Exploration Statistique Multidimensionnelle
Data Mining

PHILIPPE BESSE

NB. *Les cours et travaux pratiques (scénarios, ateliers) du site wikistat.fr sont dispensés en formation initiale à l'INSA de Toulouse dans la spécialité d'Ingénieurs en Génie Mathématique et Modélisation. Ils sont également proposés dans le cadre de stages de Formation Professionnelle Qualifiante.*

Équipe de Statistique et Probabilités
Institut de Mathématiques de Toulouse — UMR CNRS C5219

Département Génie Mathématique et Modélisation
Institut National des Sciences Appliquées de Toulouse — 31077 – Toulouse cedex 4.

Table des matières

- Introduction à la Statistique page 5
- Description unidimensionnelle page 12
- Description bidimensionnelle page 23
- Introduction exploration multidimensionnelle page 29
- Analyse en Composantes Principales page 31
- Analyse Factorielle Discriminante page 43
- Analyse Factorielle des Correspondances page 49
- Analyse Factorielle des Correspondances Multiple page 57
- Analyse Canonique des Corrélations page 65
- Positionnement Multidimensionnel page 75
- Classification non supervisée page 80
- Factorisation par matrices non négatives (NMF) page 93
- Annexe
- Compléments d'algèbre linéaire page 99

Attention ce cours est dense, la lecture de ce document ne suffira pas à la bonne compréhension des outils qui y sont décrits de façon synthétique. La présence aux cours et la participation active aux TDs sont indispensables à l'acquisition des compétences incontournables pour une utilisation raisonnable et raisonnée des techniques d'apprentissage statistique.

Statistique : Introduction

Résumé

Introduction à la Statistique et ses méthodes. Contexte et objectifs (descriptif, explicatif, prédictif) d'une analyse statistique ; les compétences nécessaires.

Ce cours est structuré en deux niveaux principaux et quelques grands thèmes :

- *L : Description et inférences statistiques élémentaires*
- *M1 : Exploration multivariée*
- *M1 : Inférence statistique*
- *M1 : Modèle linéaire et linéaire général*
- *M2 : Modèle linéaire, modèle mixte*
- *M2 : Apprentissage et modélisation*

Réflexions autour de : Statistique et Déontologie scientifique

1 Le métier de statisticien

Le développement continu des moyens informatiques de saisie, de stockage (bases de données) et de calcul permet la production, la gestion, le traitement et l'analyse d'ensembles de données de plus en plus volumineux. Par exemple, les 600 Mb de données produites en une dizaine d'heures par l'un des séquenceurs actuels représentent l'équivalent de la production mondiale déposée dans GenBank entre 1982 et 1996. Les séquenceurs arrivant sur le marché en 2010 produisent en 5 jours 200Gb par traitement. Le perfectionnement des interfaces graphiques offre aux utilisateurs, statisticiens ou non, des possibilités de mise en œuvre très simples avec des outils logiciels de plus en plus "conviviaux". Cette évolution, ainsi que la popularisation de nouvelles méthodes algorithmiques (réseaux de neurones, support vector machine, agrégation de modèles...) et outils graphiques, conduisent au développement et à la commercialisation de logiciels généraux, ou spécifiques à des métiers, qui intègrent un sous-ensemble de méthodes statistiques et algorithmiques plus ou moins exhaustif.

Une question émerge alors de façon très présente ; elle est fondamentale pour l'emploi et les débouchés des étudiants, la gestion des ressources humaines et les investissements économiques des entreprises ou encore les stratégies scientifiques des laboratoires de recherche.

> *Quelles sont les compétences nécessaires à la mise en œuvre de tels logiciels pour analyser, modéliser, interpréter des corpus de données de plus en plus complexes et volumineux produits par une entreprise ou un laboratoire ?*

Les enjeux sont en effet majeurs ; les résultats influent directement sur les prises de décision du *management* ou la validation de résultats scientifiques et leur valorisation par des publications.

2 Terminologie

Le travail du statisticien est d'abord un travail de communication avec des représentants d'autres disciplines ou d'autres métiers. Ceci nécessite beaucoup de rigueur et donc de précision dans l'emploi des mots et concepts lorsqu'il s'agit de traduire en phrases intelligibles des résultats numériques ou graphiques. En effet, de ces *interprétations* découleront des prises de décision.

2.1 Statistique, statistiques, statistique

Le mot *statistiques* avec un "s" est apparu au XVIIIème siècle pour désigner des quantités numériques : des *tables* ou *états*, issus de techniques de dénombrement et décrivant les ressources économiques (impôts...), la situation démographique (conscription...), d'un pays. La Statistique est une sous-discipline des Mathématiques qui s'est développée depuis la fin du XIXème siècle notamment à la suite des travaux de l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman...). Une *statistique* est une quantité définie par rapport à un modèle (*i.e.* une statistique de test) permettant d'inférer sur son comportement dans une situation expérimentale donnée.

2.2 Statistique descriptive, inférentielle et apprentissage

De manière approximative, il est possible de classer les méthodes statistiques en trois groupes : celui des méthodes descriptives, celui des méthodes

inférentielles et celui récent de l'apprentissage.

- La Statistique descriptive regroupe les méthodes dont l'objectif principal est la *description* des données étudiées ; cette description des données se fait à travers leur *présentation* (la plus synthétique possible), leur *représentation graphique*, et le calcul de *résumés numériques*. Dans cette optique, il n'est pas fait appel à des modèles probabilistes. On notera que les termes de statistique descriptive, *statistique exploratoire* et *analyse des données* sont quasiment synonymes.
- La statistique inférentielle. Ce terme regroupe les méthodes dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population, l'échantillon. Il s'agit donc d'induire (ou encore d'inférer) du particulier au général avec un objectif principalement *explicatif*. Ce passage ne peut se faire qu'aux moyens de modèles et d'hypothèses probabilistes. Les termes de statistique inférentielle, *statistique mathématique*, et *statistique inductive* sont eux aussi quasiment synonymes.
- L'apprentissage statistique est issu de l'interface entre deux disciplines : *Statistique* et *Machine Learning (apprentissage machine)*. L'objectif est principalement la construction d'un modèle statistique traditionnel ou algorithmique sans nécessairement d'hypothèse probabiliste, en privilégiant la *prévision* d'une variables qualitative (discrimination ou classification supervisée) ou quantitative (régression). Le contexte est souvent celui de données de grandes dimensions avec comme défi majeur le cas où le nombre de variables explicatives p est considérablement plus important que le nombre n d'observations ou taille de l'échantillon dit d'apprentissage.

D'un point de vue méthodologique, la statistique descriptive précède la statistique inférentielle ou l'apprentissage statistique dans une démarche de traitement de données : ces différents aspects de la statistique se complètent bien plus qu'ils ne s'opposent une fois que le ou les objectifs : descriptif, explicatif, prédictif sont explicités.

Le vocabulaire de la Statistique :

Population Ω (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de *champ de l'étude*.

Individu $\omega \in \Omega$ (ou *unité statistique*) : tout élément de la population.

Échantillon : sous–ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon n : cardinal du sous-ensemble correspondant.

Enquête (statistique) : opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon.

Recensement : enquête dans laquelle l'échantillon observé est la population tout entière (enquête *exhaustive*).

Sondage : enquête dans laquelle l'échantillon observé est un sous–ensemble strict de la population (enquête *non exhaustive*).

Variable (statistique) : $\Omega \xrightarrow{X} \begin{cases} \mathcal{E} \text{ si qualitative} \\ \mathbb{R} \text{ si quantitative} \end{cases}$ caractéristique (âge, salaire, sexe, glycémie...), définie sur la population et observée sur l'échantillon ; mathématiquement, il s'agit d'une application définie sur l'échantillon. Si la variable est à valeurs dans $\mathbb{R}$ (ou une partie de $\mathbb{R}$, ou un ensemble de parties de $\mathbb{R}$), elle est dite *quantitative* (âge, salaire, taille...) ; sinon elle est dite *qualitative* (sexe, catégorie socioprofessionnelle...). Si les modalités d'une variables qualitatives sont ordonnées (*i.e.* tranches d'âge), elle est dite *qualitative ordinale* et sinon *qualitative nominale*.

Données (statistiques) : ensemble des individus observés (échantillon), des variables considérées, et des observations de ces variables sur ces individus. Elles sont en général présentées sous forme de *tableaux* (individus en lignes et variables en colonnes) et stockées dans un fichier informatique. Lorsqu'un tableau ne comporte que des nombres (valeurs des variables quantitatives ou codes associés aux variables qualitatives), il correspond à la notion mathématique de *matrice*.

3 Démarche du statisticien

Le crédo de l'enseignant de statistique consiste à répéter inlassablement : un statisticien (ou les compétences qu'il représente) doit être associé *préalablement* à une étude, des expérimentations, une enquête... De la qualité du recueil et de l'organisation des données dépendra bien évidemment la *pertinence* des résultats de l'analyse. Plusieurs questions sont préalables :

3.1 Expérimentation

- Quelle est la question biologique, sociologique, épidémiologique à laquelle je veux apporter une réponse ? En particulier, quel est l'objectif (descriptif, explicatif, prédictif ou une combinaison) ?
- Quelle est la population étudiée ?
- Comment planifier des expériences ou des recueils d'informations dans des bases pré-existantes ?
- Quels sont les échantillons ?
- Précision des conditions expérimentales
- Observations et mesures

3.2 Exploration pour un objectif descriptif

Cette étape est de toute façon un préalable à tout autre objectif. Les données recueillies sont elles de qualité suffisante ? Sont-elles bien exemptes de biais ou artefacts expérimentaux ? Leurs grandes structures (groupes, corrélations...) sont-elles en accord avec les connaissances acquises sur le sujet ?

- Valeurs manquantes, erronées ou atypiques,
- Modalités trop rares,
- Distributions "anormales",
- Incohérences, liaisons non linéaires,
- Transformations, imputation, codage...

3.3 Décision pour un objectif explicatif

Telle variable ou tel facteur a-t-il une influence sur la variable d'intérêt ? Le modèle théorique est-il en accord avec les résultats expérimentaux ?

- Explicitation de l'hypothèse statistique répondant à la question biologique,
- Détermination du modèle statistique correspondant,
- Estimation des paramètres du modèle et calcul de la statistique de test,
- Prise de décision : rejet ou acceptation de l'hypothèse.

3.4 Apprentissage pour un objectif prédictif

Un modèle *explicatif* construit dans l'étape précédente peut être un bon candidat comme modèle *prédictif* mais pas nécessairement. Paradoxalement, un modèle "vrai" n'est pas nécessairement un "meilleur" modèle prédictif s'il est trop complexe, pas assez "parcimonieux". Une quantité impressionnante de méthodes ont été développées ces dernières années sans qu'il soit possible de déterminer, *a priori*, celle qui conduira aux meilleures prévisions sur le problème et les données étudiées.

4 Quel logiciel ?

Deux logiciels sont privilégiés : l'un commercial SAS car le plus répandu et le plus demandé dans les offres d'emplois ; l'autre, R, en distribution libre (licence GNU) comme outil de développement des dernières avancées méthodologiques du monde universitaire.

4.1 SAS

Mis à part le module SAS/IML de langage matriciel très peu utilisé, SAS est un logiciel de type "boîte noire" superposant des couches basses, pour lesquelles l'utilisateur écrit des lignes de code dans une syntaxe complexe, et des interfaces graphiques conviviales (SAS/INSIGHT, SAS User Guide, Sas Enterprise Miner...). Sa diffusion est telle qu'il apparaît en situation de quasi monopole dans certaines branches d'activité comme l'industrie pharmaceutique. Paradoxalement, sa complexité et son coût sont des atouts pour l'emploi de statisticiens indispensables à sa bonne utilisation et donc à sa rentabilisation. Son apprentissage est incontournable.

4.2 R

A l'opposé et à l'exception des traitements les plus rudimentaires pilotés par menu, R est avant tout un langage de programmation pour la manipulation des objets du statisticien : vecteurs, matrices, bases de données, liste de résultats, graphiques. D'un point de vue pédagogique, sa mise en œuvre oblige à l'indispensable compréhension des méthodes et de leurs limites. Il force à 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 ré-agencement ou la réorganisation de quelques commandes R offrent une combinatoire très ou-

verte de possibilités contrairement à un système clos de menus prédéfinis. Il offre par ailleurs, grâce à de nombreuses librairies facilement 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 certaines disciplines (Biologie). Les limitations de R sont d'une part celles d'un langage interprété : lenteur pour l'exécution de boucles (à éviter) et d'autre part la taille des données car elles sont toutes chargées en mémoire.

4.3 Quel choix ?

En résumé, il est bien et utile de savoir utiliser ces deux types de logiciels et il est important de comprendre que l'apprentissage syntaxique d'un logiciel est indispensable mais secondaire. Une fois les méthodes comprises et appréhendées, il est techniquement facile de passer d'un logiciel à l'autre, leurs fonctionnalités étant structurellement les mêmes. La difficulté principale ne réside pas dans l'obtention de sorties ou résultats mais dans leur *compréhension*.

5 Domaines d'application

Toutes les méthodes et techniques utilisées nécessitent d'être illustrées sur des exemples simples ou "académiques", pour ne pas dire simplistes, afin d'en comprendre les fondements. Néanmoins, leur apprentissage effectif requiert leur utilisation effective sur des jeux de données en vraie grandeur, issus de différents domaines d'applications. Ce n'est qu'à cette condition que peuvent être appréhendées les difficultés de mise en œuvre, les limites, les stratégies d'interprétation mais aussi la grande efficacité de ces outils.

Ils sont tirés des principaux domaines d'application de la Statistique.

5.1 Sciences de la Vie

Depuis les travaux pionniers de Sir Ronald Fisher, les disciplines des Sciences de la Vie ont toujours motivé les développements de la Statistique : modèles de durée de vie, modèles épidémiologiques, dynamique de population... Les techniques de séquençage et les technologies d'instrumentation à haut débit (transcriptomique, protéomique, métabolomique...) viennent renforcer lourdement cette tendance en posant des défis redoutables au statisticien : que faire lorsque les transcriptions (quantités d'ARN messagers) de milliers de gènes (les variables statistiques) sont simultanément observées pour seulement quelques dizaines d'échantillons biologiques ?

La figure : 1 est un exemple original d'emploi de l'analyse canonique (objectif descriptif). Cette méthode permet de mettre en relation deux paquets de variables (gènes et concentrations d'acides gras) observées sur les mêmes individus (souris).

Le jeu de données utilisé provient de l'Unité de Pharmacologie-Toxicologie de l'INRA de Toulouse. Il concerne 40 souris réparties en 2 génotypes (sauvages et génétiquement modifiées : PPARα déficientes) et 5 régimes alimentaires (dha, efad, lin, ref, tsol). Le plan est équilibré complet : quatre souris par combinaison des deux facteurs.

dha régime enrichi en acides gras de la famille Oméga 3 et particulièrement en acide docosahexaénoïque (DHA), à base d'huile de poisson ;

efad (Essential Fatty Acid Deficient) : régime constitué uniquement d'acides gras saturés, à base d'huile de coco hydrogénée ;

lin régime riche en Oméga 3, à base d'huile de lin ;

ref régime dont l'apport en Oméga 6 et en Oméga 3 est adapté des Apports Nutritionnels Conseillés pour la population française, soit sept fois plus d'Oméga 6 que d'Oméga 3 ;

tsol riche en Oméga 6, à base d'huile de tournesol.

Les expressions des gènes ainsi que des concentrations de 21 acides gras sont mesurées au niveau du foie après euthanasie. Ce jeu de données aux problématiques statistiques très riches est très souvent repris tout au long des présentations des différentes méthodes.

