	$1-\alpha$	α
H_1	β	$1-\beta$

La quantité $1-\beta$ est une probabilité de bonne décision appelé *puissance* du test.

Remarque : Accepter H_0 ne signifie pas que cette hypothèse est vraie mais seulement que les observations disponibles ne sont pas incompatibles avec cette hypothèse et que l'on n'a pas de raison suffisante de lui préférer l'hypothèse H_1 compte tenu des résultats expérimentaux.

4.2 Tests multiples

Supposons que p moyennes $(m_1, m_2, ..., m_p)$ soient à comparer et que ces p moyennes soient respectivement estimées par : $\bar{X}_1, \bar{X}_2, ..., \bar{X}_p$ et que ces moyennes soient estimées sur des échantillons de tailles respectives $n_1, n_2, ..., n_p$. En comparant les moyennes deux à deux, il faut faire $\frac{p(p-1)}{2}$ comparaisons. Chaque comparaison de 2 moyennes est effectuée en utilisant la règle de décision suivante : si

$$\frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\hat{\sigma}^2(1/n_i + 1/n_j)}} > t_{1-\alpha/2;n_i+n_j-2}$$

alors on rejette l'hypothèse $H_0 : m_i = m_j$.

Si deux comparaisons sont réalisées avec un risque de première espèce de α, il est faux de penser que la décision globale peut être prise avec un risque α. Cela provient du fait qu'une succession de tests de risque α ne permet pas de prendre une décision globale avec ce même risque.

	2	3	4	5	6
Erreur nominale de type I	5%	5%	5%	5%	5%
Erreur globale de type I	5%	12.2 %	20.3 %	28.6%	36.6%

La méthode de Bonferroni est une méthode qui ne permet pas un strict contrôle de α mais en revanche elle en donne une majoration. L'idée de Bonferroni est de se placer dans le pire des cas (pour α). Par exemple si on a $p = 5$ moyennes à comparer, il faut effectuer 10 comparaisons. Pour avoir un risque global α, il faut que chacune des 10 comparaisons soit effectuée avec un risque $\alpha' = \alpha/10$.

En pratique, Bonferroni fournit une liste de gènes différentiellemnt exprimés dans laquelle on contrôle le nombre de faux positifs. Quand le nombre des gènes est grand, cette liste est souvent vide.

A l'opposé, le LSD (Least Square Difference), c'est à dire le test de Student sans correction, est le plus laxiste : il va détecter des gènes différentiellement exprimés qui en fait ne le sont pas.

En général, on présente ces taux d'erreurs dans le tableau suivant :

Réalité	*Décision*		
	H_0 vraie	H_1 vraie	Total
H_0 vraie	U	V	m_0
H_1 vraie	T	S	m_1
	W	R	m

où m tests sont effectués. Pour une analyse de biopuces dans laquelle on teste les effets différentiels de m gènes, m_1 est le nombre de gènes déclarés différentiellement exprimés, alors que R est le nombre réel (mais inconnu) de gènes différentiellement exprimés.

Diverses méthodes sont proposées pour contrôler ces divers taux d'erreurs.

Le FWER (Family Wise Error Rate) représente la probabilité d'effectuer au moins une erreur de première espèce sur l'ensemble des comparaisons :

$$P[V \geq 1] = m_0\alpha.$$

On prend donc un seuil nominal de $\alpha' = \alpha/m_0$.

Au même titre que Bonferroni, plus il y a de tests (soit de gènes à tester), moins on rejette H_0 (moins de gènes déclarés différentiellement exprimés). La notion suivante est très utile pour pallier à cet inconvénient.

La FDR (False Discovery Rate) contrôle l'espérance du taux de faux positifs, ou le nombre de faux positifs parmi les différences déclarées significatives. Pratiquement, on ordonne les m p-values des m tests (les gènes) et on recherche le plus haut rang k des p-values tel que $p - value(k) \geq \alpha k/m$..

Il existe d'autres approches récentes ou en cours de développement pour contrôler la FDR positive, le nombre moyen d'erreurs, etc ...

5 Modèle linéaire mixte gaussien

Dans les modèles linéaires classiques (cf. chapitre 7), toutes les variables explicatives sont supposées déterministes. La plupart du temps, cette situation simple ne permet pas de répondre de façon effective aux questions posées. Pour illustrer cette insuffisance, prenons quelques exemples.

5.1 Exemple 1

Supposons que l'on cherche à comparer 2 traitements A et B ; 4 élevages ont été sélectionnés pour participer à cet essai. Dans chaque élevage un échantillon d'animaux a été tiré au hasard, une moitié des animaux de l'échantillon ont reçu le traitement A et l'autre moitié le traitement B. Les données brutes ont été analysées et les analyses ont montré que le traitement B a une plus grande efficacité que le traitement A. Que peut-on conclure ?. Pour répondre convenablement à cette question, il est nécessaire de préciser la nature du facteur élevage :

- si les élevages ont été choisis, le facteur élevage est un facteur fixe et les résultats de l'analyse ne peuvent pas être extrapolés à d'autres élevages,
- si les élevages ont été tirés au hasard parmi tous les élevages susceptibles d'utiliser ces produits, le facteur élevage est alors un facteur aléatoire et les résultats de cette analyse peuvent être extrapolés aux autres élevages.