5.2 Marketing

La prospection ou *fouille de données (data mining)* est une appellation issue des services marketing spécialisés dans la gestion de la relation client (GRC) (*client relation management* ou CRM). Elle désigne un ensemble de techniques statistiques souvent regroupées dans un logiciel spécialement conçu à cet effet et vendu avec un slogan racoleur (SAS Enterprise Miner) :

> *Comment trouver un diamant dans un tas de charbon sans se salir les mains.*

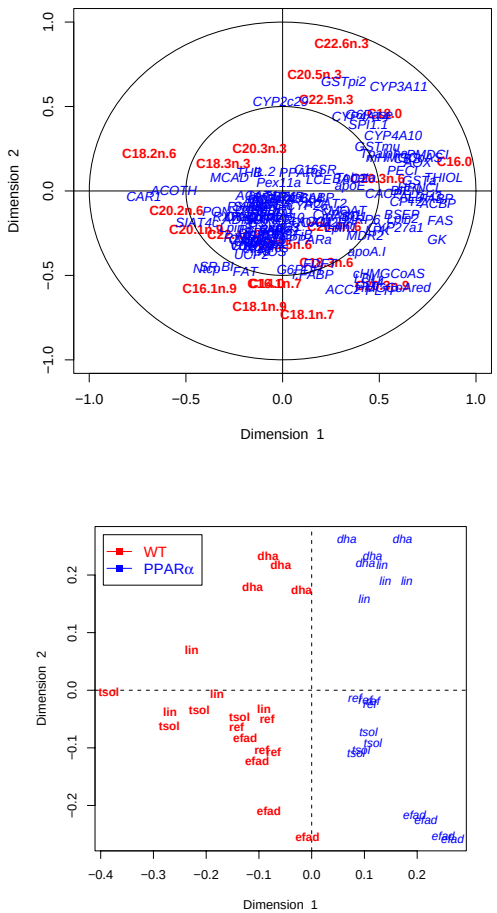

FIGURE 1 – *Souris : premier plan des facteurs canoniques : représentation conjointe des relations gènes et acides gras puis des souris selon le génotype et le régime suivi.*

Les entreprises commerciales du tertiaire (banques, assurances, téléphonie, marketing directe, publipostage, ventes par correspondance...) sont en effet très motivées pour tirer parti et amortir, par une aide à la décision quantifiée, les coûts de stockage des téraoctets que leur service informatique s'emploie à administrer.

Le contexte informationnel de la fouille de données est celui des *data warehouses*. Un entrepôt de données, dont la mise en place est assurée par un gestionnaire de données (data manager), est un ensemble de bases relationnelles extraites des données brutes de l'entreprise et relatives à une problématique.

Chaque banque, assurance... dispose d'un fichier client qui, pour des raisons comptables, enregistre tous leurs mouvements et comportements. Les données anonymes en provenance d'une banque décrivent tous les soldes et produits financiers (emprunt, contrats d'assurance vie...) détenus par les clients ainsi que l'historique mensuel des mouvements, nombre d'opérations, de jours à découvert... La base initiale étudiée comprend 1425 clients décrits par 32 variables explicitées dans une vignette décrivant les données.

Le graphique représenté est un grand classique du marketing bancaire. L'objectif (descriptif) de statistique multidimensionnelle est de construire des classes ou segments de clients homogènes quant à leur comportement bancaire. Une fois les classes construites et l'ensemble des clients affectés, l'agent commercial sait quel langage adopter, quels produits proposer au client qu'il a en face de lui. Après une analyse factorielle des correspondances multiples, les clients caractérisés par leur nouvelles coordonnées sont regroupés en classes dont l'explicitation est facilitée par la représentation des modalités de ces classes dans le plan factoriel de l'analyse des correspondances multiples (figure 2). Un autre objectif (apprentissage) est abordé sur ces mêmes données pour la recherche de scores d'appétences ou d'attrition. Les applications marketing sont très nombreuses (intérêts de certains clients pour des produits financiers, risque pour d'autres clients de changer de fournisseur en téléphonie). Elles le sont également dans les applications financières : risque de défaut de paiement d'un client, de ruine d'une entreprise.

5.3 Industrie

Pour des raisons culturelles et historiques trop longues à développer (culture déterministe des Écoles d'ingénieurs...), la Statistique a une place très mi-

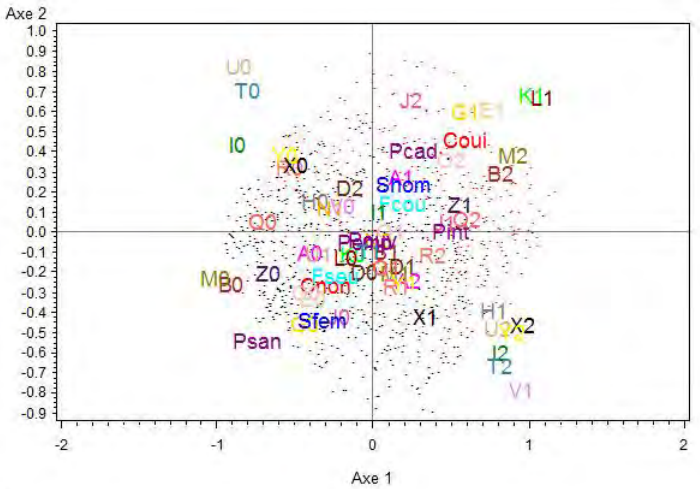

FIGURE 2 – *Banque : représentation des classes de clients, w1 à w5, dans le premier plan factoriel de l'analyse des correspondances multiples*

neures dans l'industrie française sauf en cas d'obligation légale : essais cliniques pour l'autorisation de mise sur le marché des médicaments, contrôle de qualité et fiabilité des matériaux pour la conformité aux normes ISO... La Statistique est ainsi plus vécue comme une contrainte, un contrôle, que comme une aide à la décision. D'autre part, les exemples développés dans le cadre de thèses sont, outre les questions de confidentialité, souvent trop complexes à expliciter pour s'adapter à la simple illustration de ce cours. Néanmoins, il faut être conscient que chacune des techniques abordées, en particulier celles de biostatistique, se transposent directement : durée de vie et fiabilité des matériaux, fouille de données et traçabilité pour la détection de défaillances... Le contexte est souvent techniquement très complexe en terme de modélisation physique mais plus favorable sur le plan statistique, du fait notamment d'un plus grand nombre d'observations que dans le domaine de la santé.

5.4 *Big Data*

Les entreprises industrielles sont actuellement confrontées à la même situation que celles du tertiaire il y a vingt ans : afflux automatique et stockage massif de données. La situation et donc les métiers de la Statistique évoluent considérablement dans ce domaine. Après une période où la question principale est : comment organiser et structurer les matériels et bases de données, la question suivante est : que faire, quelles analyses développées pour les valoriser et aider à la décision ? Prospection numérique dans l'industrie pétrolière, web mining des sites marchands en pleine explosion, utilisation massive des repérages GPS de flottes de véhicules, bâtiments intelligents bardés de capteurs, imagerie 3D... Les applications et problèmes nécessitent en plus, par rapport au data mining maintenant classique, une réflexion approfondie sur les structures de données : fonctions, surfaces, graphes...

6 Quelles compétences ?

Les compétences acquises doivent permettre de répondre avec assurance aux questions suivantes ou alors conduire à une proposition de redéfinition de la problématique envisagée si celle-ci est trop mal engagée.

- Quelle est précisément la question posée ?
- Quelle méthode utiliser avec quelles limites ?
- Comment la mettre en œuvre ?

- Comprendre les sorties du logiciel utilisé.
- Quelle décision ?

Un argument tendancieux, pour ne pas dire fallacieux, est souvent avancé : *il n'est pas besoin d'être mécanicien pour conduire une voiture*. C'est vrai, il n'est pas nécessaire d'être informaticien pour utiliser un ordinateur. En revanche, toute étude statistique nécessite des choix fondamentaux : transformation des données, sélection de variables, choix de méthodes, valeurs des options et paramètres de ces méthodes... qu'il n'est pas prudent de laisser faire, par défaut, au logiciel utilisé. Ces choix ne sont pas anodins et autrement plus difficiles à déterminer que le choix du carburant dans une voiture. Ils doivent être conduits en connaissance de cause par opposition à une stratégie de Shadok (cf. figure 3) qui est un mode d'apprentissage de type "jeux vidéos" : exclusivement par essais — erreurs. Elles est utile, mais pas en toute circonstance, car il ne suffit pas d'obtenir un résultat pour qu'il soit pertinent ou même simplement juste.

FIGURE 3 – *Shadok : devise numéro 1*

Statistique descriptive unidimensionnelle

Résumé

Les objectifs et la démarche d'un première exploration d'un jeu de données, les outils de la description statistique d'une variable quantitative (indicateur de tendance centrale, de dispersion, histogramme, diagramme-boîte), puis d'une variable qualitative (fréquences).

Retour au plan.

1 Introduction

l'objectif des outils de Statistique descriptive élémentaire est de fournir 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, 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 descriptives à réaliser sur un jeu de données.

1.1 Démarche

Toute étude sophistiquée d'un corpus de données doit être précédée 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 de dépister les sources de problèmes :

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

C'est ensuite la recherche de prétraitements des données afin de corriger les sources de problèmes et les rendre exploitables par des techniques plus sophistiquées :

- transformation : logarithme, puissance, réduction, rangs... des variables,
- codage en classe ou recodage de classes,
- imputations ou non des données manquantes,
- lissage, décompositions (ondelettes, Fourier) de courbes,

Ensuite, les techniques exploratoires multidimensionnelles permettent des

- représentations graphiques synthétiques,
- réductions de dimension pour la compression ou le résumé des données,
- recherches et représentations de typologies des observations.

1.2 Avertissement

Attention le côté rudimentaire voire trivial des outils de statistique descriptive uni et bidimensionnelle 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, les méthodes descriptives ne supposent, *a priori*, aucun modèle sous-jacent, de type probabiliste. Ainsi, lorsque l'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.

Le 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 ou transformation des données pour éventuellement les remettre en cause.

2 Variable quantitative

2.1 Variable quantitative discrète

Introduction

En général, on appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs entières (plus rarement décimales). Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible (sauf exception, moins d'une vingtaine). Citons, par exemple, le nombre d'enfants dans une population de familles, le nombre d'années d'études après le bac dans une population d'étudiants...

On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise ; la série statistique brute est donnée ci-dessous (il s'agit de données fictives).

43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31
62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59
57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

Présentation des données

Le tableau statistique C'est un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X ; ces observations sont rangées par ordre croissant et non répétées ; nous les noterons $\{x_l\ ;\ l = 1, \ldots, r\}$. Dans une seconde colonne, on dispose, en face de chaque valeur x_l, le nombre de réplications qui lui sont associées ; ces réplications sont appelées *effectifs* et notées n_l. Les effectifs n_l sont souvent remplacés par les quantités $f_l = \frac{n_l}{n}$, appelées *fréquences* (rappelons que n désigne le nombre total d'observations, c'est–à–dire le cardinal de Ω : $n = \sum_{l=1}^{r} n_l$).

x_l	n_l	N_l	$f_l(\%)$	$F_l(\%)$
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,91
52	3	38	6,25	79,16
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100,00

TABLE 1 – *Effectifs, effectifs cumulés, fréquences et fréquences cumulées.*

Les effectifs cumulés et les fréquences cumulées Il peut être utile de compléter le tableau statistique en y rajoutant soit les effectifs cumulés, soit les fréquences cumulées. Ces quantités sont respectivement définies de la façon suivante :

$$N_l = \sum_{j=1}^{l} n_j \text{ et } F_l = \sum_{j=1}^{l} f_j.$$

On notera que $N_r = n$ et $F_r = 1$.

Illustration Dans le tableau statistique (1), on a calculé, sur les données présentées dans l'exemple 2.1, les effectifs, effectifs cumulés, fréquences et fréquences cumulées.

Remarque. —

- Comme c'est le cas ci-dessus, les fréquences sont souvent exprimées en pourcentages.
- Le choix entre effectifs (resp. effectifs cumulés) et fréquences (resp. fréquences cumulées) est très empirique ; il semble naturel de choisir les effectifs lorsque l'effectif total n est faible et les fréquences lorsqu'il est plus important ; la limite approximative de 100 paraît, dans ces conditions, assez raisonnable.

La présentation tige–et–feuille (ou "stem–and–leaf") Cette façon particulière de présenter les données est assez commode, dans la mesure où elle préfigure déjà un graphique. Elle est illustrée ci–dessous sur le même exemple que précédemment.

2	4 6 6 9 9 9
3	1 1 3 3 3 3 7 7 8 8 8 8
4	1 1 1 3 3 3 5 6 6 6 6 6 9 9 9
5	0 2 2 2 7 7 7 7 7 9 9
6	0 0 2

Elle consiste donc, dans la présentation des données, à séparer la partie des dizaines de celle des unités. En face de la partie des dizaines, chaque unité est répétée autant de fois qu'il y a d'observations de la valeur correspondante. Bien entendu, cette présentation doit être adaptée de façon appropriée lorsque les données sont d'un autre ordre de grandeur.

Représentations graphiques

Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques, qui sont en fait complémentaires : le diagramme en bâtons et le diagramme cumulatif (en escaliers).

Le diagramme en bâtons Il permet de donner une vision d'ensemble des observations réalisées. La figure 1 donne le diagramme en bâtons des données de l'exemple 2.1.

Le diagramme cumulatif Il figure les effectifs cumulés (resp. les fréquences cumulées) et permet de déterminer simplement le nombre (resp. la proportion)

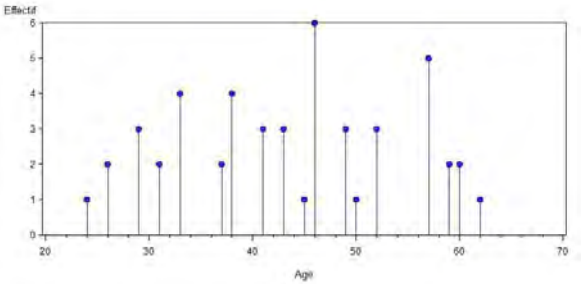

FIGURE 1 – *Diagramme en bâtons*

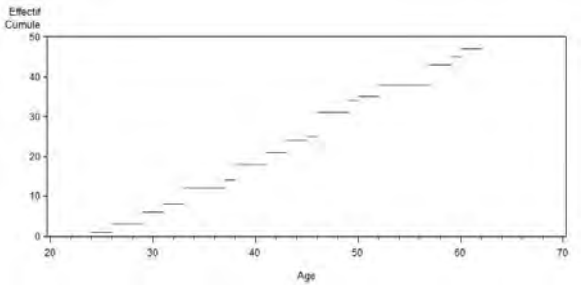

FIGURE 2 – *Diagramme cumulatif*

d'observations inférieures ou égales à une valeur donnée de la série. Lorsqu'il est relatif aux fréquences, c'est en fait le graphe de la *fonction de répartition empirique* F_X définie de la façon suivante :

$$F_X(x) = \begin{cases} 0 & \text{si } x < x_1, \\ F_l & \text{si } x_l \leq x < x_{l+1}, \quad l = 1, \ldots, r-1, \\ 1 & \text{si } x \geq x_r. \end{cases}$$

Le diagramme cumulatif relatif à l'exemple 2.1 est donné par la figure 2.

Notion de quantile

Définition La fréquence cumulée F_l ($0 \leq F_l \leq 1$) donne la proportion d'observations inférieures ou égales à x_l. Une approche complémentaire consiste à se donner a priori une valeur α, comprise entre 0 et 1, et à rechercher x_α vérifiant $F_X(x_\alpha) \simeq \alpha$. La valeur x_α (qui n'est pas nécessairement unique) est appelée quantile (ou *fractile*) d'ordre α de la série. Les quantiles les plus utilisés sont associés à certaines valeurs particulières de α.

La médiane et les quartiles La médiane est le quantile d'ordre $\frac{1}{2}$; elle partage donc la série des observations en deux ensembles d'effectifs égaux. Le premier quartile est le quantile d'ordre $\frac{1}{4}$, le troisième quartile celui d'ordre $\frac{3}{4}$ (le second quartile est donc confondu avec la médiane).

Les autres quantiles Les *quintiles*, *déciles* et *centiles* sont également d'usage assez courant.

Le diagramme-boîte (ou "box–and–whisker 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. La figure 3 donne le diagramme–boîte de l'exemple 2.1. Dans cet exemple, on a obtenu $x_{\frac{1}{4}} = 35$, $x_{\frac{1}{2}} = 44$ et $x_{\frac{3}{4}} = 52$; on notera que l'obtention, d'une part de $x_{\frac{1}{4}}$ et $x_{\frac{1}{2}}$, d'autre part de $x_{\frac{3}{4}}$, ne s'est pas faite de la même façon (en fait, avec une variable discrète, la détermination des quantiles est souvent approximative comme on peut le constater avec cet exemple).

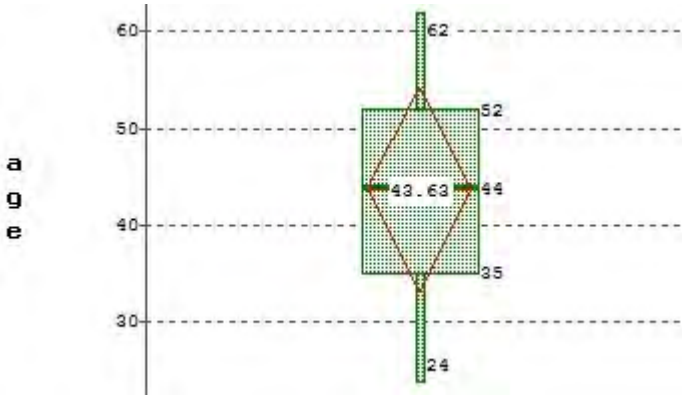

FIGURE 3 – *Diagramme-boîte et moyenne en rouge*

Caractéristiques numériques

Les caractéristiques (ou résumés) numériques introduites ici servent à synthétiser la série étudiée au moyen d'un petit nombre de valeurs numériques. On distingue essentiellement les caractéristiques de tendance centrale (ou encore de *position* ou de *localisation*) et les caractéristiques de dispersion.

Tendance centrale Leur objectif est de fournir un ordre de grandeur de la série étudiée, c'est–à–dire d'en situer le centre, le milieu. Les deux caractéristiques les plus usuelles sont :

- la *médiane*,
- la *moyenne* (ou moyenne arithmétique).

Formule de la moyenne pour une variable quantitative discrète :

$$\overline{x} = \frac{1}{n}\sum_{l=1}^{r} n_l x_l = \sum_{l=1}^{r} f_l x_l.$$

Dispersion Elles servent à préciser la variabilité de la série, c'est–à–dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale.

- L'*étendue* ($x_r - x_1$),
- l'*intervalle inter-quartiles* ($x_{\frac{3}{4}} - x_{\frac{1}{4}}$),
- l'*écart-moyen à la médiane* ($\frac{1}{n}\sum_{l=1}^{r} n_l \left|x_l - x_{\frac{1}{2}}\right|$),
- l'*écart-moyen à la moyenne* ($\frac{1}{n}\sum_{l=1}^{r} n_l |x_l - \overline{x}|$),

sont des caractéristiques de dispersion que l'on rencontre parfois.

Mais, la caractéristique de loin la plus utilisée est l'*écart–type*, racine carrée positive de la *variance*. Formules de la variance :

$$\begin{aligned} var(X) = \sigma_X^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l(x_l - \overline{x})^2 \\ &= \frac{1}{n}\sum_{l=1}^{r} n_l(x_l)^2 - (\overline{x})^2. \end{aligned}$$

L'écart–type de X sera donc noté σ_X.

Illustration En utilisant toujours l'exemple 2.1, on a calculé :

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{l=1}^{r} n_l x_l = \frac{2094}{48} = 43,625 \simeq 43,6 \text{ ans} ; \\ \sigma_X^2 &= \frac{1}{n}\sum_{l=1}^{r} n_l(x_l)^2 - (\overline{x})^2 = \frac{96620}{48} - (43,625)^2 \simeq 109,7760 ; \\ \sigma_X &= \sqrt{\sigma_X^2} \simeq 10,5 \text{ ans}. \end{aligned}$$

Remarque. — Toutes les caractéristiques numériques introduites ici (médiane, moyenne, variance, écart–type...) sont dites *empiriques*, c'est–à–dire calculées sur un échantillon Ω ; par opposition, on parle, par exemple, de moyenne *théorique* (ou espérance mathématique) pour désigner le concept de moyenne relatif à une variable aléatoire réelle.

2.2 Variable quantitative continue

Généralités

Une variable quantitative est dite continue lorsque les observations qui lui sont associées ne sont pas des valeurs précises mais des intervalles réels. Cela signifie que, dans ce cas, le sous–ensemble de $\mathbb{R}$ des valeurs possibles de la variable étudiée a été divisé en r intervalles contigus appelés *classes*.

En général, les deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative sont le grand nombre d'observations distinctes (un traitement en discret serait dans ce cas peu commode) et le caractère "sensible" d'une variable (il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis). Deux exemples de variables quantitatives fréquemment considérées comme continues sont l'âge et le revenu (pour un groupe d'individus).

Nous noterons $(b_0\,;b_1)$,...,$(b_{r-1}\,;b_r)$ les classes considérées. Les nombres b_{l-1} et b_l sont appelés les *bornes* de la $l^{\text{ième}}$ classe ; $\frac{b_{l-1}+b_l}{2}$ est le *centre* de cette classe et $(b_l - b_{l-1})$ en est l'*amplitude* (en général notée a_l).

Présentation des données

On utilise encore un tableau statistique analogue à celui vu au paragraphe précédent, en disposant dans la première colonne les classes rangées par ordre croissant. Les notions d'effectifs, de fréquences, d'effectifs cumulés et de fréquences cumulées sont définies de la même façon que dans le cas discret. On notera que l'on n'utilise pas dans ce cas la présentation tige–et–feuille car les valeurs exactes de la série sont inconnues.

Le tableau ci-dessous donne, pour l'année 1987, la répartition des exploitations agricoles françaises selon la SAU (surface agricole utilisée) exprimée en hectares (Tableaux Économiques de Midi–Pyrénées, INSEE, 1989, p. 77) ; la SAU est ici une variable quantitative continue comportant 6 classes.

SAU (en ha)	fréquences (%)
moins de 5	24,0
de 5 à 10	10,9
de 10 à 20	17,8
de 20 à 35	20,3
de 35 à 50	10,2
plus de 50	16,8

Représentations graphiques

Les deux graphiques usuels remplaçant respectivement dans ce cas le diagramme en bâtons et le diagramme cumulatif sont l'histogramme et la courbe cumulative.

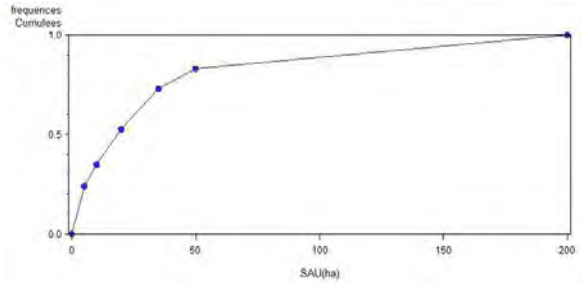

FIGURE 4 – *Courbe cumulative*

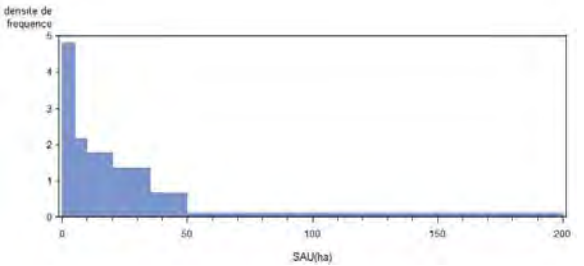

FIGURE 5 – *Histogramme (classes d'effectifs égaux) des répartitions des SAU*

Courbe cumulative C'est encore une fois le graphe de la *fonction de répartition empirique*, cette dernière devant maintenant être précisée au moyen d'*interpolations linéaires*.

On appelle fonction de répartition empirique de la variable continue X la fonction F_X définie par :

$$F_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ F_{l-1} + \frac{f_l}{b_l - b_{l-1}}(x - b_{l-1}) & \text{si } b_{l-1} \leq x < b_l, \quad l = 1, \ldots, r, \\ 1 & \text{si } x \geq b_r \end{cases}$$

(on a supposé $F_0 = 0$).

La courbe cumulative relative à l'exemple 2.2 est donnée par la figure 4. On notera que dans cet exemple, comme c'est souvent le cas avec une variable quantitative continue, il a fallu fixer arbitrairement la borne inférieure de la première classe (il était naturel ici de prendre $b_0 = 0$) ainsi que la borne supérieure de la dernière classe (on a choisi $b_6 = 200$, mais d'autres choix étaient possibles).

Histogramme La fonction de répartition empirique est, dans le cas continu, une fonction dérivable sauf, éventuellement, aux points d'abscisses $b_0, b_1, \ldots, b_r$. Sa fonction dérivée, éventuellement non définie en ces points, est appelée *densité empirique* de X et notée f_X. On obtient :

$$f_X(x) = \begin{cases} 0 & \text{si } x < b_0, \\ \frac{f_l}{b_l - b_{l-1}} & \text{si } b_{l-1} < x < b_l, \quad l = 1, \ldots, r, \\ 0 & \text{si } x \geq b_r. \end{cases}$$

Le graphe de f_X est alors appelé histogramme de la variable X. Un histogramme est donc la juxtaposition de rectangles dont les bases sont les amplitudes des classes considérées ($a_l = b_l - b_{l-1}$) et dont les hauteurs sont les quantités $\frac{f_l}{b_l - b_{l-1}}$, appelées *densités de fréquence*. L'aire du $l^{\text{ième}}$ rectangle vaut donc f_l, fréquence de la classe correspondante.

L'histogramme correspondant aux données de l'exemple 2.2 est présenté dans la figure 5.

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*. Les démonstrations du ca-

ractère optimal de ces estimations fonctionnelles, liée à l'optimalité du choix de la valeur du paramètre de lissage, font appel à des outils théoriques plus sophistiquées sortant du cadre de ce cours (Eubank, 1988, Silverman, 1986).

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 optimisée 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 :

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

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.

Quantiles

Les quantiles x_α d'une variable continue peuvent être déterminés de façon directe à partir de la courbe cumulative. Cela signifie que, par le calcul, on doit commencer par déterminer la classe dans laquelle se trouve le quantile cherché, puis le déterminer dans cette classe par interpolation linéaire (voir l'illustration plus loin).

Moyenne et écart-type

La moyenne, la variance et l'écart–type d'une variable continue se déterminent de la même manière que dans le cas discret ; dans les formules, on doit prendre pour x_l les centres de classes au lieu des observations (qui ne sont pas connues). Les valeurs obtenues pour ces caractéristiques sont donc assez approximatives ; cela n'est pas gênant dans la mesure où le choix de traiter une variable quantitative comme continue correspond à l'acceptation d'une certaine imprécision dans le traitement statistique.

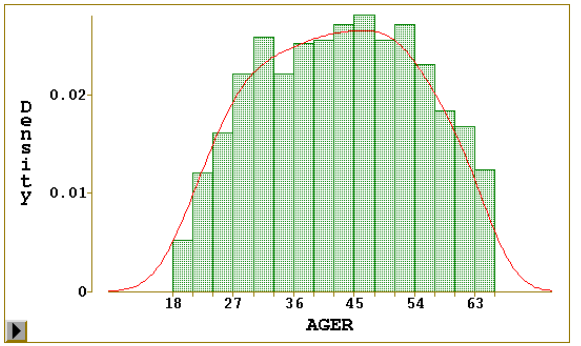

FIGURE 6 – *Histogramme (classes amplitudes égales) des répartitions des âges et estimation non paramétrique de la densité par la méthode du noyau (en rouge).*

Illustration

La médiane de la variable présentée dans l'exemple 2.2 se situe dans la classe (10 ; 20), puisque la fréquence cumulée de cette classe (52,7) est la première à dépasser 50. On détermine la médiane en faisant l'interpolation linéaire suivante (l'indice l ci–dessous désigne en fait la troisième classe) :

$$\begin{aligned} x_{\frac{1}{2}} &= b_{l-1} + a_l \frac{50 - F_{l-1}}{F_l - F_{l-1}} \\ &= 10 + 10\frac{15,1}{17,8} \\ &\simeq 18,5 \text{ ha.} \end{aligned}$$

La moyenne vaut :

$$\overline{x} = \sum_{l=1}^{r} f_l x_l = \frac{3080,5}{100} \simeq 30,8 \text{ ha.}$$

Remarque. —

Dans cet exemple, il convient de noter trois choses :

- tout d'abord, pour le calcul de la moyenne, nous avons choisi $x_6 = 100$, plutôt que 125, car cette valeur nous a semblé plus proche de la réalité ;
- ensuite, il se trouve que, dans ce cas, on peut calculer la *vraie* valeur de la moyenne, connaissant la SAU totale en France (31 285 400 ha) et le nombre total d'exploitations agricoles (981 720) ; on obtient 31,9 ha, ce qui signifie que l'approximation obtenue ici est très correcte ;
- enfin, le fait que la médiane soit sensiblement plus faible que la moyenne caractérise les séries fortement concentrées sur les petites valeurs.

2.3 Variables quantitatives et logiciels

Le volume des données et la pratique généralisée des logiciels statistiques induit une prise en compte particulière des notions précédentes. Par principe, le codage des valeurs, mêmes réelles, est toujours discret, et la précision fonction du nombre de chiffres significatifs pris en compte En conséquences, tous les calculs des indicateurs (moyenne, variance, quantile...) sont traités avec les formules considérant les valeurs comme connues et discrètes, sans pour autant s'intéresser aux fréquences des valeurs car ces dernières sont généralement distinctes les unes des autres. En revanche, les graphiques produits (histogramme, courbe cumulative mais pas l'estimation fonctionnelle) sont issus de découpages automatiques en classes d'amplitudes égales, pas toujours très judicieux, selon les principes des variables continues.

3 Variable qualitative

3.1 Variables nominales et ordinales

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 dans une population d'étudiants), la variable est dite *ordinale*. Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite *nominale*.

3.2 Traitements statistiques

Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques. Encore faut–il noter que les notions d'effectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).

3.3 Représentations graphiques

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 :

- le *diagramme en colonnes,*
- le *diagramme en barre,*
- le *diagramme en secteurs.*

Les figures 8, 7 et 9 présentent chacun de ces trois graphiques sur les données de l'exemple 3.3.

> Le tableau ci–dessous donne la répartition de la population active occupée (ayant effectivement un emploi) selon la CSP (catégorie socioprofessionnelle), en France, en mars 1988 (Tableaux de l'Économie Française, INSEE, 1989, p. 59).

CSP	effectifs en milliers	fréquences (%)
1. agriculteurs exploitants	1312	6,1
2. artisans, commerçants, chefs d'entreprises	1739	8,1
3. cadres, professions intellectuelles supérieures	2267	10,6
4. professions intermédiaires	4327	20,1
5. employés	5815	27,0
6. ouvriers	6049	28,1

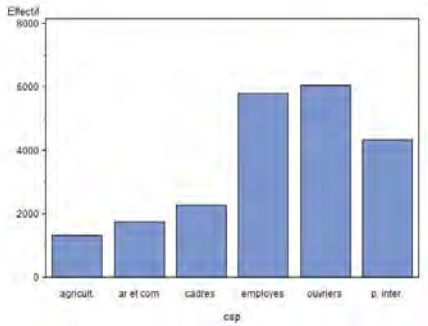

FIGURE 7 – *Diagramme en colonnes*

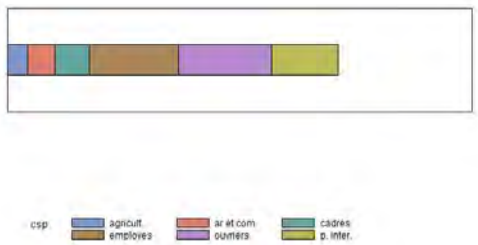

FIGURE 8 – *Diagramme en barre*

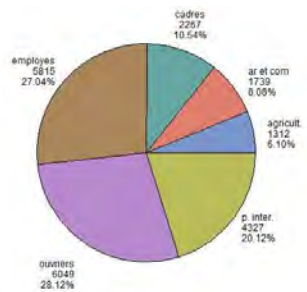

FIGURE 9 – *Diagramme en secteurs*

4 Détection de 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 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 à faire 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.