Dans une analyse de variance, on s'intéresse à l'effet particulier de chacun des niveaux de la variable explicative sur la variable à expliquer. Cette façon de procéder suppose que l'on introduise dans le modèle tous les niveaux du facteur susceptibles d'avoir un intérêt. Mais cela n'est pas toujours possible. Par exemple, si on s'intéresse aux performances au champ d'une variété de blé, ou aux performances de croissance (ou production laitière) des animaux d'une race particulière, il est impossible de tester ces performances sur tous les champs ou animaux possibles. On peut également vouloir s'intéresser à l'effet d'un régime alimentaire sur la croissance des porcs, on ne pourra pas le tester sur tous les porcs. A chaque fois, pour réaliser l'expérience, il faudra prendre quelques individus (ici, des champs ou des porcs) et chercher à étendre les résultats obtenus à la population entière. Si on suppose que les individus ont été tirés au hasard dans la population, on ne s'intéresse plus à l'effet particulier associé à tel individu particulier, mais à la distribution de l'ensemble des effets possibles. L'effet associé à l'individu n'est plus un effet fixe mais devient un effet aléatoire et il faut en tenir compte dans l'analyse. Le modèle linéaire étudié contient un mélange d'effets fixes et d'effets aléatoires, on parle alors de modèle linéaire mixte. Le modèle linéaire mixte constitue une extension du modèle linéaire classique. D'une manière générale, on pourra y faire appel chaque fois que l'on désirera *étendre à une population toute entière des résultats obtenus sur quelques individus pris au hasard dans cette population.*

5.2 Exemple 2

On a relevé les durées de gestation de 16 filles de 30 taureaux qui avaient été tirés au sort dans la population devant être étudiée. On voudrait savoir dans quelle mesure la durée de gestation est un caractère héréditaire. On considère que ce caractère se transmet (aussi) par les pères : un taureau ayant de bons gènes les transmettra à ces filles, qui seront donc meilleures en moyenne que des vaches descendantes de "mauvais" taureaux. Il s'agit de répondre, grâce à un échantillon comportant peu de taureaux, à une question concernant toute la population. Pour pouvoir étendre les résultats obtenus sur l'échantillon, il faut que celui-ci soit représentatif de toute la population et donc qu'il ait été obtenu par tirage au sort (indépendants et équiprobables). Il en découle que les taureaux de l'échantillon sont aléatoires et leurs effets sur leurs descendants sont a fortiori aléatoires.

Le modèle s'écrira

$$y_{ij} = \mu + a_i + e_{ij} \quad j = 1, ...16 \quad i = 1, ...30$$

où y_{ij} représente la durée de gestation de la fille j du père i, μ est la moyenne générale, a_i est l'effet du père i, supposé aléatoire puisque le père est un individu tiré aléatoirement, e_{ij} est le résidu.

On suppose les distributions suivantes :

$$\begin{aligned} a_i &\sim N(0, \sigma_a^2) \\ e_{ij} &\sim N(0, \sigma_e^2) \end{aligned}$$

Les a_i et les e_{ij} sont supposés mutuellement indépendants.

On appelle σ_a^2 et σ_e^2 les composantes de la variance. La quantité $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$ est la part de variance "génétique" de la variance totale.

Il faut bien comprendre que $\mu + a_i$ n'est pas l'espérance de y, mais son espérance conditionnelle : $\mathrm{E}(y_{ij}|a_i) = \mu + a_i$. De la même manière, la variance conditionnelle de y vaut $\mathrm{Var}(y_{ij}|a_i) = \sigma_e^2$.

5.3 Exemple 3

Une compagnie pharmaceutique veut tester les performances d'une méthode de spectroscopie (NIR = Near Infrared Reflectance) permettant de déterminer le contenu en substance active de comprimés. La méthode utilisée en routine (HPLC) est plus coûteuse et laborieuse. Pour cela, 10 comprimés ont été tirés au hasard et les 2 méthodes ont été utilisées sur chacun des comprimés.

Comprimé	HPLC	NIR	Différence
1	10.4	10.1	0.3
2	10.6	10.8	-0.2
3	10.2	10.2	0.0
4	10.1	9.9	0.2
5	10.3	11.0	-0.7
6	10.7	10.2	0.2
7	10.3	10.2	0.1
8	10.9	10.9	0.0
9	10.1	10.4	-0.3
10	9.8	9.9	-0.1
	$\bar{y}_1 = 10.34$	$\bar{y}_2 = 10.36$	$\bar{d} = 0.05$
	$s_1 = 0.3239$	$s_2 = 0.4033$	$s_d = 0.2953$

Text de Student

L'analyse la plus simple consiste à considérer les données comme un échantillon apparié et d'utiliser le test de Student correspondant. La moyenne et l'écart-type des différences d sont égaux à $\bar{d} = 0.05$ et $s_d = 0.2953$ respectivement. La statistique du test de Student vaut $t = \frac{\bar{d}}{SE_{\bar{d}}} = \frac{-0.05}{0.2953/\sqrt{(10)}} = -0.535$ qui donne une p-value égale à 0.61. On en conclut qu'il n'y a pas différence significative entre les 2 méthodes de mesure. L'intervalle de confiance à 95% du biais de la méthode vaut $\bar{d} \pm t_{0.975}(9) SE_{\bar{d}} = -0.05 \pm 0.21$.

Sous R, par exemple, on obtient

```
> t.test(d)
One-sample t-Test
data : d
t = -0.5354, df = 9, p-value = 0.6054
alternative hypothesis : true mean is not equal to 0
95 percent confidence interval :
-0.2612693 0.1612693
sample estimates :
```

```
mean of x
-0.05
```

Anova

Le modèle d'analyse de variance pour cette situation s'écrit :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2)$$

avec μ la moyenne générale, α_i l'effet du $i$ème comprimé, β_j l'effet de la méthode j.

La statistique du test F de Fisher est égal au carré de la statistique de Student : $F = t^2 = (-0.535)^2 = 0.29$.

L'estimation de l'écart-type résiduel est donnée par $\hat{\sigma} = 0.209 = \sqrt{2}s_d$. L'incertitude sur la moyenne des différences est donnée par $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\sigma^2(1/10 + 1/10)} = 0.0934$, exactement comme dans l'approche simple.

Si maintenant on s'intéresse à la précision de la valeur moyenne pour la méthode NIR, l'ANOVA donne $SE(\bar{y}_1) = \hat{\sigma}/\sqrt{10} = 0.066$. D'un autre côté, si on considère l'échantillon des 10 mesures de NIR, on obtient $s_1 = 0.4012$, et donc $SE(\bar{y}_1) = s_1/\sqrt{10} = 0.127$, valeur très différente de 0.066.