4.1 Marketing bancaire

Les données de patrimoine, de revenu, comme également celles de concentration présente des distributions très disymétriques (figure 10 accompagnées de nombres importants de valeurs atypiques. Le diagramme boîte est un outil efficace pour identifier ce problème avant d'y remédier par une transformation appropriée, ici le logarithme.

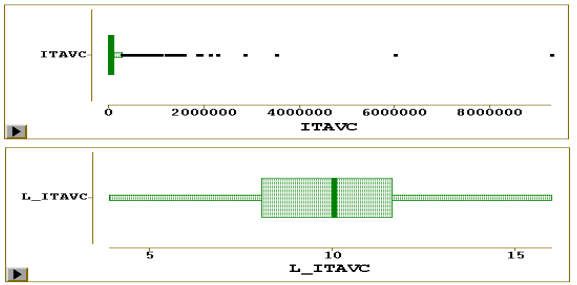

FIGURE 10 – *Banque : La simple transformation* ($\log(50+x)$)*, de la variable cumulants les avoirs, résout bien les problèmes posés par l'allure "log-normale" de sa distribution avec son cortège de valeurs atypiques.*

4.2 Données génomiques

Le diagramme boîte parallèle est également très efficace pour visualiser simultanément les distributions d'un grand nombre de variables, par exemple de centaines voire de milliers de gènes, dont l'expression a été observée dans différentes conditions expérimentales. Dans cet exemple, la représentation des diagrammes en boîtes pour les souris, ordonnées selon le génotype et le régime suivi (Fig. 11) 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 12 comme les plus sur-exprimés, se retrouvent dans les valeurs extrêmes pour chaque souris sur la figure 11.

Les diagrammes en boîtes pour chaque gène (Fig. 12) 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,

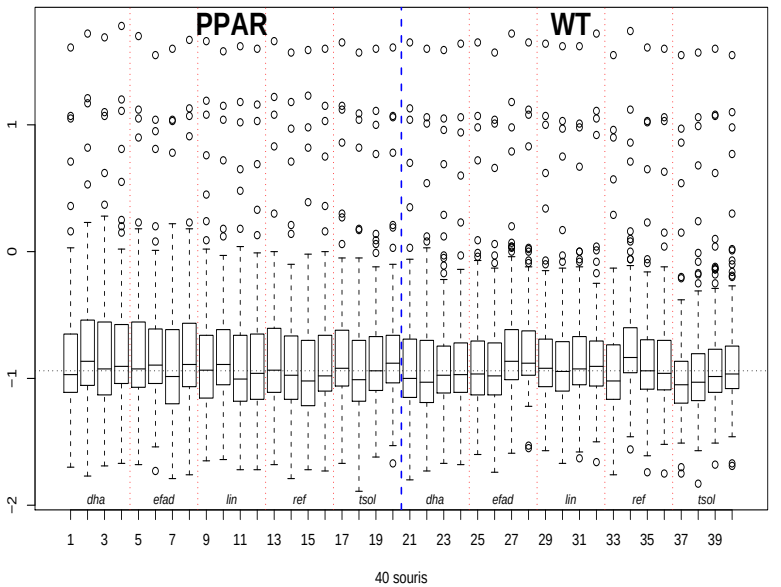

FIGURE 11 – *Souris : diagrammes en boîtes pour les 40 souris. La ligne verticale et épaisse sépare les souris selon leur génotype. Les lignes verticales et fines séparent les souris selon le régime qu'elles ont suivi. La ligne horizontale représente la médiane de l'ensemble des valeurs.*

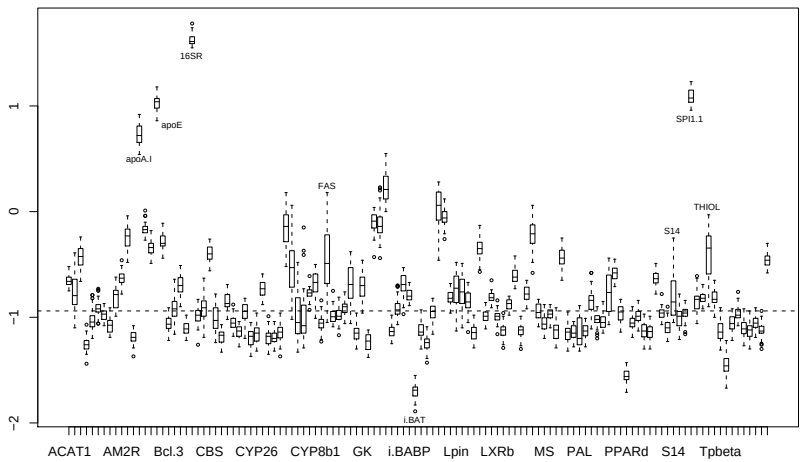

FIGURE 12 – *Souris : Diagrammes-boîtes parallèles représentant simultanément les distributions des logarithmes des expressions des gènes.*

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.

Statistique descriptive bidimensionnelle

Résumé

Liaisons entre variables quantitatives (corrélation et nuages de points), qualitatives (contingence, mosaïque) et de types différents (rapport de corrélation). Introduction au cas multidimensionnel.

Retour au plan.

1 Introduction

Dans cette section, on s'intéresse à l'étude simultanée de deux variables X et Y, étudiées sur le même échantillon, toujours noté Ω. 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*).

2 Deux variables quantitatives

2.1 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 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*.

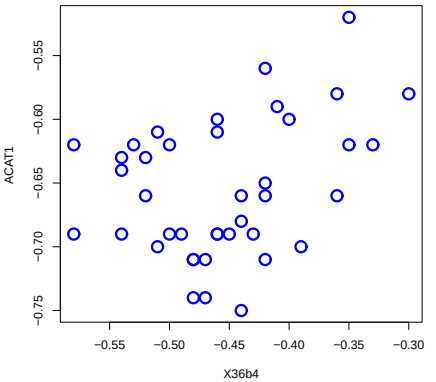

FIGURE 1 – *Souris : Nuage de points illustrant la faible liaison linéaire entre les expressions de deux gènes (corrélation de 0,33).*

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).

2.2 Rappel : variables centrées et réduites

Si X est une variable quantitative de moyenne $\overline{x}$ et d'écart–type σ_X, on appelle variable centrée associée à X la variable $X - \overline{x}$ (elle est de moyenne

nulle et d'écart–type σ_X), et variable centrée et réduite (ou tout simplement variable réduite) associée à X la variable $\frac{X-\overline{x}}{\sigma_X}$ (elle est de moyenne nulle et d'écart–type égal à un). Une variable centrée et réduite s'exprime sans unité.

2.3 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. En particulier, on en déduit les deux formules suivantes :

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y),$$

$$[cov(X,Y)]^2 \leq var(X)var(Y)\,;$$

(cette dernière propriété est l'inégalité de Cauchy-Schwarz).

la covariance 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 (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. Les valeurs -1 et $+1$ correspondent à une liaison linéaire parfaite entre X et Y (existence de réels a, b et c tels que : $aX+bY+c=0$).

Notons pour mémoire la possibilité d'utiliser d'autres indicateurs de liaison entre variables quantitatives. Construits sur les rangs (corrélation de Spearman) ils sont plus robustes faces à des situations de non linéarité ou des valeurs atypiques mais restent très réducteurs.

3 Une variable quantitative et une qualitative

3.1 Notations

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

$$x_1, \ldots, x_\ell, \ldots, x_m$$

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 :

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

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

3.2 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

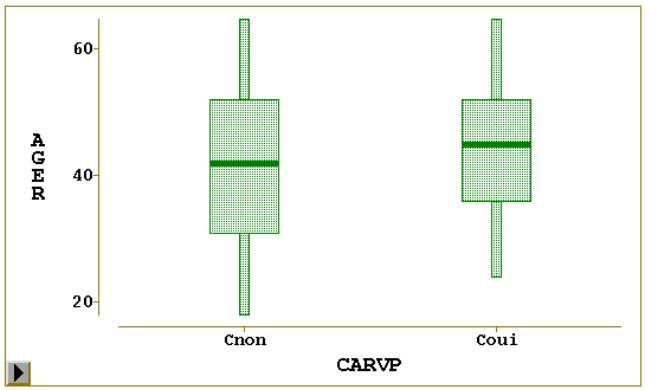

FIGURE 2 – *Banque : Diagrammes-boites illustrant les différences de distribution des âges en fonction de la possession d'une carte Visa Premier.*

des diagrammes-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.

3.3 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.

$$\overline{y} = \frac{1}{n}\sum_{\ell=1}^{m} n_\ell \overline{y_\ell}\,;$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{\ell=1}^{m} n_\ell(\overline{y_\ell}-\overline{y})^2 + \frac{1}{n}\sum_{\ell=1}^{m} n_\ell \sigma_\ell^2 = \sigma_E^2 + \sigma_R^2\,.$$

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).

3.4 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.

4 Deux variables qualitatives

4.1 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.

4.2 Représentations graphiques des profils

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*.

On appelle ℓ-ème profil-ligne l'ensemble des fréquences de la variable Y conditionnelles à la modalité x_ℓ de X (c'est-à-dire définies au sein de la sous-population Ω_ℓ de Ω associée à cette modalité). Il s'agit donc des quantités :

$$\{\frac{n_{\ell 1}}{n_{\ell+}}, \ldots, \frac{n_{\ell h}}{n_{\ell+}}, \ldots, \frac{n_{\ell c}}{n_{\ell+}}\}.$$

On définit de façon analogue le h-ème profil-colonne :

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\}.$$

La représentation graphique des profils-lignes ou des profils-colonnes, au moyen, par exemple, de diagrammes en barre parallèles (*mosaïc plot*), donne alors une idée assez précise de la variation conjointe des deux variables.

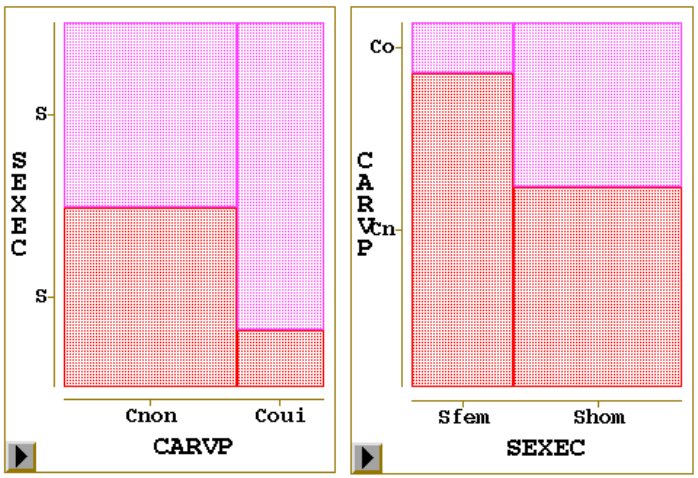

FIGURE 3 – *Banque : Diagrammes en barres des profils lignes et colonnes (mosaïque plot) de la table de contingence croisant le sexe et la possession de la carte Visa Premier. La superficie de chaque case est en plus proportionnelle à l'effectif de la cellule associée.*

4.3 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.

4.3.1 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.

4.3.2 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 : $0 \leq T \leq C \leq 1$.

Enfin, la p-valeur d'un test d'indépendance (test du χ^2) est aussi utilisée pour comparer des liaisons entre variables.

5 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.

5.1 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.

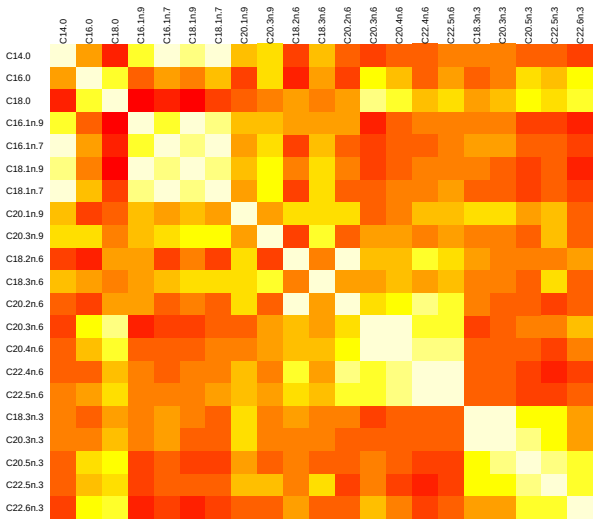

FIGURE 4 – *Souris : représentation graphique des corrélations entre les variables de concentration de lipides par des intensités de couleur.*

5.2 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.

5.3 La matrice des coefficients de Tschuprow (ou de Cramer)

Considérons maintenant le cas où l'on étudie simultanément plusieurs variables qualitatives (p variables, $p \geq 3$). La matrice des coefficients de Tschuprow est la matrice carrée d'ordre p, symétrique, comportant des 1 sur la diagonale et, en dehors de la diagonale, les coefficients de Tschuprow entre les variables prises deux à deux. Il s'agit donc d'une matrice du même type que la matrice des corrélations (elle est d'ailleurs, elle aussi, semi définie positive), et son utilisation pratique est analogue. Notons que l'on peut, de la même façon, utiliser les coefficients de Cramer au lieu des coefficients de Tschuprow.

5.4 Le tableau de Burt

Le tableau de Burt est une généralisation particulière de la table de contingence dans le cas où l'on étudie simultanément p variables qualitatives. Notons $X^1, \ldots, X^p$ ces variables, appelons c_j le nombre de modalités de $X^j, j = 1, \ldots, p$ et posons $c = \sum_{j=1}^{p} c_j$. Le tableau de Burt est en fait une matrice carrée $c \times c$, constituée de p^2 sous–matrices. Chacune des p sous–matrices diagonales est relative à l'une des p variables ; la $j^{\text{ième}}$ d'entre elles est carrée d'ordre c_j, diagonale, et comporte sur la diagonale les effectifs marginaux de X^j. La sous–matrice figurant dans le bloc d'indice $(j, j'), j \neq j'$, est la table de contingence construite en mettant X^j en lignes et $X^{j'}$ en colonnes ; le tableau de Burt est donc symétrique. Il apparaît en fait comme l'analogue qualitatif du tableau des nuages.

Introduction à la Statistique exploratoire multidimensionnelle

Résumé

Cette vignette fait suite à celles, plus élémentaires, de Statistique descriptive unidimensionnelle, bidimensionnelle et multidimensionnelle pour aborder les principales méthodes factorielles de réduction de dimension et de représentation optimale ainsi que celles de classification non supervisée.

Plan du cours :

- *Introduction*
- *Analyse en Composantes Principales*
- *Analyse Canonique des Corrélations*
- *Analyse Factorielle Discriminante*
- *Analyse Factorielle des Correspondances*
- *Analyse Factorielle des Correspondances Multiple*
- *Positionnement Multidimensionnel*
- *Classification non supervisée*
- *Factorisation par matrices non négatives (NMF)*
- *Compléments d'algèbre linéaire*

1 Historique

Les bases théoriques de ces méthodes sont anciennes et sont principalement issues de "psychomètres" américains : Spearman (1904) et Thurstone (1931, 1947) pour l'Analyse en Facteurs, Hotteling (1935) pour l'Analyse en Composantes Principales et l'Analyse Canonique, Hirschfeld (1935) et Guttman (1941, 1959) pour l'Analyse des Correspondances. Pratiquement, leur emploi ne s'est généralisé qu'avec la diffusion des moyens de calcul dans le courant des années 60. Sous l'appellation *"Multivariate Analysis"* elles poursuivent des objectifs sensiblement différents à ceux qui apparaîtront en France. Un individu ou unité statistique n'y est souvent considéré que pour l'information qu'il apporte sur la connaissance des liaisons entre variables au sein d'un échantillon statistique dont la distribution est le plus souvent soumise à des hypothèses de normalité.

En France, l'expression *"Analyse des Données"* recouvre les techniques ayant pour objectif la *description statistique des grands tableaux* (n lignes, où n varie de quelques dizaines à quelques milliers, p colonnes, où p varie de quelques unités à quelques dizaines). Ces méthodes se caractérisent par une utilisation *intensive* de l'ordinateur, leur objectif *exploratoire* et une absence quasi systématique d'hypothèses de nature *probabiliste* au profit des propriétés et résultats de géométrie euclidienne. Elles insistent sur les représentations graphiques en particulier de celles des individus qui sont considérés au même titre que les variables.

Depuis la fin des années 1970, de nombreux travaux ont permis de rapprocher ou concilier les deux points de vue en introduisant, dans des espaces multidimensionnels appropriés, les outils probabilistes et la notion de *modèle*, usuelle en statistique *inférentielle*. Les techniques se sont ainsi enrichies de notions telles que l'estimation, la convergence, la stabilité des résultats, le choix de critères...

L'objectif essentiel de ces méthodes est l'aide à la compréhension de volumes de données souvent considérables. Réduction de dimension, représentation graphique optimale, recherche de facteurs ou variables latentes... sont des formulations équivalentes.

2 Méthodes

Les méthodes de *Statistique exploratoire multidimensionnelle* se classifient selon leur objectif (réduction de dimension ou classification) et le type des données à analyser (quantitatives et/ou qualitatives) :

- Description et réduction de dimension (méthodes factorielles) :
  1. Analyse en Composantes Principales (p variables quantitatives),
  2. Analyse Factorielle Discriminante (p variables quantitatives, 1 variable qualitative),
  3. Analyse Factorielle des Correspondances simple (2 variables qualitatives) et Multiple (p variables qualitatives),
  4. Analyse Canonique (p et q variables quantitatives),

5. *Multidimensional Scaling* (M.D.S.) ou positionnement multidimensionnel ou analyse factorielle d'un tableau de distances.

   Toutes les précédentes méthodes sont basées sur des outils classiques de géométrie euclidienne qui sont développés dans les rappels et compléments d'algèbre linéaire.

6. *Non negative Matrix Factorisation* ou NMF.

- Méthodes de classification :
  1. Classification ascendante hiérarchique,
  2. Algorithmes de réallocation dynamique,
  3. Cartes de Kohonen (réseaux de neurones).

Les références introductives les plus utiles pour ce cours sont : Bouroche & Saporta (1980)[2], Jobson (1992)[3], Droesbeke, Fichet & Tassi (1992)[1], Mardia, Kent & Bibby (1979)[5], Saporta (2006)[6], Lebart, Morineau & Piron (1995)[4].

Références

[1] P.C. Besse et A. Pousse, *Extension des analyses factorielles*, Modèles pour l'Analyse des Données Multidimensionnelles (J.J. Droesbeke et al., réds.), Economica, 1992, p. 129–158.

[2] J.M. Bouroche et G. Saporta, *L'Analyse des Données*, Que Sais-je, PUF, 1980.

[3] J.D. Jobson, *Applied Multivariate Data Analysis*, t. II : Categorical and multivariate methods, Springer-Verlag, 1992.

[4] L. Lebart, A. Morineau et M. Piron, *Statistique exploratoire multidimensionnelle*, Dunod, 1995.

[5] K.V. Mardia, J.T. Kent et J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[6] G. Saporta, *Probabilités, Analyse des Données et Statistique*, deuxième éd., Technip, 2006.

Analyse en Composantes Principales (ACP)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes. Construction du modèle statistique associé, estimation. Représentations graphiques des individus, des variables et simultanée ; qualité de représentation.

Travaux pratiques *de complexité croissante par l'études de données de températures puis de données socio-économiques cubiques.*

Retour au plan du cours.

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 (cf. l'introduction élémentaire à l'ACP). 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 et graphiques, 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)[2] 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 illustrée dans ce chapitre à travers l'étude de données élémentaires. Elles sont constituées 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 ; elles sont homogènes et il est inutile de les réduire.

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.

D'un point de vue plus "mathématique", l'ACP correspond à l'approximation d'une matrice (n, p) par une matrice de même dimensions mais de rang $q < p$ (cf. rappels d'algèbre linéaire) ; q étant souvent de petite valeur 2, 3 pour la construction de graphiques facilement compréhensibles.

2 Espaces vectoriels

2.1 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 :

$$\forall i = 1, \ldots, n \ : w_i \ > \ 0 \text{ et } \sum_{i=1}^{n} w_i = 1 \ ;$$

$$\forall i = 1, \ldots, n \ : x_i^j \ = \ X^j(i), \text{ mesure de } X^j \text{ sur le } i^{ème} \text{ individu.}$$

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*.

2.2 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 :	$\overline{\mathbf{X}}$	$=$	$\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$.
Écart-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 :	$\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\left\|\mathbf{x}^j\right\|_{\mathbf{D}}\left\|\mathbf{x}^k\right\|_{\mathbf{D}}}$	$=$	$\cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$.

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*.

2.3 Objectifs

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.

3 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 $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

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*, le modèle étant dit *fonctionnel*. Ceci s'écrit en résumé :

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

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.

Si les $\mathbf{z}_i$ sont considérés comme *aléatoires*, le modèle est alors dit *structurel* ; on suppose que $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ est un échantillon statistique i.i.d. Les unités statistiques jouent des rôles symétriques, elles ne nous intéressent que pour l'étude des relations entre les variables. On retrouve alors le principe de l'analyse en facteurs (ou en facteurs communs et spécifiques, ou *factor analysis*).

3.1 Estimation

PROPOSITION 1. — *L'estimation des paramètres de (1) est fournie par l'ACP de* $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ *c'est-à-dire par la décomposition en valeurs singulières de* $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ *:*

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

Preuve

Sans hypothèse sur la distribution de l'erreur, une estimation par les moindres carrés conduit à résoudre le problème :

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

Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$.

$$\sum_{i=1}^n w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^n w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{x}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2 \ ;$$

le problème (2) conduit alors à prendre $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et devient équivalent à résoudre :

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p},\ \text{rang}(\mathbf{Z}) = q \right\}. \tag{3}$$

La fin de la preuve est une conséquence immédiate du théorème d'approximation matricielles (cf. rappels d'algèbre linéaire). □

- 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 :

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

Remarques

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

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

2. Les espaces principaux sont uniques sauf, éventuellement, dans le cas de valeurs propres multiples.
3. 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*.

Sous l'hypothèse que la distribution de l'erreur est gaussienne, une estimation par maximum de vraisemblance conduit à la même solution.

3.2 Autre définition

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 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* ;

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

- 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*.

4 Graphiques

4.1 Individus

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

4.1.1 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. — *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.*

4.1.2 Qualités

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

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

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 :

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

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

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

Pour éviter de consulter un tableau qui risque d'être volumineux (n lignes), les étiquettes de chaque individu sont affichées 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.

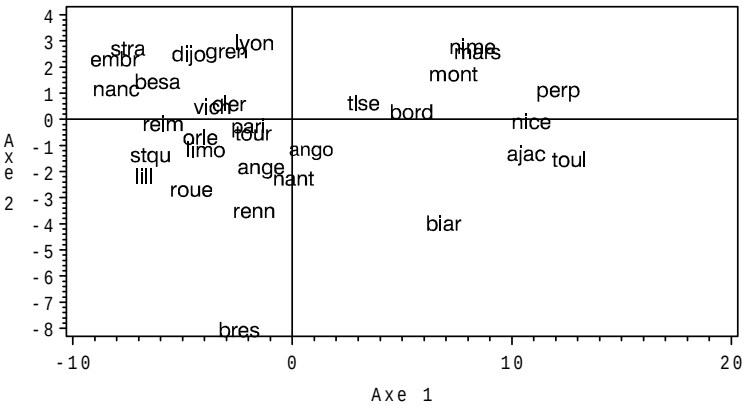

FIGURE 1 – *Températures : premier plan des individus.*

4.1.3 Contributions

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

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

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" ou atypiques 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.

4.1.4 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 :

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

4.2 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).

4.2.1 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 :

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

PROPOSITION 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 $\mathbf{V}\mathbf{\Lambda}^{1/2}$.*

4.2.2 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 :

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

4.2.3 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 $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

4.2.4 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.

4.3 Biplot

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

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

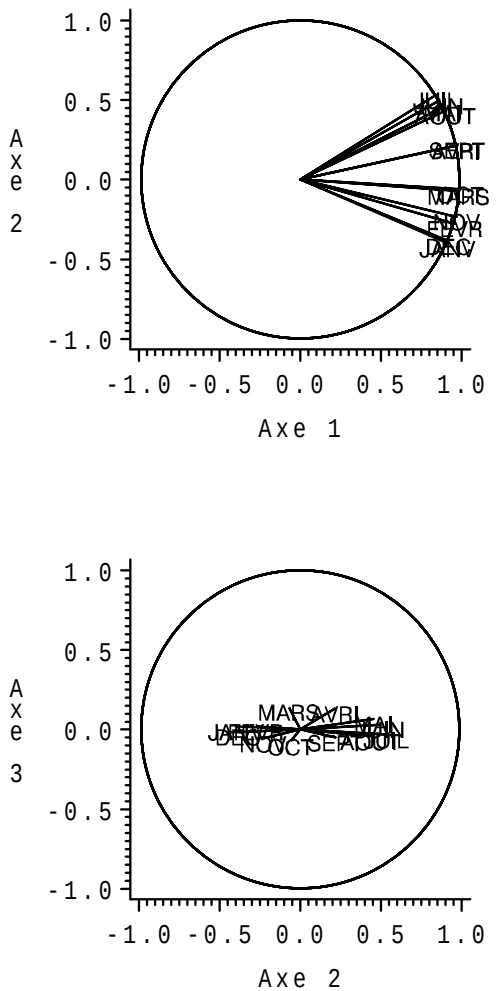

FIGURE 2 – *Températures : Premier et deuxième plan des variables.*

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.

1. 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 $X^j(\omega_i)$;
2. 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

1. 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.
2. 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. L'usage théoriquement abusif fait finalement superposer les deux représentations isométriques lignes et colonnes.

5 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.

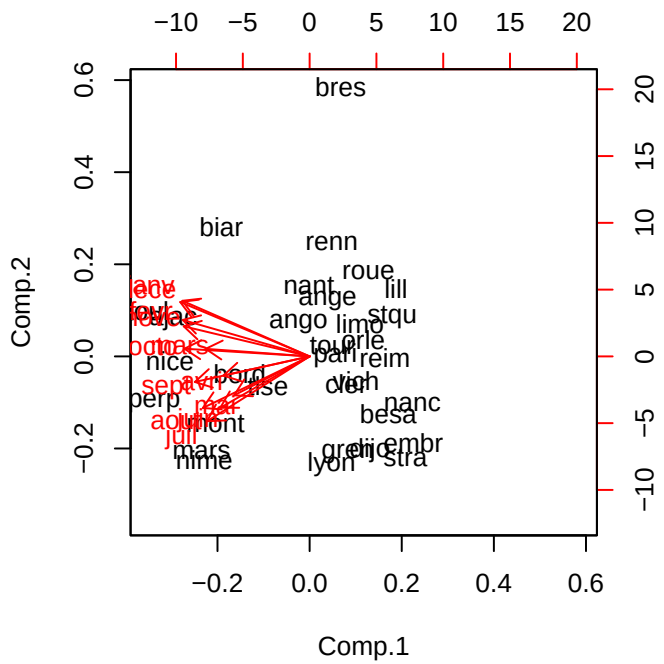

FIGURE 3 – Températures : Représentation simultanée ou biplot du premier plan.

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.

5.1 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.

5.2 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.

5.3 Éboulis

C'est le graphique (figures 4) 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 $\widehat{E_q}$.

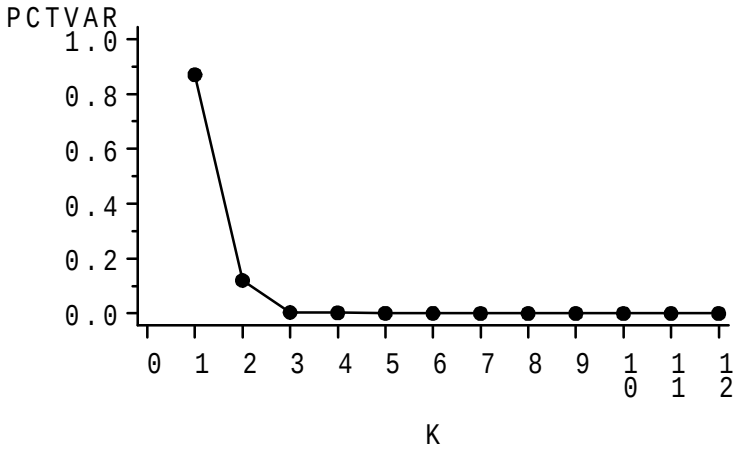

FIGURE 4 – *Températures : éboulis des valeurs propres.*

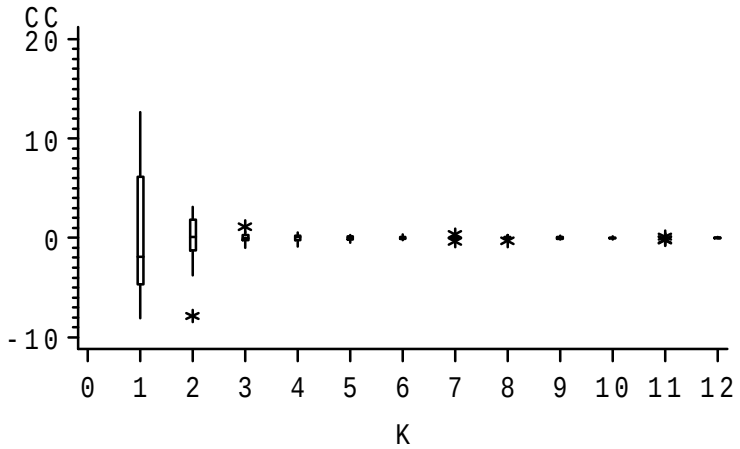

FIGURE 5 – *Températures : composantes en boîtes.*

5.4 Diagrammes boîtes

Un graphique (figure 5) présentant, en parallèle, les diagrammes boîtes 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.

5.5 Stabilité

La présentation de l'ACP, comme résultat de l'estimation d'un modèle, offre une autre approche au problème du choix de dimension. La qualité des estimations est évaluée de façon habituelle en statistique par un risque moyen quadratique définissant un critère de stabilité du sous-espace de représentation. Il est défini comme l'espérance d'une distance entre le modèle "vrai" et l'estimation qui en est faite. Besse (1992)[1] propose d'étudier la qualité de l'estimation du sous-espace de représentation $\widehat{E_q}$ en considérant la fonction perte :

$$L_q = Q(E_q, \widehat{E_q}) = \frac{1}{2}\left\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\right\|^2_{\mathbf{M,D}} = q - \mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q},$$

où Q mesure la distance entre deux sous-espaces par la distance usuelle entre les matrices de projection qui leur sont associées. C'est aussi la somme des carrés des coefficients de corrélation canonique entre les ensembles de composantes ou de variables principales qui engendrent respectivement E_q et son estimation $\widehat{E_q}$.

Un risque moyen quadratique est alors défini en prenant l'espérance de la fonction perte :

$$R_q = EQ(E_q, \widehat{E_q}). \tag{4}$$

Sans hypothèse sur la distribution de l'erreur, seules des techniques de réechantillonnage (bootstrap, jackknife) permettent de fournir une estimation de ce risque moyen quadratique. Leur emploi est justifié, car le risque est invariant par permutation des observations, mais coûteux en temps de calcul.

On se pose donc la question de savoir pour quelles valeurs de q les représentations graphiques sont fiables, c'est-à-dire stables pour des fluctuations de l'échantillon. Besse (1992)[1] propose d'utiliser une approximation de l'estimateur par jackknife ; elle fournit, directement à partir des résultats de l'A.C.P. (valeurs propres et composantes principales), une estimation satisfaisante du risque :

$$\widehat{R_{JKq}} = \widehat{R_{\mathbf{P}q}} + O((n-1)^{-2}).$$

$\widehat{R_{\mathbf{P}q}}$ est une approximation analytique de l'estimateur jackknife qui a pour expression :

$$\widehat{R_{\mathbf{P}q}} = \frac{1}{n-1} \sum_{k=1}^{q} \sum_{j=q+1}^{p} \frac{\frac{1}{n}\sum_{i=1}^{n}(c_i^k)^2(c_i^j)^2}{(\lambda_j - \lambda_k)^2} \tag{5}$$

où c_i^j désigne le terme général de la matrice des composantes principales $\mathbf{C}$.