L'ANOVA sous-estime beaucoup l'incertitude sur l'estimation de l'effet moyen de NIR. C'est ainsi parce que la variance σ^2 mesure la variabilité résiduelle après que les effets des comprimés ont été corrigés. La différence conceptuelle entre les 2 approches est que l'ANOVA considère que les 10 comprimés n'ont pas été tirés au hasard, alors que la seconde (échantillon des 10 mesures NIR) si. L'ANOVA n'est valide que si l'on s'intéresse aux effets spécifiques des 10 comprimés.

L'idée du modèle mixte est de combiner les 2 approches : utiliser un modèle linéaire et y considérer certains facteurs comme aléatoires.

Modèle mixte

On considère maintenant le modèle linéaire mixte

$$y_{ij} = \mu + a_i + \beta_j + \varepsilon_{ij}$$

où a_i est l'effet aléatoire du $i$ème comprimé. Ses effets sont supposés indépendants et identiquement distribués : $a_i \sim N(0, \sigma_a^2)$.

On peut montrer que

	ANOVA	Modèle mixte
$E(y_{ij})$	$\mu + \alpha_i + \beta_j$	$\mu + \beta_j$
$Var(y_{ij})$	σ^2	$\sigma^2 + \sigma_a^2$
$Cov(y_{ij}, y_{i'j'}), j \neq j'$	0	σ_a^2 si $i = i'$et 0 sinon

L'écart-type attendu de la moyenne des valeurs de NIR vaut dans le cadre du modèle linéaire mixte :

$$SE(\bar{y}_1) = \sqrt{\hat{\sigma}_a^2 + \hat{\sigma}^2}/\sqrt{10} = 0.115,$$

ce qui est conforme à ce que l'on attendait.

Notons que les mesures HPLC et NIR sur un même comprimé sont corrélées dans le modèle mixte. Un modèle linéaire mixte permet donc aussi de prendre en compte de façon simple des corrélations entre observations.

5.4 Définition

Un modèle linéaire mixte se définit comme un modèle linéaire dans lequel toute ou partie des paramètres associés à certaines unités expérimentales sont traités comme des variables aléatoires du fait de l'échantillonnage de ces unités dans une population plus large.

6 Exemple : nutrition chez la souris

6.1 Analyses de variance et modèle mixte

6.2 Principe des analyses de variance

L'analyse de variance (ANOVA) permet d'apprécier l'effet d'une ou plusieurs variables qualitatives (les facteurs) sur une variable quantitative (la variable réponse, ici le niveau d'expression des gènes). Dans le domaine de l'analyse transcriptomique, cette approche a été largement développée, en particulier par Kerr *et al.* (2000). Pour l'analyse de nos données, un modèle d'ANOVA à trois facteurs (génotype, régime, gène) permet de mettre en évidence des effets d'interaction d'ordre 3 très significatifs à l'aide du test de Fisher. Cela signifie qu'il existe des gènes régulés simultanément par le régime et le génotype, les effets du régime et du génotype étant non additifs. Le modèle d'ANOVA considéré s'écrit

$$y_{ijkl} = g_i + r_j + G_k + gr_{ij} + gG_{ik} + rG_{jk} + grG_{ijk} + e_{ijkl} \tag{8.2}$$

où y_{ijkl} représente le logarithme du niveau d'expression du gène k ($k = 1, ..., 120$), pour le régime j ($j = 1, ..., 5$) et le génotype i ($i = 1, 2$), mesuré chez la souris l ($l = 1, ..., 4$) ; g_i représente l'effet du génotype i, r_j celui du régime j, G_k celui du gène k, gr_{ij} représente l'effet de l'interaction du génotype i et du régime k, gG_{ik} l'effet de l'interaction du génotype i et du gène k, rG_{jk} l'effet de l'interaction du regime j et du gène k et grG_{ijk} représente l'interaction d'ordre 3 combinant le génotype i, le régime j et le gène k. On suppose que les résidus e_{ijkl} du modèle sont indépendants et identiquement distribués suivant une loi normale de moyenne nulle et de variance σ^2. L'écriture d'un tel modèle suppose que les gènes sont tous de même variabilité. Cette hypothèse est discutable (en effet la figure 2.9 montre clairement quelques gènes fortement variables) ; nous verrons par la suite comment lever cette hypothèse. À partir de ce modèle, on peut estimer les effets principaux des 120 gènes, effectuer des comparaisons de moyennes à l'aide du test de Fisher, puis opérer des corrections pour des tests multiples afin de repérer les gènes surexprimés ou sous-exprimés selon le génotype et le régime.

Dans cette séquence de tests, les variances des gènes sont supposées égales, contrairement aux tests de Student de comparaison de moyennes par régime et génotype pour un gène fixé. Ce dernier cas revient à écrire une modèle d'ANOVA par gène, sous la forme suivante

$$y_{ijl} = g_i + r_j + gr_{ij} + e_{ijl} \tag{8.3}$$

où les notations utilisées ici sont identiques à celles du modèle (8.2). Ici, il est nécessaire de faire autant d'analyses de variance que de gènes étudiés (soit 120 dans notre exemple) mais nous disposerons d'une variance estimée par gène. Toutefois une telle analyse n'est pas toujours recommandée car en règle générale le nombre d'observations par gène est très faible, ce qui conduit à des estimations de variance très peu précises. Notons cependant que ces 120 analyses conduisent à 120 estimations des 10 effets $genotype_i \times regime_j$. Un modèle équivalent, mais utilisant simultanément l'ensemble des données pour estimer les paramètres, s'écrit comme le modèle (8.2) en posant