Ce résultat souligne l'importance du rôle que joue l'écart $(\lambda_q - \lambda_{q+1})$ dans la stabilité du sous-espace de représentation. Le développement est inchangé dans le cas d'une ACP réduite ; de plus, il est valide tant que

$$n > \frac{\|\mathbf{S}\|_2^2}{\inf\{(\lambda_k - \lambda_{k+1}); k = 1, \ldots, q\}}.$$

La figure 6 montrent la stabilité du sous-espace de représentation en fonction de la dimension q pour l'A.C.P. des données de températures. Comme souvent, le premier axe est très stable tandis que le premier plan reste fiable. Au delà, les axes étant très sensibles à toute perturbation des données, ils peuvent être associés à du bruit. Ces résultats sont cohérents avec les deux critères graphiques précédents mais souvent, en pratique, le critère de stabilité conduit à un choix de dimension plus explicite.

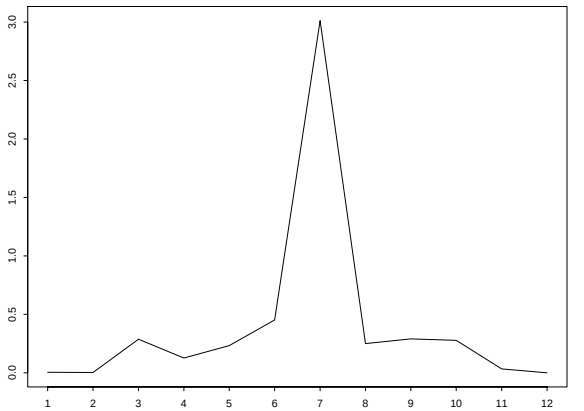

FIGURE 6 – *Températures : stabilité des sous-espaces.*

6 Interprétation

Les macros SAS utilisées, 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 d'aide à l'interprétation (contributions, cosinus carrés). 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 !

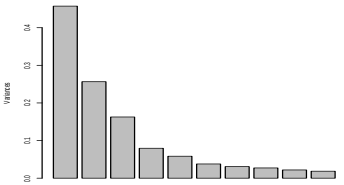

FIGURE 7 – *Souris : éboulis des dix premières valeurs propres de l'ACP.*

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.

7 Exemple : Données génomiques

L'éboulis des premières valeurs propres (figure 7) conduit à considérer trois dimensions représentant environ les deux tiers de l'inertie globale mais nous limiterons l'interprétation un peu sommaire au premier plan.

La figure 8 représente conjointement souris et gènes (biplot). 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 le plan factoriel. Pour des données plus volumineuses (puces pangénomiques) d'autres outils (version parcimonieuse ou creuse de l'ACP) sont à considérer.

Le premier plan (Fig. 8) doit être interprété globalement puisque sa deuxième bissectrice sépare exactement les souris `WT` des souris `PPAR`. Les gènes à coordonnées négatives sur l'axe 2 et positives sur l'axe1 sont sensiblement plus exprimés chez les souris `WT`, en particulier `CYP3A11`, `CYP4A10`, `CYP4A14`, `THIOL`, `PMDCI`, `GSTpi2`, `L.FABP` et `FAS` (négatif sur les deux axes). À l'inverse, les gènes à forte coordonnée négative sur l'axe 2 s'expriment davantage chez les souris `PPAR`, par exemple, `S14` et `CAR1`. Ceci est en partie connu des biologistes.

Sur cette représentation, seules les souris `WT` présentent des comportement

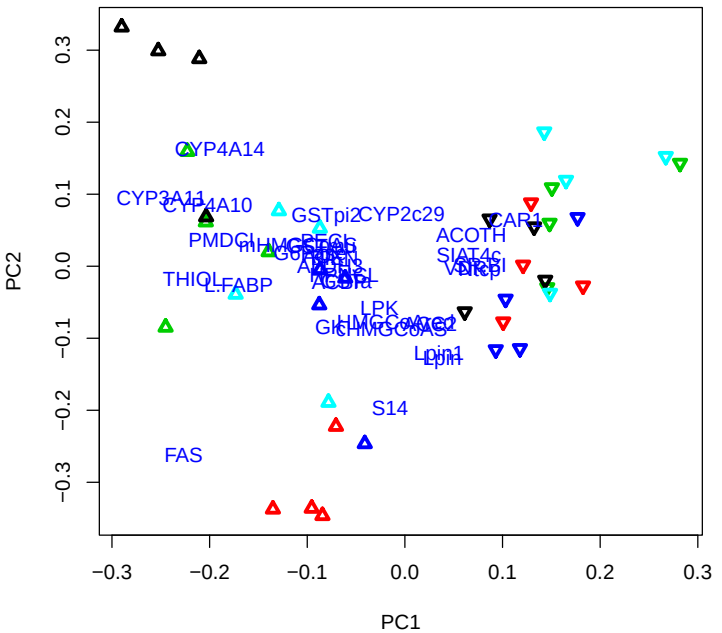

FIGURE 8 – *Représentations conjointe sur le premier plan principal. Les souris identifiés par leur génotype (*`WT` *triangles vers le haut,* `PPAR` *vers le bas) et leur régime (principalement noir-*`dha` *et rouge-*`efad`*).*

sensiblement différents au regard des régimes. Le phénomène le plus marquant est l'opposition, chez ces souris `WT`, entre les régimes `dha` (triangles noirs), dont les coordonnées sont toutes positives, et `efad` (triangles rouges), 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` ; dans le second cas (régime `efad` chez les mêmes souris), il s'agit des gènes `FAS` et `S14`. 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 la littérature.

Références

[1] P.C. Besse, *PCA stability and choice of dimensionality*, Statistics & Probability Letters **13** (1992), 405–410.

[2] I. Jolliffe, *Principal Component Analysis*, 2nd edition éd., Springer-Verlag, 2002.

Analyse factorielle discriminante (AFD)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de variables quantitatives et d'une variable qualitative. Construction du modèle statistique associé, estimation. Représentation graphique optimale des classes des individus, liens avec d'autres définitions de l'AFD.

Travaux pratiques *de complexité croissante par l'études de données socio-économiques.*

Retour au plan du cours.

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 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 :

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

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 (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 $\mathbf{z}_ell$ 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 2. — *L'estimation des paramètres E_q et $\mathbf{z}_\ell$ du modèle 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 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

5.1 Les insectes de Lubitsch

Cette méthode est illustrée par une comparaison des sorties graphiques issues d'une ACP et d'une AFD. Les données décrivent trois classes d'insectes

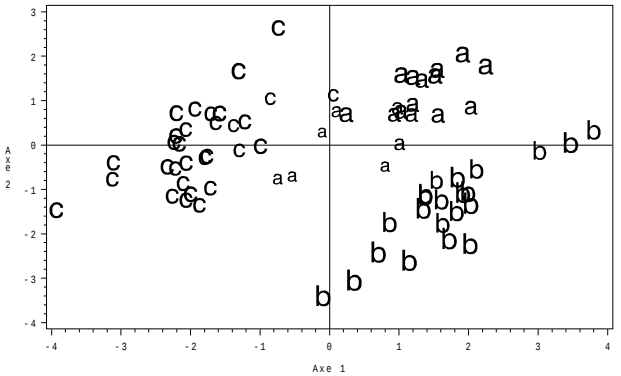

FIGURE 1 – *Insectes : premier plan factoriel de l'ACP.*

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 "scolaire", comme les fameux iris de Fisher conduit à une discrimination assez évidente. La comparaison entre l'ACP et l'AFD met clairement en évidence le rôle de la distance $\mathbf{S}_R^{-1}$ que la forme des nuages de chaque classe en analyse discriminante.

5.2 Données génomiques

Les données génomiques pose évidemment des problèmes à l'analyse discriminante ; le grand nombre de gènes/variables par rapport au nombre de souris/individus rend impossible l'inversion de la matrice des covariances intra-classes. Aussi, en s'aidant de la sélection de variables suggérée par l'analyse en composantes principales, une analyse factorielle discriminante a été calculée sur les seules souris sauvages (`WR`) pour qui les régimes apparaissaient déjà bien différenciés sur l'ACP. Les variables ne sont pas représentées mais les rapprochements déjà évoqués pour l'ACP sont confirmés et précisés.

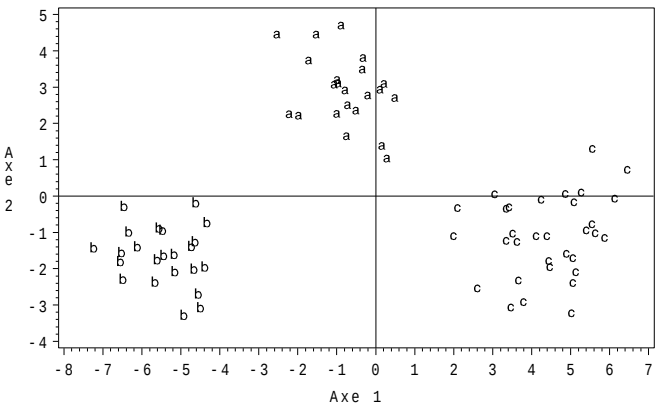

FIGURE 2 – *Insectes : premier plan factoriel de l'AFD.*

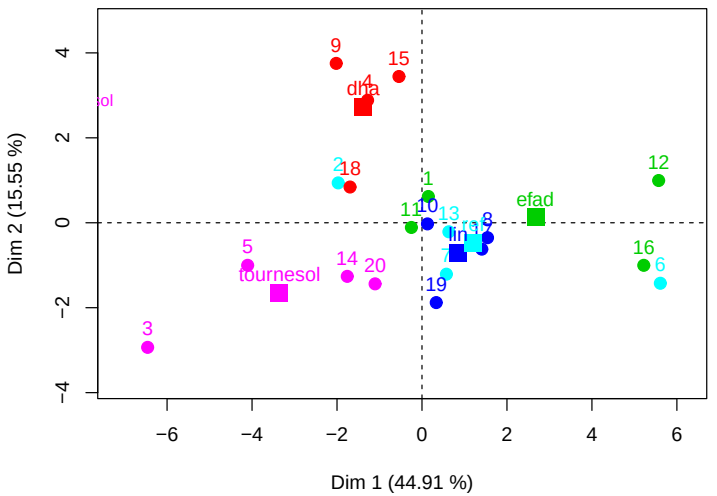

FIGURE 3 – *Souris : Les souris de génotype* `WT` *dans le premier plan factoriel de l'AFD calculée avec une sélection de variables d'expression de gènes conditionnellement au régime.*

Analyse Factorielle des Correspondances (AFC)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique d'une table de contingence définie par deux variables qualitatives. Définition à partir de l'analyse en composantes principales des profils. Définition du modèle statistique associé, estimation. Représentation graphique simultanée des modalités des variables.

Travaux pratiques *de complexité croissante par l'études de données élémentaires.*

Retour au plan du cours.

1 Introduction

1.1 Données

On considère dans cette vignette deux variables qualitatives observées simultanément sur n individus affectés de poids identiques $1/n$. On suppose que la première variable, notée X, possède r modalités notées $x_1, \ldots, x_\ell, \ldots, x_r$, et que la seconde, notée Y, possède c modalités notées $y_1, \ldots, y_h, \ldots, y_c$.

La table de contingence associée à ces observations, de dimension $r \times c$, est notée $\mathbf{T}$; son élément générique est $n_{\ell h}$, effectif conjoint. Elle se présente sous la forme suivante d'une table de contingence présentée dans le Tableau 1).

1.2 Notations

Les quantités $\{n_{\ell+} = \sum_{h=1}^{c} n_{\ell h}; \ell = 1, \ldots, r\}$ et $\{n_{+h} = \sum_{\ell=1}^{r} n_{\ell h}; h = 1, \ldots, c\}$ sont les *effectifs marginaux* vérifiant $\sum_{\ell=1}^{r} n_{\ell+} = \sum_{h=1}^{c} n_{+h} = n$. De façon analogue, on définit les notions de *fréquences conjointes* ($f_{\ell h} = n_{\ell h}/n$) et de *fréquences marginales* rangées dans les vecteurs :

TABLE 1 – *Table de contingence*

	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

$$\begin{aligned} \mathbf{g}_r &= [f_{1+}, \ldots, f_{r+}]', \\ \text{et } \mathbf{g}_c &= [f_{+1}, \ldots, f_{+c}]'. \end{aligned}$$

Elles permettent de définir les matrices :

$$\begin{aligned} \mathbf{D}_r &= \text{diag}(f_{1+}, \ldots, f_{r+}), \\ \text{et } \mathbf{D}_c &= \text{diag}(f_{+1}, \ldots, f_{+c}). \end{aligned}$$

On sera également amené à considérer les profils–lignes et les profils–colonnes déduits de $\mathbf{T}$. Le ℓ-ième profil-ligne est

$$\{\frac{n_{\ell 1}}{n_{\ell+}}, \ldots, \frac{n_{\ell h}}{n_{\ell+}}, \ldots, \frac{n_{\ell c}}{n_{\ell+}}\}.$$

Il est considéré comme un vecteur de $\mathbb{R}^c$ et les r vecteurs ainsi définis sont disposés en colonnes dans la matrice $c \times r$

$$\mathbf{A} = \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}.$$

De même, le h-ième profil-colonne est

$$\{\frac{n_{1h}}{n_{+h}}, \ldots, \frac{n_{\ell h}}{n_{+h}}, \ldots, \frac{n_{rh}}{n_{+h}}\},$$

vecteur de $\mathbb{R}^r$, et la matrice $r \times c$ des profils-colonnes est

$$\mathbf{B} = \frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}.$$

1.3 Liaison entre deux variables qualitatives

DÉFINITION 1. — *On dit que deux variables X et Y sont* non liées *relativement à T si et seulement si :*

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

Il est équivalent de dire que tous les profils-lignes sont égaux, ou encore que tous les profils-colonnes sont égaux.

Cette notion est cohérente avec celle d'indépendance en probabilités. En effet, soit $\Omega = \{1, \ldots, n\}$ l'ensemble des individus observés et $(\Omega, \mathcal{P}(\Omega), P)$ l'espace probabilisé associé où P est l'équiprobabilité ; $\mathcal{M}_X = \{x_1, \ldots, x_r\}$ et $\mathcal{M}_Y = \{y_1, \ldots, y_c\}$ désignent les ensembles de modalités, ou valeurs prises par les variables X et Y. On note $\widetilde{X}$ et $\widetilde{Y}$ les variables aléatoires associées aux 2 variables statistiques X et Y :

$$\begin{aligned} \widetilde{X} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), \\ \widetilde{Y} &: (\Omega, \mathcal{P}(\Omega), P) \mapsto (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y)) ; \end{aligned}$$

P_X, P_Y et P_{XY} désignent respectivement les probabilités images définies par $\widetilde{X}, \widetilde{Y}$ et le couple $(\widetilde{X}, \widetilde{Y})$ sur $(\mathcal{M}_X, \mathcal{P}(\mathcal{M}_X)), (\mathcal{M}_Y, \mathcal{P}(\mathcal{M}_Y))$ et $(\mathcal{M}_X \times \mathcal{M}_Y, \mathcal{P}(\mathcal{M}_X) \times \mathcal{P}(\mathcal{M}_Y))$; ce sont les probabilités empiriques. Alors, X et Y sont *non liées* si et seulement si $\widetilde{X}$ et $\widetilde{Y}$ sont *indépendantes en probabilité* (la vérification est immédiate).

On suppose maintenant qu'il existe une liaison entre X et Y que l'on souhaite étudier. La représentation graphique des profils-lignes ou des profils-colonnes, au moyen de diagrammes en barres parallèles, ainsi que le calcul de coefficients de liaison (Cramer ou Tschuprow) donnent une première idée de la variation conjointe des deux variables. Le test du χ^2 permet de plus de s'assurer du caractère significatif de cette liaison. Il est construit de la manière suivante :

l'hypothèse nulle est $H0$: $\widetilde{X}$ et $\widetilde{Y}$ sont indépendantes en probabilités ;

l'hypothèse alternative est $H1$: les variables $\widetilde{X}$ et $\widetilde{Y}$ ne sont pas indépendantes.