$$\text{var}(e_{ijkl}) = \sigma^2_{e,k}. \tag{8.4}$$

D'autre part, entre le modèle (8.2), supposant toutes les variances des gènes égales, et le modèle (8.4) supposant une variance différente pour chaque gène, il est possible d'ajuster un modèle intermédiaire prenant en compte les hétérogénéités de variances de l'expression des gènes, en définissant simplement des groupes de gènes de variabilité homogène (Robert-Granié *et al.*, 1999 ; Foulley *et al.*, 2000 ; San Cristobal *et al.*, 2002). Ainsi, sur les 120 gènes analysés, un histogramme des variances nous a conduit à définir trois groupes de gènes ayant des variabilités très différentes : un groupe contenant les gènes `FAS`, `G6Pase`, `PAL` et `S14`, présentant des variabilités résiduelles importantes (variances supérieures à 0.02) ; un deuxième groupe à variabilité modérée (variances comprises entre 0.009 et 0.02), comprenant les gènes `CYP2c29`, `CYP3A11`, `CYP4A10`, `CYP4A14`, `CYP8b1`, `GSTmu`, `GSTpi2`, `L-FABP`, `Lpin`, `Lpin1`, `TRa` et `cHMGCoAS` ; enfin un dernier groupe à faible variabilité (variances inférieures à 0.009), contenant l'ensemble des autres gènes. À partir de ces trois groupes de gènes, nous pouvons construire un modèle dont la variance dépend de cette nouvelle variable à trois classes. Le modèle s'écrit encore comme les modèles (8.2) et (8.4) en posant cette fois

$$\text{var}(e_{ijkl}) = \sigma^2_h, \tag{8.5}$$

où $h = \{1, 2, 3\}$ représente l'indice d'hétérogénéité de variance.

Nous pouvons ainsi comparer les gènes différentiellement exprimés selon les 3 modèles :

- Modèle (8.2), modèle d'ANOVA avec une unique variance pour l'ensemble des gènes ;
- Modèle (8.4), modèle d'ANOVA avec une variance différente par gène ;
- Modèle (8.5), modèle d'ANOVA avec trois groupes de variances différentes.

Notons que le modèle (8.4) implique l'estimation de 120 variances différentes, alors que le modèle (8.5) ne nécessite l'estimation que de trois paramètres de variances ; ce dernier est donc beaucoup plus économe en nombre de paramètres à estimer. Enfin, d'un point de vue technique et opérationnel, la mise en oeuvre de ces modèles peut être réalisée en utilisant la fonction `lme` du logiciel statistique R ou la procédure `mixed` du logiciel SAS.

6.3 Synthèse des tests multiples

L'objectif de l'analyse statistique est de déterminer quels sont les gènes différentiellement exprimés entre les 2 génotypes et les 5 régimes. Quelle que soit la méthode statistique utilisée, il existera une probabilité non nulle (risque de première espèce α) de détecter des faux positifs (gènes déclarés différentiellement exprimés alors qu'ils ne le sont pas) et une autre probabilité non nulle (risque de deuxième espèce β) de ne pas être capable de détecter des gènes réellement différentiellement exprimés (faux négatifs). Il est bien entendu souhaitable de minimiser ces deux probabilités d'erreur sachant que, toutes choses égales par ailleurs, la seconde augmente quand la première diminue et réciproquement. Le test de Student est couramment utilisé pour tester l'égalité de deux moyennes (l'hypothèse nulle étant de considérer que les moyennes des intensités des signaux d'un gène donné dans chaque condition 1 et 2 sont égales). Ainsi, quand la statistique de Student excède un certain seuil (dépendant du risque de première espèce α choisi, généralement 5%), les niveaux d'expression du gène étudié entre les deux populations testées sont considérées comme significativement différentes. Lorsque l'on souhaite tester plus de deux conditions, le test de Fisher, qui est une extension du test de Student, est utilisé. L'hypothèse nulle constitue l'absence d'expression différentielle d'un gène entre les diverses conditions et l'hypothèse alternative montre une différence d'expression.

Enfin, prendre un risque de 5% dans une expérimentation où 10 000 gènes, par exemple, sont étudiés simultanément peut conduire à obtenir 500 faux positifs, ce qui est parfaitement inacceptable. C'est pourquoi ont été proposées des modifications du test de Student adaptées à l'analyse du transcriptome (méthodes de Bonferroni, FWER, FDR...). Le lecteur souhaitant des détails sur ces approches peut se référer, par exemple, à Benjamini & Hochberg (1995), Bland & Altman (1995), Dudoit *et al.* (2002) ou Speed (2003).

La méthode de Bonferroni, rappelons le, est une méthode qui ne permet pas un strict contrôle de α, mais qui en donne une majoration. Pour avoir un risque global α, il faut que chacune des p comparaisons soit effectuée avec un risque $\alpha' = \alpha/p$. En pratique, Bonferroni fournit une liste de gènes différentiellemnt exprimés dans laquelle on contrôle le nombre de faux positifs. Mais, lorsque le nombre des gènes est grand, cette liste est souvent vide.

En général, on présente ces taux d'erreurs dans le tableau 4.2.

Pour revenir à notre étude, à partir de chaque modèle proposé dans le paragraphe précédent, nous pouvons rechercher les gènes différentiellement exprimés entre les deux génotypes à régime fixé (120 comparaisons pour chacun des 5 régimes) ou entre régime à génotype fixé (1200 comparaisons par génotype), ce qui conduit à effectuer 3000 comparaisons. Le tableau 8.2 présente le nombre de gènes sélectionnés selon les trois modèles considérés et selon le test ou l'ajustement utilisée (Student, Bonferroni, Benjamini-Hochberg qui correspond à l'approche FDR).

TAB. 8.2 – *Nombre de gènes sélectionnés selon le modèle et le test utilisés.*

Tests	Modèle (8.2)	Modèle (8.4)	Modèle (8.5)
Student à 5%	85	103	97
Student à 1%	55	65	67
Benjamini-Hochberg à 5%	44	56	59
Benjamini-Hochberg à 1%	35	40	38
Bonferroni à 5%	53	62	65
Bonferroni à 1 pour mille	18	19	21

On peut remarquer que le nombre de gènes sélectionnés est peu différent selon le modèle utilisé et que, globalement, les trois modèles sélectionnent le même groupe de gènes. Les petites différences sont principalement liées à l'ordre de sélection de ces gènes.

D'autre part, on peut, à partir de critères de sélection de modèle tels que le critère d'Akaike (AIC ; Akaike, 1974) ou le critère de Schwarz (BIC ; Schwarz, 1978), ou encore en effectuant un test du rapport de vraisemblance, choisir le modèle le plus adéquat.

Le tableau 8.3 présente les valeurs des critères AIC et BIC pour les trois modèles mis en compétition.

TAB. 8.3 – *Valeurs des critères AIC et BIC.*

Modèles	-2AIC	-2BIC
(8.2)	-6576.9	-6570.7
(8.4)	-6946.6	-6612.1
(8.5)	-7044.5	-7036.2

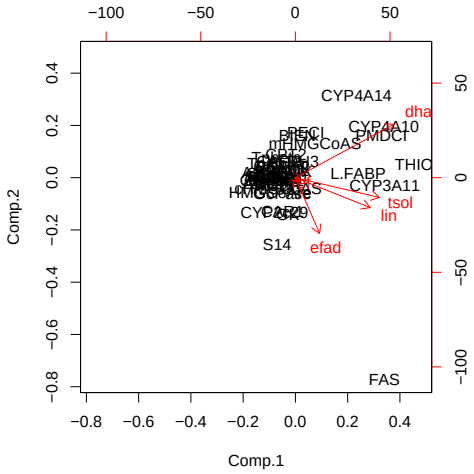


FIG. 8.5 – *Souris : représentation sur le premier plan principal de l'ACP du logarithme des p-value des gènes différentiellement exprimés entre les deux génotypes à régime fixé.*

Le meilleur modèle est celui pour lequel les valeurs des critères -2AIC ou -2BIC sont les plus petits. Dans les deux cas, il s'agit du modèle (8.5).

Le test du rapport de vraisemblance consiste, quant à lui, à comparer deux modèles emboîtés (par exemple, (8.2) *vs* (8.4)) ; l'hypothèse nulle considérée suppose alors que toutes les variances sont égales. La statistique du rapport de vraisemblance nécessite de calculer la différence entre les logarithmes des vraisemblances sous chacun des deux modèles. Sous l'hypothèse nulle, cette statistique suit asymptotiquement une loi de khi-deux dont le nombre de degré de liberté est égal à la différence des nombres de paramètres à estimer sous chacun des deux modèles considérés. Si nous effectuons ces différents tests du rapport de vraisemblance ((8.2) *vs* (8.4), (8.2) *vs* (8.5), (8.4) *vs* (8.5)), il en ressort que le modèle (8.5), avec trois groupes de variances, est encore le meilleur.

À partir de ce modèle (8.5), on peut estimer les différents effets du modèle, et s'intéresser aux différences d'expression des gènes entre génotypes à régime fixé ou encore aux différences d'expression des gènes entre régimes à génotype fixé.

En raison de la multiplicité des tests, la correction proposée par Benjami & Hochberg (1995) a été utilisée. Lorsque nous considérons les différences d'expression des gènes entre génotypes à régime fixé, l'hypothèse nulle représente l'absence d'expression différentielle d'un gène entre les deux génotypes. On peut visualiser l'ensemble des résulats des *p-values* de ces différents tests en effectuant une ACP centrée sur le logarithme des *p-values*, les gènes en ligne et les régimes en colonne. La figure 8.5 présente le premier plan principal des gènes différentiellement exprimés entres les deux génotypes à régime fixé. Les deux premiers axes principaux représentent 93% de la variance totale. Pour des raisons de visibilité, les résultats sont présentés sur les 59 gènes différentiellement exprimés selon le modèle (8.5) et en utilisant la correction de Benjamini & Hochberg à 5% (Tab. 8.2).

On observe que les gènes `CYP3A11`, `CYP4A10`, `CYP4A14`, `L.FABP`, `PMDCI` et `THIOL` différencient les deux génotypes pour les régimes `dha`, `lin` et `tsol`. Certains de ces gènes présentent des expressions constitutives différentielles entre les souris des deux génotypes. De plus ces gènes sont régulés positivement par ces trois régimes riches en acides gras polyinsaturés d'une

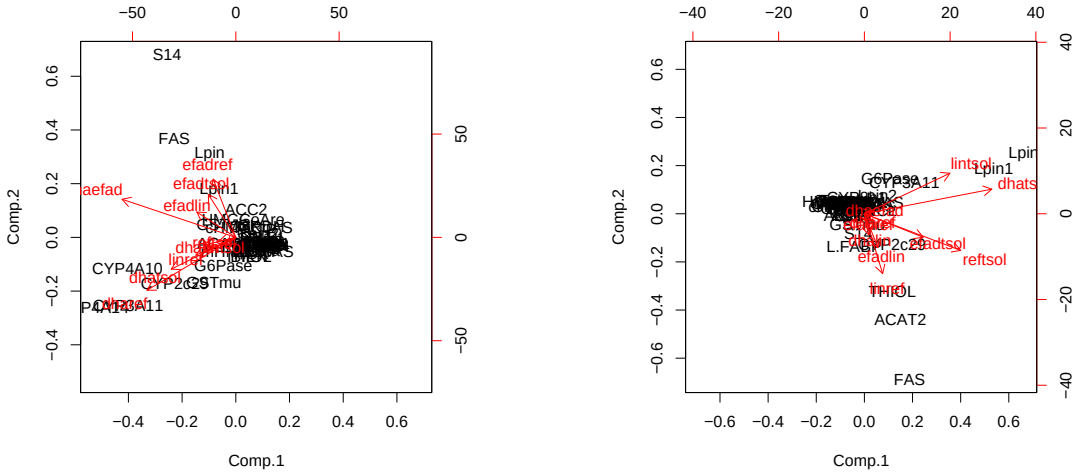


FIG. 8.6 – *Souris : représentation sur le premier plan principal de l'ACP du logarithme des p-value des gènes différentiellement exprimés entre les régimes pour le génotype* `WT` *à gauche et* `PPAR` *à droite.*

famille particulière (*Oméga 3* pour `dha` et `lin` et *Oméga 6* pour `tsol`) chez les souris `WT` alors que la régulation de plusieurs de ces gènes est altérée chez les souris PPAR. Les gènes `mHMGCoAS`, `PECI` et `BIEN` apparaissent dans le contraste entre génotypes pour le régime `dha`, alors que les gène `S14` et `FAS` apparaissent pour le régime `efad`. Les souris des deux génotypes présentent là encore des régulations différentielles de ces gènes, soulignant ainsi le rôle du récepteur PPARα dans ces modulations d'expression provoquées par les régimes alimentaires.