La statistique de test est alors

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

elle suit asymptotiquement (pour les grandes valeurs de n), et si l'hypothèse $H0$ est vraie, une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté. On rejette donc $H0$ (et l'on conclut au caractère significatif de la liaison) si χ^2 dépasse une valeur particulière (valeur ayant une probabilité faible et fixée a priori – en général 0,05 –être dépassée par une loi de χ^2 à $(r-1)(c-1)$ degrés de liberté).

1.4 Objectifs

Pour préciser la liaison existant entre les variables X et Y, on souhaite définir un modèle statistique susceptible de fournir des paramètres dont la représentation graphique (de type biplot) illustrera les "*correspondances*" entre les modalités de ces 2 variables. Cette approche sera développée au paragraphe 3.

Une autre approche, très courante dans la littérature francophone, consiste à définir l'Analyse Factorielle des Correspondances (AFC) comme étant le résultat d'une double Analyse en Composantes Principales

- l'ACP des profils–lignes,
- l'ACP des profils–colonnes,

relativement à la métrique dite du χ^2. Cette approche est présentée au paragraphe 2.

Remarque. — :

1. Toute structure d'ordre existant éventuellement sur les modalités de X ou de Y est ignorée par l'AFC
2. Tout individu présente une modalité et une seule de chaque variable.
3. Chaque modalité doit avoir été observée au moins une fois ; sinon, elle est supprimée.

2 Double ACP

2.1 Métriques du Chi2

Les correspondances entre modalités évoquées au paragraphe précédant se trouvent exprimées en termes de distances au sens d'une certaine métrique. Ainsi, chaque modalité x_ℓ de X est caractérisée par son profil–ligne représenté par le vecteur $\mathbf{a}^\ell$ de l'espace $\mathbb{R}^c$ muni de la base canonique (les coordonnées de $\mathbf{a}^\ell$ sont les éléments de la ℓ-ième colonne de $\mathbf{A}$). De même, chaque modalité y_h de Y est caractérisée par son profil–colonne représenté par le vecteur $\mathbf{b}^h$ de l'espace $\mathbb{R}^r$ muni de la base canonique.

Ces espaces sont respectivement munis des métriques, dites du χ^2, de matrices $\mathbf{D}_c^{-1}$ et $\mathbf{D}_r^{-1}$. Ainsi, la distance entre deux modalités x_ℓ et x_i de X s'écrit

$$\|\mathbf{a}^\ell - \mathbf{a}^i\|^2_{\mathbf{D}_c^{-1}} = \sum_{h=1}^{c} \frac{1}{f_{+h}}(a_h^\ell - a_h^i)^2,$$

et de même pour les modalités de Y. La métrique du χ^2 introduit les inverses des fréquences marginales des modalités de Y comme *pondérations* des écarts entre éléments de deux profils relatifs à X (et réciproquement) ; elle attribue donc plus de poids aux écarts correspondants à des modalités de *faible effectif* (rares) pour Y.

2.2 ACP des profils–colonnes

On s'intéresse ici à l'ACP du triplet $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$. Dans cette ACP, les "individus" sont les modalités de Y, caractérisées par les profils–colonnes de $\mathbf{T}$, pondérées par les fréquences marginales correspondantes et rangées en lignes dans la matrice $\mathbf{B}'$.

PROPOSITION 2. — *Les éléments de l'ACP de* $(\mathbf{B}', \mathbf{D}_r^{-1}, \mathbf{D}_c)$ *sont fournis par l'analyse spectrale de la matrice carrée,* $\mathbf{D}_r^{-1}$*–symétrique et semi–définie positive* $\mathbf{BA}$.

Preuve Elle se construit en remarquant successivement que :

1. le barycentre du nuage des profils–colonnes est le vecteur $\mathbf{g}_r$ des fréquence marginales de X,
2. la matrice $\mathbf{BD}_c\mathbf{B}' - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r'$ joue le rôle de la matrice des variances–covariances,
3. la solution de l'ACP est fournie par la D.V.S. de $(\mathbf{B}' - \mathbf{1}\mathbf{g}_r', \mathbf{D}_r^{-1}, \mathbf{D}_c)$, qui conduit à rechercher les valeurs et vecteurs propres de la matrice $(\mathbf{SM})$

$$\mathbf{BD}_c\mathbf{B}'\mathbf{D}_r^{-1} - \mathbf{g}_r\mathbf{D}_c\mathbf{g}_r' = \mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1} \quad (\text{ car } \mathbf{B}'\mathbf{D}_r^{-1} = \mathbf{D}_c^{-1}\mathbf{A})$$

4. les matrices $\mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$ et $\mathbf{BA}$ ont les mêmes vecteurs propres associées aux mêmes valeurs propres, à l'exception du vecteur $\mathbf{g}_r$ associé à la valeur propre $\lambda_0 = 0$ de $\mathbf{BA} - \mathbf{g}_r\mathbf{g}_r'\mathbf{D}_r^{-1}$ et à la valeur propre $\lambda_0 = 1$ de $\mathbf{BA}$.

□

On note $\mathbf{U}$ la matrice contenant les vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés de $\mathbf{BA}$. La représentation des "individus" de l'ACP réalisée fournit une représentation des modalités de la variable Y. Elle se fait au moyen des lignes de la matrice des "composantes principales" $(\mathbf{XMV})$:

$$\mathbf{C}_c = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{U}.$$

2.3 ACP des profils–lignes

De façon symétrique (ou duale), on s'intéresse à l'ACP des "individus" modalités de X ou profils–lignes (la matrice des données est $\mathbf{A}'$), pondérés par les fréquences marginales des lignes de $\mathbf{T}$ (la matrice diagonale des poids est $\mathbf{D}_r$) et utilisant la métrique du χ^2. Il s'agit donc de l'ACP de $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$.

PROPOSITION 3. — *Les éléments de l'ACP de* $(\mathbf{A}', \mathbf{D}_c^{-1}, \mathbf{D}_r)$ *sont fournis par l'analyse spectrale de la matrice carrée,* $\mathbf{D}_c^{-1}$*–symétrique et semi–définie positive* $\mathbf{AB}$.

On obtient directement les résultats en permutant les matrices $\mathbf{A}$ et $\mathbf{B}$, ainsi que les indices c et r. Notons $\mathbf{V}$ la matrice des vecteurs propres de la matrice $\mathbf{AB}$; les coordonnées permettant la représentation les modalités de la variable X sont fournies par la matrice :

$$\mathbf{C}_r = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}.$$

Sachant que $\mathbf{V}$ contient les vecteurs propres de $\mathbf{AB}$ et $\mathbf{U}$ ceux de $\mathbf{BA}$, un théorème de l'annexe (st-m-explo-alglin Compléments d'algèbre linéaire)

montre qu'il suffit de réaliser une seule analyse, car les résultats de l'autre s'en déduisent simplement :

$$\begin{aligned}\mathbf{V} &= \mathbf{AU}\mathbf{\Lambda}^{-1/2},\\ \mathbf{U} &= \mathbf{BV}\mathbf{\Lambda}^{-1/2};\end{aligned}$$

$\mathbf{\Lambda}$ est la matrice diagonale des valeurs propres (exceptée $\lambda_0 = 0$) communes aux deux ACP

$$\begin{aligned}\mathbf{C}_c &= \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{U} = \mathbf{B}'\mathbf{D}_r^{-1}\mathbf{BV}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{ABV}\mathbf{\Lambda}^{-1/2} = \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2},\\ \mathbf{C}_r &= \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V} = \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2}.\end{aligned}$$

On en déduit les formules dites de *transition* :

$$\begin{aligned}\mathbf{C}_c &= \mathbf{B}'\mathbf{C}_r\mathbf{\Lambda}^{-1/2},\\ \mathbf{C}_r &= \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2}.\end{aligned}$$

La représentation simultanée habituellement construite à partir de ces matrices (option par défaut de SAS) n'est pas a priori justifiée. On lui donnera un sens dans les paragraphes suivants.

3 Modèles pour une table de contingence

On écrit d'abord que chaque fréquence $f_{\ell h}$ de $\mathbf{T}$ correspond à l'observation d'une probabilité théorique $p_{\ell h}$; on modélise donc la table de contingence par cette distribution de probabilités. On précise ensuite le modèle en explicitant l'écriture de $p_{\ell h}$. Différents modèles classiques peuvent être considérés.

3.1 Le modèle log–linéaire

Il consiste à écrire :

$$\ln(p_{\ell h}) = \mu + \alpha_\ell + \beta_h + \gamma_{\ell h}$$

avec des contraintes le rendant identifiable. Ce modèle, très classique, est développé par ailleurs.

3.2 Le modèle d'association

Il est encore appelé RC-modèle, ou modèle de Goodman :

$$p_{\ell h} = \gamma.\alpha_\ell.\beta_h.\exp\left(\sum_{k=1}^{q}\phi_k.\mu_{\ell k}.\nu_{hk}\right).$$

Ce modèle, muni des contraintes nécessaires, permet de structurer les interactions et de faire des représentations graphiques des lignes et des colonnes de $\mathbf{T}$ au moyen des paramètres $\mu_{\ll k}$ et ν_{hk}. Ces paramètres peuvent être estimés par maximum de vraisemblance ou par moindres carrés.

3.3 Le modèle de corrélation

On écrit ici :

$$p_{\ell h} = p_{\ell +}p_{+h} + \sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k, \tag{1}$$

avec $q \leq \inf(r-1, c-1)$, $\lambda_1 \geq \cdots \geq \lambda_q > 0$ et sous les contraintes d'identifiabilité suivantes :

$$\begin{aligned}\sum_{\ell=1}^{r}u_\ell^k = \sum_{h=1}^{c}v_h^k &= 0,\\ \mathbf{u}^{k'}\mathbf{D}_r^{-1}\mathbf{u}^j = \mathbf{v}^{k'}\mathbf{D}_c^{-1}\mathbf{v}^j &= \delta_{kj}.\end{aligned}$$

Remarque. — :

1. Le modèle (1) ci-dessus est équivalent au modèle considéré par Goodman :

$$p_{\ell h} = p_{\ell +}p_{+h}\left(1 + \sum_{k=1}^{q}\sqrt{\lambda_k}\xi_\ell^k\eta_h^k\right), \tag{2}$$

   moyennant une homothétie sur les paramètres.
2. La quantité $\sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k$ exprime l'écart à l'indépendance pour la cellule considérée.

3. Le modèle suppose que cet écart se décompose dans un sous–espace de dimension $q < \min(c-1, r-1)$.
4. Les estimations des paramètres $p_{\ell+}, p_{+h}, \lambda_k, \mathbf{u}^k, \mathbf{v}^k$ peuvent être réalisées par maximum de vraisemblance [1] ou par moindres carrés. Dans le contexte de la statistique descriptive, qui est celui de ce cours, il est naturel de retenir cette dernière solution.

3.4 Estimation Moindres Carrés dans le modèle de corrélation

3.4.1 Critère

Considérons les espaces $\mathbb{R}^c$ et $\mathbb{R}^r$ munis de leur base canonique et de leur métrique du χ^2 respectives et notons $\mathbf{P}$ le tableau des probabilités théoriques définies selon le modèle (1). Le critère des moindres carrés s'écrit alors :

$$\min_{\mathbf{P}} \left\| \frac{1}{n}\mathbf{T} - \mathbf{P} \right\|^2_{\mathbf{D_r^{-1}D_c^{-1}}}. \tag{3}$$

3.4.2 Estimation

PROPOSITION 4. — *L'estimation des paramètres de (1) en résolvant (3) est fournie par la D.V.S. de* $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$ *à l'ordre* q. *Les probabilités marginales* $p_{\ell+}$ *et* p_{+h} *sont estimées par* $f_{\ell+}$ *et* f_{+h} *tandis que les vecteurs* $\mathbf{u}^k$ *(resp.* $\mathbf{v}^k$*) sont vecteurs propres de la matrice* $\mathbf{BA}$ *(resp.* $\mathbf{AB}$*) associés aux valeurs propres* λ_k.

On obtient ainsi, d'une autre façon, l'AFC de la table de contingence $\mathbf{T}$.

Preuve Elle se construit à partir de la D.V.S. de $(\frac{1}{n}\mathbf{T}, \mathbf{D}_c^{-1}, \mathbf{D}_r^{-1})$:

$$\frac{1}{n}t_\ell^h = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k} u_\ell^k v_h^k,$$

où les vecteurs $\mathbf{u}^k$ (resp. $\mathbf{v}^k$) sont vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés (resp. $\mathbf{D}_c^{-1}$–orthonormés) de la matrice

$$\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1}\frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1} = \mathbf{BA} \text{ (resp. } \frac{1}{n}\mathbf{T}'\mathbf{D}_r^{-1}\frac{1}{n}\mathbf{T}\mathbf{D}_c^{-1} = \mathbf{AB}),$$

associés aux valeurs propres λ_k.

De plus, le vecteur $\mathbf{g}_r = \mathbf{u}^0$ (resp. $\mathbf{g}_c = \mathbf{v}^0$) est vecteur propre $\mathbf{D}_r^{-1}$–normé (resp. $\mathbf{D}_c^{-1}$–normé) de la matrice $\mathbf{BA}$ (resp. $\mathbf{AB}$) associé à la valeur propre $\lambda_0 = 1$. Enfin, les matrices $\mathbf{AB}$ et $\mathbf{BA}$ sont stochastiques [2] et donc les valeurs propres vérifient :

$$1 = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_q > 0.$$

En identifiant les termes, l'approximation de rang $(q+1)$ de la matrice $\mathbf{P}$ s'écrit donc :

$$\widehat{\mathbf{P}}_q = \mathbf{g}_r\mathbf{g}_c' + \sum_{k=1}^{q} \sqrt{\lambda_k}\mathbf{u}^k\mathbf{v}^{k'}$$

et les propriétés d'orthonormalité des vecteurs propres assurent que les contraintes du modèle sont vérifiées.

□

4 Représentations graphiques

4.1 Biplot

La décomposition de la matrice $\frac{1}{n}\mathbf{T}$ se transforme encore en :

$$\frac{f_{\ell h} - f_{\ell+}f_{+h}}{f_{\ell+}f_{+h}} = \sum_{k=0}^{\min(r-1,c-1)} \sqrt{\lambda_k}\frac{u_\ell^k}{f_{\ell+}}\frac{v_h^k}{f_{+h}}.$$

En se limitant au rang q, on obtient donc, pour chaque cellule (ℓ, h) de la table $\mathbf{T}$, une approximation de son écart relatif à l'indépendance comme produit scalaire des deux vecteurs

$$\frac{u_\ell^k}{f_{\ell+}}\lambda_k^{1/4} \text{ et } \frac{v_h^k}{f_{+h}}\lambda_k^{1/4},$$

termes génériques respectifs des matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/4} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/4},$$

1. On suppose alors que les n $p_{\ell h}$ sont les paramètres de lois de Poisson indépendantes conditionnellement à leur somme qui est fixée et égale à n.

2. Matrice réelle, carrée, à termes positifs, dont la somme des termes de chaque ligne (ou chaque colonne) vaut 1.

qui sont encore les estimations des vecteurs ξ_ℓ et η_h du modèle 2. Leur représentation (par exemple avec $q = 2$) illustre alors la *correspondance* entre les deux modalités x_ℓ et y_h : lorsque deux modalités, éloignées de l'origine, sont voisines (resp. opposées), leur produit scalaire est de valeur absolue importante ; leur cellule conjointe contribue alors fortement et de manière positive (resp. négative) à la dépendance entre les deux variables.

L'AFC apparaît ainsi comme la meilleure reconstitution des fréquences $f_{\ell h}$, ou encore la meilleure représentation des écarts relatifs à l'indépendance. La représentation simultanée des modalités de X et de Y se trouve ainsi pleinement justifiée.