La même approche sur les effets différentiels entre couples de régimes, à génotype fixé, est réalisée. Les représentations de la figure 8.6 présentent le premier plan principal des gènes différentiellement exprimés entre régime pour le génotype `WT` (à gauche) et pour le génotype `PPAR` (à droite). Les deux premiers axes, pour chacune des figures, représentent respectivement 79% et 78% de la variance totale. Les gènes `Lpin` et `Lpin1` apparaissent dans des contrastes impliquant le régime `efad` pour le génotype `WT`, et le régime `tsol` pour le génotype `PPAR`. Le gène `CYP3A11` est impliqué dans le régime `dha`, quel que soit le génotype. Les gènes `FAS` et `S14` apparaissent dans les contrastes impliquant le régime `efad` pour le génotype `WT`, alors que le même gène `FAS` apparaît dans les contrastes impliquant le régime `ref` pour le génotype `PPAR`. L'ensemble de ces résultats confirme les résultats obtenus pour l'ACP.

6.4 Modèle mixte

Les souris étant issues d'une lignée consanguine, elles ont été considérées dans un premier temps comme des répétitions indépendantes et identiquement distribuées. Cependant, à l'aide d'un modèle linéaire mixte, chaque souris peut être considérée comme un tirage aléatoire dans une population plus large de souris. Le modèle linéaire mixte mis en oeuvre s'écrit

$$y_{ijkl} = g_i + r_j + G_k + gr_{ij} + gG_{ik} + rG_{jk} + grG_{ijk} + souris_l + e'_{ijkl}, \qquad (8.6)$$

où $souris_l$ représente l'effet aléatoire de la souris l, avec $souris_l \sim \mathcal{N}(0, \sigma_s^2)$, les différentes réalisations étant indépendantes, et e_{ijkl} représente les résidus, avec $e'_{ijkl} \sim \mathcal{N}(0, \sigma_e^2)$, les résidus étant indépendants entre eux et indépendants de l'effet aléatoire souris.

Dans ce cas, les estimations des composantes de la variance sont pour la variance « souris » de 0.001 et pour la variance résiduelle de 0.007. La variabilité indivuelle est très faible. La variance des observations est identique à celle obtenue à l'aide d'une ANOVA (modèle à effets fixes) puisque nous sommes dans le cadre d'un plan équilibré et que la méthode d'estimation pour le modèle mixte est la méthode du maximum de vraisemblance restreinte (REML). Nous pouvons également étendre ce modèle aux cas de variances résiduelles hétérogènes, comme c'était le cas dans le modèle (8.5).

L'application du modèle linéaire mixte est beaucoup plus appropriée dans le cas où les variabilités dues à la technique, à la diversité génétique, aux gènes de la biopuce, ont un intérêt. C'est le cas dans l'étude transcriptomique décrite dans Bonnet *et al.* (2004) dans laquelle le logarithme du signal est modélisé en fonction des facteurs membrane, truie, aiguille (ou bloc), jour d'hybridation, et des covariables logarithme de l'intensité du bruit de fond et de l'hybridation en sonde vecteur. Après une étape de choix de modèle (à l'aide du test de Fisher), le modèle linéaire mixte permet d'appréhender et de quantifier la part de variabilité due aux différentes sources de variation. La part de variabilité due à la diversité génétique représente 8%, celle due à la technique 4% et celle due aux gènes 75%. Toute inférence basée sur ce modèle sera valide pour tout animal, toute membrane... car l'échantillonnage des animaux, des membranes... de cette étude, dans une population plus large d'animaux, membranes... est pris en compte. Considérer les membranes (par exemple) comme effets fixes dans ce modèle aurait entraîné des conclusions valides uniquement sur les membranes de l'expérience. De plus, une structure de covariance non diagonale est prise en compte par ce modèle mixte puisque deux signaux d'une même membrane seront corrélés, la corrélation étant égale à $\sigma^2_{membrane}/\sigma^2_{totale}$.

En guise de conclusion

Ce document explore déjà une grande variétés d'approches statistiques en tâchant de les adapter au mieux aux caractéristiques très particulières des données d'expression. Néanmoins, beaucoup d'autres approches sont encore possibles mais, faute de place ou de compétences, elles ont été laissées de côté comme les modèles de mélange. Baccini et col. (2005) proposent d'ailleurs, sur le même jeu de données, d'autres approchent à base de discrimination : plutôt que de rechercher quelles sont les gènes dont on peut dire qu'ils sont significativement différenciellement exprimés, on peut rechercher un sous-ensemble de gènes permettant la construction d'un meilleur modèle de prédiction des groupes d'échantillons. Par exemple quels sont les gènes qui permettent de discriminer au mieux les deux génotypes de souris ou encore, plus difficle, les différents régimes.

Déjà au niveau de ce cours, l'étude déroulée sur l'exemple des données de nutrition permet de mettre en exergue le fait qu'il n'existe pas "une" méthode unique qui permettraient de traiter des données d'expression. La question "Quelle méthode dois-je utiliser pour traiter mes données d'expression ?" n'a pas de sens. En revanche, à une question précise du type "Puis-je effectuer une partition des gènes ?", une méthode statistique (ici la classification) peut apporter des éléments de réponses par des sorties numériques et/ou des graphiques mais la réponse précise à la question ne peut être apportée que par le praticien qui sait interpréter les résultats statistiques en termes biologiques. Finalement, chaque méthode ou technique statistique associée à différents jeux d'options (comme différentes métriques) fournit différentes *optiques* pour appréhender les données. Seule une forte interaction entre le biologiste et le statisticien est susceptible d'aboutir à des interprétations cohérentes des résultats afin d'ébaucher des nouvelles pistes de recherche pertinentes et prometteuses.

Bibliographie

[1] A. BACCINI et P. BESSE : Data mining : 1. exploration statistique, 2000. www.ups-tlse.fr/Besse/enseignement.html.

[2] Y. BENJAMINI et Y. HOCHBERG : Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, (85):289–300, 1995.

[3] P. BESSE : Pratique de la modélisation statistique, 2000. www.ups-tlse.fr/Besse/enseignement.html.

[4] P. BESSE : Data mining : 2. modélisation statistique et apprentissage, 2004. www.ups-tlse.fr/Besse/enseignement.html.

[5] J. BLAND et D. ALTMAN : Multiple significance tests : the bonferroni method. *British medical Journal*, (310), 1995.

[6] A. BONNET, F. BENNE, C. DANTEC, N. GOBERT, P.O. FRAPPART, M. SANCRISTOBAL, F. HATEY et G TOSSER-KLOPP : Identification of genes and gene networks involved in pig ovarian follicular development, by using c-dna microarrays. *In III International Workshop on the Development and Function of Reproductive organs*, 2004.

[7] L. BREIMAN : Random forests. *Machine Learning*, 45:5–32, 2001.

[8] L. BREIMAN, J. FRIEDMAN, R. OLSHEN et C. STONE : *Classification and regression trees*. Wadsworth & Brooks, 1984.

[9] P.B. BROCKHOFF : Statistical models with random effects, 2004. www.dina.dk/ per/Netmaster/courses/st113/Intro/index.html.

[10] P. DALGAARD : *Introductory Statistics with R*. Springer, 2003.

[11] S. DRAGHICI : *Data Analysis Tools for DNA Microarrays*. Mathematical Biology and Medicine Series. Chapman & Hall/CRC, 2003.

[12] S. DUDOIT, Y. YANG, T. SPEED et M. CALLOW : Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica Sinica*, pages 111–139, 2002.

[13] J.J. FARAWAY : Practical regression and anova using r, 2002. www.stat.lsa.umich.edu/ faraway/book/.

[14] J.-L. FOULLEY, F. JAFFREZIC et C. ROBERT-GRANIÉ : Em-reml estimation of covariances parameters in gaussian mixed models for longitudinal data analysis. *Genetics Selection Evolution*, (32):129–141, 2000.

[15] T. HASTIE, R. TIBSHIRANI et J FRIEDMAN : *The elements of statistical learning : data mining, inference, and prediction*. Springer, 2001.

[16] I. JOLLIFFE : *Principal Component Analysis*. Springer-Verlag, 2nd edition édition, 2002.

[17] L. KAUFMAN et J. ROUSSEEUW, P. : *Finding groups in data*. Wiley, 1990.

[18] Churchill G. KERR K., Martin M. : Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, pages 819–837, 2000.

[19] S.S. LEE, T. PINEAU, J. DRAGO, E.J. LEE, J.W. OWENS, D.L. KROETZ, P.M. FERNANDEZ-SALGUERO, H. WESTPHAL et F.J. GONZALEZ : Targeted disruption of the alpha isoform of the peroxisome proliferator-activated receptor gene in mice results in abolishment of the pleiotropic effects of peroxisome proliferators. *Molecular and Cellular Biology*, 15(6):3012–22, 1995.

[20] K.V. MARDIA, J.T. KENT et J.M. BIBBY : *Multivariate Analysis*. Academic Press, 1979.

[21] P.G.P. MARTIN et COL. : A nutrigenomic approach in mice reveals new aspects of pparα-deficient phenotype with important implications in pharmacology. *Gene Expression*, à paraître.

[22] P.G.P. MARTIN, F. LASSERRE, C. CALLEJA, A. VAN ES, A. ROULET, D. CONCORDET, M. CANTIELLO, R. BARNOUIN, B. GAUTHIER et T. PINEAU : Transcriptional modulations by rxr agonists are only partially subordinated to pparalpha signaling and attest additional, organ-specific, molecular cross-talks. *Gene Expression*, à paraître.

[23] G.J. MCLACHLAN, K.-A. DO et C. AMBROISE : *Analysing microarray gene expression data*. Wiley, 2004.

[24] C. ROBERT-GRANIÉ, B. BONAITI, D. BOICHARD et Barbat A. : Accounting for variance heterogeneity in french dairy cattle genetic evaluation. *Livestock Production Science*, (60): 343–357, 1999.

[25] M. SAN CRISTOBAL, C. ROBERT-GRANIÉ et JL. FOULLEY : Hétéroscédasticité et modèles linéaires mixtes : théorie et applications en génétique quantitative. *Journal de la Société Française de Statistique*, (143), 2002.

[26] S.R. SEARLE : *Linear Models*. Wiley, 1971.

[27] T. SPEED : *Statistical Analysis of Gene Expression Microarray Data*. Interdisciplinary Statistics. Chapman & Hall/CRC, 2003.

[28] V.N. VAPNIK : *Statistical learning theory*. Wiley Inter science, 1999.

Chapitre A

Annexes

1 Analyse canonique

A : Corrélations entre gènes

	PMDCI	THIOL	CYP3A11	CYP4A10	CYP4A14
PMDCI	1.00	0.84	0.79	0.85	0.75
THIOL	0.84	1.00	0.74	0.80	0.70
CYP3A11	0.79	0.74	1.00	0.76	0.74
CYP4A10	0.85	0.80	0.76	1.00	0.89
CYP4A14	0.75	0.70	0.74	0.89	1.00
Lpin	-0.15	0.07	-0.02	-0.23	-0.29
Lpin1	-0.12	0.09	-0.03	-0.20	-0.28
GSTmu	0.42	0.57	0.62	0.44	0.53
GSTpi2	0.44	0.36	0.60	0.42	0.42
S14	0.09	0.33	-0.03	0.08	-0.11

	Lpin	Lpin1	GSTmu	GSTpi2	S14
PMDCI	-0.15	-0.12	0.42	0.44	0.09
THIOL	0.07	0.09	0.57	0.36	0.33
CYP3A11	-0.02	-0.03	0.62	0.60	-0.03
CYP4A10	-0.23	-0.20	0.44	0.42	0.08
CYP4A14	-0.29	-0.28	0.53	0.42	-0.11
Lpin	1.00	0.97	0.11	-0.15	0.58
Lpin1	0.97	1.00	0.06	-0.12	0.59
GSTmu	0.11	0.06	1.00	0.45	0.09
GSTpi2	-0.15	-0.12	0.45	1.00	-0.27
S14	0.58	0.59	0.09	-0.27	1.00

B : Corrélations entre acides gras

	C16_0	C18_0	C18_1n_9	C18_1n_7	C18_2n_6	C20_4n_6
C16_0	1.00	0.56	-0.20	0.11	-0.66	0.18
C18_0	0.56	1.00	-0.84	-0.57	-0.08	0.55
C18_1n_9	-0.20	-0.84	1.00	0.80	-0.23	-0.36
C18_1n_7	0.11	-0.57	0.80	1.00	-0.56	-0.17
C18_2n_6	-0.66	-0.08	-0.23	-0.56	1.00	0.15
C20_4n_6	0.18	0.55	-0.36	-0.17	0.15	1.00
C22_5n_6	-0.06	0.27	-0.23	-0.03	0.28	0.83
C18_3n_3	-0.37	-0.08	-0.22	-0.31	-0.06	-0.37
C22_6n_3	0.45	0.57	-0.60	-0.57	-0.03	-0.18

C20_5n_3	0.26	0.44	-0.56	-0.46	-0.19	-0.39
C22_5n_3	0.18	0.29	-0.41	-0.38	-0.22	-0.44

	C22_5n_6	C18_3n_3	C22_6n_3	C20_5n_3	C22_5n_3
C16_0	-0.06	-0.37	0.45	0.26	0.18
C18_0	0.27	-0.08	0.57	0.44	0.29
C18_1n_9	-0.23	-0.22	-0.60	-0.56	-0.41
C18_1n_7	-0.03	-0.31	-0.57	-0.46	-0.38
C18_2n_6	0.28	-0.06	-0.03	-0.19	-0.22
C20_4n_6	0.83	-0.37	-0.18	-0.39	-0.44
C22_5n_6	1.00	-0.32	-0.39	-0.44	-0.44
C18_3n_3	-0.32	1.00	-0.02	0.48	0.40
C22_6n_3	-0.39	-0.02	1.00	0.59	0.59
C20_5n_3	-0.44	0.48	0.59	1.00	0.70
C22_5n_3	-0.44	0.40	0.59	0.70	1.00

C : Corrélations entre gènes et acides gras

	C16_0	C18_0	C18_1n_9	C18_1n_7	C18_2n_6	C20_4n_6
PMDCI	0.70	0.68	-0.41	-0.14	-0.48	0.33
THIOL	0.75	0.41	-0.06	0.13	-0.62	0.19
CYP3A11	0.62	0.59	-0.52	-0.36	-0.34	-0.01
CYP4A10	0.60	0.51	-0.33	-0.16	-0.33	0.15
CYP4A14	0.45	0.39	-0.30	-0.19	-0.24	0.01
Lpin	0.18	-0.23	0.38	0.28	-0.46	-0.36
Lpin1	0.19	-0.21	0.38	0.27	-0.45	-0.31
GSTmu	0.48	0.14	-0.04	0.13	-0.35	-0.16
GSTpi2	0.35	0.55	-0.55	-0.53	0.07	0.12
S14	0.33	-0.19	0.44	0.48	-0.49	-0.05

	C22_5n_6	C18_3n_3	C22_6n_3	C20_5n_3	C22_5n_3
PMDCI	0.07	-0.08	0.41	0.34	0.17
THIOL	0.01	-0.24	0.29	0.23	0.12
CYP3A11	-0.17	0.06	0.72	0.60	0.51
CYP4A10	0.03	-0.08	0.44	0.38	0.10
CYP4A14	-0.08	-0.04	0.53	0.36	0.15
Lpin	-0.45	0.16	0.01	-0.02	0.29
Lpin1	-0.45	0.15	-0.00	-0.05	0.23
GSTmu	-0.13	-0.33	0.49	0.15	0.38
GSTpi2	-0.14	-0.09	0.66	0.42	0.36
S14	0.01	-0.07	-0.37	-0.16	-0.04

2 Modèle linéaire

D : Quelques rappels sur les lois

Loi du Chi-deux

Si $X_1, X_2, \ldots, X_n$ sont des variables aléatoires $N(0,1)$ et indépendantes alors

$$Q_n = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

avec $\mathrm{E}(Q_n) = n$ et $\mathrm{Var}(Q_n) = 2n$

Remarques :

La somme de 2 chi-deux indépendantes est aussi un chi-deux.

Une variable du chi-deux est toujours positive.

Loi de Student

Si $X \sim N(0,1)$ et $Q \sim \chi^2_n$ avec X et Q deux variables indépendantes alors

$$T_n = \frac{X}{\sqrt{Q/n}} \sim t_n$$

Remarque :

Si $n \to +\infty$, t_n tend vers une loi normale réduite.

Loi de Fisher

Si $Q_1 \sim \chi^2_{n_1}$ et $Q_2 \sim \chi^2_{n_2}$ avec Q_1 et Q_2 deux variables indépendantes alors

$$F_{n_1;n_2} = \frac{Q_1/n_1}{Q_2/n_2} \sim F^{n_1}_{n_2}$$

Théorème de Cochran

Soient $X \sim N(\mu, \Sigma)$; A et B deux matrices carrées, symétriques ($A' = A$ et $B' = B$) et idempotentes ($AA = A$ et $BB = B$) d'ordre n ; a un vecteur aléatoire de $\mathbb{R}^n$; $Q_1 = X'AX$ et $Q_2 = X'BX$, deux formes quadratiques,

Alors,

i. $A\Sigma B = 0 \Longrightarrow Q_1$ et Q_2 indépendantes,

ii. $A\Sigma a = 0 \Longrightarrow Q_1$ et $a'X$ indépendantes,

iii. $\| X - \mu \|^2 \sim \chi^2_n$,

iv. Si $rang(A) = r$, $\Sigma = I$ et $\mu = 0$ alors $X'AX \sim \chi^2_r$.