4.2 Double ACP

Chacune des deux ACP réalisée permet une représentation des "individus" (modalités) approchant, au mieux, les distances du χ^2 entre les profils–lignes d'une part, les profils–colonnes d'autre part. Les coordonnées sont fournies cette fois par les matrices (de composantes principales)

$$\mathbf{C}_r = \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2} \text{ et } \mathbf{C}_c = \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}.$$

Même si la représentation simultanée n'a plus alors de justification, elle reste couramment employée. En fait, les graphiques obtenus diffèrent très peu de ceux du biplot ; ce dernier sert donc de "caution" puisque les interprétations des graphiques sont identiques. On notera que cette représentation issue de la double ACP est celle réalisée par la plupart des logiciels statistiques (c'est en particulier le cas de SAS).

4.3 Représentations barycentriques

D'autres représentations simultanées, appelées barycentriques, sont proposées en utilisant les matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda}^{1/2} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda},$$

ou encore les matrices

$$\mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2}.$$

Si l'on considère alors, par exemple, la formule de transition

$$\mathbf{C}_r = \mathbf{A}'\mathbf{C}_c\mathbf{\Lambda}^{-1/2} \Longleftrightarrow \mathbf{C}_r\mathbf{\Lambda}^{1/2} = \mathbf{A}'\mathbf{C}_c \Longleftrightarrow \mathbf{D}_r^{-1}\mathbf{U}\mathbf{\Lambda} = \mathbf{A}'\mathbf{D}_c^{-1}\mathbf{V}\mathbf{\Lambda}^{1/2},$$

on voit que dans la seconde des représentations ci–dessus, chaque modalité x_ℓ de X est représentée par un vecteur qui est barycentre de l'ensemble des vecteurs associés aux modalités de Y, chacun d'eux ayant pour poids l'élément correspondant du l-ième profil–ligne. Là encore, la représentation simultanée s'en trouve parfaitement justifiée. Malheureusement, dans la pratique, les représentations barycentriques sont souvent illisibles ; elles sont, de ce fait, très peu utilisées.

4.4 Autre représentation

La pratique de l'AFC montre que l'interprétation des graphiques est toujours la même, quelle que soit la représentation simultanée choisie parmi les 3 ci–dessus.

On peut ainsi envisager d'utiliser, pour une représentation simultanée des modalités de X et de Y, les coordonnées fournies respectivement par les lignes des matrices

$$\mathbf{D}_r^{-1}\mathbf{U} \text{ et } \mathbf{D}_c^{-1}\mathbf{V}.$$

L'interprétation du graphique sera toujours la même et les matrices ci–dessus, outre leur simplicité, présentent l'avantage de conduire a une représentation graphique qui reste invariante lorsque l'on utilise la technique d'Analyse Factorielle des Correspondances Multiples sur les données considérées ici.

4.5 Aides à l'interprétation

Les qualités de représentation dans la dimension choisie et les contributions des modalités de X ou de Y se déduisent aisément de celles de l'ACP Ces quantités sont utilisées à la fois pour choisir la dimension de l'AFC et pour interpréter ses résultats dans la dimension choisie.

4.5.1 Mesure de la qualité globale

Pour une dimension donnée q ($1 \leq q \leq d = \inf(r-1, c-1)$), la qualité globale des représentations graphiques en dimension q se mesure par le rapport entre la somme des q premières valeurs propres de l'AFC et leur somme complète de 1 à d.

Compte–tenue de la propriété $\sum_{k=1}^{d} \lambda_k = \Phi^2$ (voir en 6.1), la qualité de la

représentation dans la k–ième dimension s'écrit

$$\frac{n\lambda_k}{\chi^2}.$$

On parle encore de part du khi–deux expliquée par la k–ième dimension (voir les sorties du logiciel SAS).

4.5.2 Mesure de la qualité de chaque modalité

Pour chaque modalité de X (resp. de Y), la qualité de sa représentation en dimension q se mesure par le cosinus carré de l'angle entre le vecteur représentant cette modalité dans $\mathbb{R}^c$ (resp. dans $\mathbb{R}^r$) et sa projection $\mathbf{D}_c^{-1}$–orthogonale (resp. $\mathbf{D}_r^{-1}$–orthogonale) dans le sous–espace principal de dimension q.

Ces cosinus carrés s'obtiennent en faisant le rapport des sommes appropriées des carrés des coordonnées extraites des lignes de $\mathbf{C}_r$ (resp. de $\mathbf{C}_c$).

4.5.3 Contributions à l'inertie totale

L'inertie totale (en dimension d) du nuage des profils–lignes (resp. des profils–colonnes) est égale à la somme des d valeurs propres. La part due au i–ième profil–ligne (resp. au j–ième profil–colonne) valant $f_{\ell+}\sum_{k=1}^{d}(c_{r\ell}^k)^2$ (resp. $f_{+h}\sum_{k=1}^{d}(c_{ch}^k)^2$), les contributions à l'inertie totale s'en déduisent immédiatement.

4.5.4 Contributions à l'inertie selon chaque axe

Il s'agit de quantités analogues à celles ci–dessus, dans lesquelles il n'y a pas de sommation sur l'indice k. Ces quantités sont utilisées dans la pratique pour sélectionner les modalités les plus importantes, c'est–à–dire celles qui contribuent le plus à la définition de la liaison entre les 2 variables X et Y.

4.5.5 Remarque

En général, on n'interprète pas les axes d'une AFC (en particulier parce qu'il n'y a pas de variable quantitative intervenant dans l'analyse). L'interprétation s'appuie surtout sur la position relative des différentes modalités repérées comme les plus importantes.

5 Exemple

L'exemple des données bancaires ainsi que les données d'expression génomique se prête mal à l'illustration d'une analyse des correspondances, aucun couple de variable qualitative ne conduit à des représentations intéressantes.

La table de contingence étudiée à titre d'exemple décrit la répartition des exploitations agricoles de la région Midi–Pyrénées dans les différents départements en fonction de leur taille. Elle croise la variable qualitative *département*, à 8 modalités, avec la variable *taille de l'exploitation*, quantitative découpée en 6 classes. Les données, ainsi que les résultats numériques obtenus avec la procédure `corresp` de SAS/STAT, sont fournis en annexe.

La figure 5 présente le premier plan factoriel utilisant les coordonnées obtenues par défaut, c'est–à–dire celles de la double ACP.

6 Compléments

6.1 Propriétés

- *Formule de reconstitution des données.* On appelle ainsi l'approximation d'ordre q (c'est–à–dire fournie par l'AFC en dimension q) de la table des fréquences initiales ($\frac{1}{n}\mathbf{T}$) :

$$f_{\ell h} \simeq f_{\ell+}f_{+h}\sum_{k=1}^{q}\sqrt{\lambda_k}u_\ell^k v_h^k.$$

- Les valeurs propres vérifient :

$$\sum_{k=1}^{d}\lambda_k = \Phi^2.$$

En effet, on vérifie facilement :

$$\mathrm{tr}\mathbf{AB} = \sum_{k=0}^{d}\lambda_k = 1 + \frac{\chi^2}{n} = 1 + \Phi^2;$$

d'où le résultat.

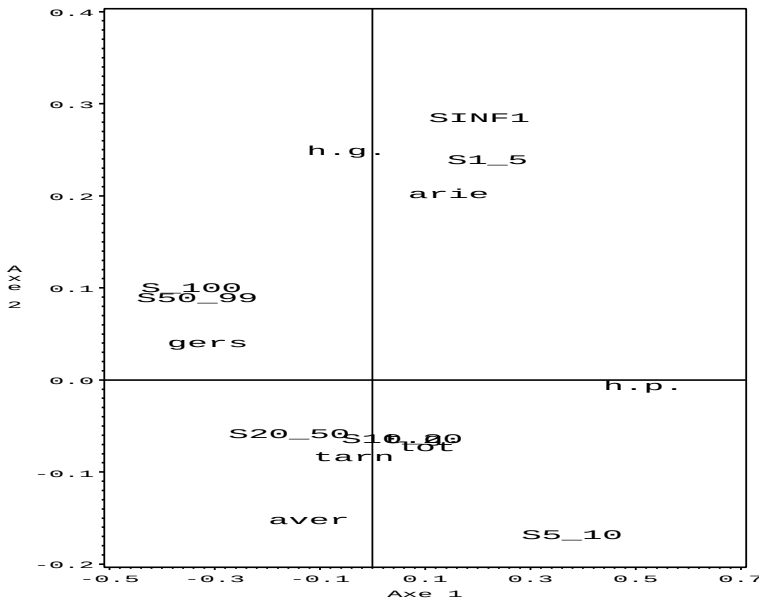

FIGURE 1 – *Répartition des exploitations agricoles par taille et par département. Premier plan de l'AFC.*

6.2 Invariance

- Les tables de contingence $\mathbf{T}$ et $\alpha\mathbf{T}, \alpha \in \mathbb{R}_+^*$, admettent la même AFC (évident).
- *Propriété d'équivalence distributionnelle* : si deux lignes de $\mathbf{T}$, ℓ et i, ont des effectifs proportionnels, alors les représentations de x_ℓ et x_i sont confondues (leurs profils sont identiques) et le regroupement de x_ℓ et x_i en une seule modalité (en additionnant les effectifs) laisse inchangées les représentations graphiques (même chose pour les colonnes de $\mathbf{T}$). Cette propriété est une conséquence de la métrique du χ^2.

6.3 Choix de la dimension

Le choix de la dimension pose les mêmes problèmes qu'en ACP De nombreuses techniques empiriques ont été proposées (essentiellement : part d'inertie expliquée, éboulis des valeurs propres). Il existe également une approche probabiliste qui peut donner des indications intéressantes. Nous la détaillons ci–dessous.

Posons

$$\widehat{n_{\ell h}^q} = n f_{\ell +} f_{+h} + n \sum_{k=1}^{q} \sqrt{\lambda_k} u_\ell^k v_h^k,$$

estimation d'ordre q de l'effectif conjoint de la cellule (ℓ, h). Alors, sous certaines conditions (échantillonnage, n grand, modèle multinomial ...), on peut montrer que

$$K_q = \sum_{\ell=1}^{r} \sum_{h=1}^{c} \frac{(n_{\ell h} - \widehat{n_{\ell h}^q})^2}{\widehat{n_{\ell h}^q}} \simeq n \sum_{k=q+1}^{d} \lambda_k$$

suit approximativement une loi de χ^2 à $(r-q-1)(c-q-1)$ degrés de liberté. On peut donc retenir pour valeur de q la plus petite dimension pour laquelle K_q est inférieure à la valeur limite de cette loi. Le choix $q = 0$ correspond à la situation où les variables sont proche de l'indépendance en probabilités ; les fréquences conjointes sont alors bien approchées par les produits des fréquences marginales.

Analyse factorielle multiple des correspondances (AFCM)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de données qualitatives complexes. Cette méthode est une généralisation de l'Analyse Factorielle des Correspondances, permettant de décrire les relations entre p ($p > 2$) variables qualitatives simultanément observées sur n individus. Elle est aussi souvent utilisée pour la construction de scores *comme préalable à une méthode de classification (kmeans) nécessitant des données quantitatives.*

Travaux pratiques *de complexité croissante par l'études de données élémentaires, puis épidémiologiques avec interactions.*

Retour au plan du cours.

1 Codages de variables qualitatives

1.1 Tableau disjonctif complet

Soit X une variable qualitative à c modalités. On appelle *variable indicatrice* de la k-ième modalité de x ($k = 1, \ldots, c$), la variable $X_{(k)}$ définie par

$$X_{(k)}(i) = \begin{cases} 1 & \text{si } X(i) = \mathcal{X}_k, \\ 0 & \text{sinon,} \end{cases}$$

où i est un individu quelconque et $\mathcal{X}_k$ est la k–ième modalité de X. On notera n_k l'effectif de $\mathcal{X}_k$.

On appelle *matrice des indicatrices* des modalités de X, et l'on notera $\mathbf{X}$, la matrice $n \times c$ de terme général :

$$x_i^k = X_{(k)}(i).$$

On vérifie :

$$\sum_{k=1}^{c} x_i^k = 1, \forall i \text{ et } \sum_{i=1}^{n} x_i^k = n_k.$$

Considérons maintenant p variables qualitatives $X^1, \ldots, X^p$. On note c_j le nombre de modalités de X^j, $c = \sum_{j=1}^{p} c_j$ et $\mathbf{X}_j$ la matrice des indicatrices de X^j.

On appelle alors *tableau disjonctif complet* la matrice $\mathbf{X}$, $n \times c$, obtenue par concaténation des matrices $\mathbf{X}_j$:

$$\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_p].$$

$\mathbf{X}$ vérifie :

$$\sum_{k=1}^{c} x_i^k = p, \forall i \text{ et } \sum_{i=1}^{n} \sum_{k=1}^{c} x_i^k = np.$$

D'autre part, la somme des éléments d'une colonne de $\mathbf{X}$ est égale à l'effectif marginal de la modalité de la variable X^j correspondant à cette colonne.

1.2 Tableau de Burt

On observe toujours p variables qualitatives sur un ensemble de n individus. On appelle *tableau de Burt* la matrice $\boldsymbol{\mathcal{B}}$, $c \times c$, définie par :

$$\boldsymbol{\mathcal{B}} = \mathbf{X}'\mathbf{X}.$$

On peut écrire $\boldsymbol{\mathcal{B}} = [\boldsymbol{\mathcal{B}}_{jl}]$ ($j = 1, \ldots, p$; $l = 1, \ldots, p$) ; chaque bloc $\boldsymbol{\mathcal{B}}_{jl}$, de dimension $c_j \times c_l$, est défini par :

$$\boldsymbol{\mathcal{B}}_{jl} = \mathbf{X}_j'\mathbf{X}_l.$$

Si $j \neq l$, $\boldsymbol{\mathcal{B}}_{jl}$ est la table de contingence obtenue par croisement des variables X^j en lignes et X^l en colonnes. Si $j = l$, le bloc diagonal $\boldsymbol{\mathcal{B}}_{jj}$ est lui–même une matrice diagonale vérifiant :

$$\boldsymbol{\mathcal{B}}_{jj} = \text{ diag } (n_1^j, \ldots, n_{c_j}^j).$$

La matrice $\boldsymbol{\mathcal{B}}$ est symétrique, d'effectifs marginaux $n_l^j p$ et d'effectif total np^2.

1.3 La démarche suivie dans ce chapitre

La généralisation de l'AFC à plusieurs variables qualitatives repose sur certaines propriétés observées dans le cas élémentaire où $p = 2$. On s'intéresse

tout d'abord aux résultats fournis par l'AFC usuelle réalisée sur le tableau disjonctif complet $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ relatif à 2 variables qualitatives X^1 et X^2 ; $\mathbf{X}$ est alors considéré comme une table de contingence (paragraphe 2). Ensuite, on suit la même démarche avec l'AFC réalisée sur le tableau de Burt $\mathcal{B}$ relatif à X^1 et X^2 (paragraphe 3). Enfin, en utilisant les propriétés obtenues dans les deux premiers cas, on généralise cette double approche à un nombre quelconque p de variables qualitatives ; on définit ainsi l'Analyse Factorielle des Correspondances Multiples (paragraphe 4).

2 AFC du tableau disjonctif complet relatif à 2 variables

2.1 Données

On note toujours X^1 et X^2 les 2 variables qualitatives considérées et r et c leurs nombres respectifs de modalités.

Les matrices intervenant dans l'AFC usuelle sont reprises ici avec les mêmes notations, mais surlignées. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2] ; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n ; \\
\overline{\mathbf{D}}_c &= \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\mathbf{\Delta} ; \\
\overline{\mathbf{A}} &= \frac{1}{2n}\overline{\mathbf{T}}'\overline{\mathbf{D}}_r^{-1} = \frac{1}{2}\mathbf{X}' ; \\
\overline{\mathbf{B}} &= \frac{1}{2n}\overline{\mathbf{T}}\,\overline{\mathbf{D}}_c^{-1} = \frac{1}{n}\mathbf{X}\mathbf{\Delta}^{-1}.
\end{aligned}$$

On considère ici l'AFC comme une double ACP : celle des profils–lignes $\overline{\mathbf{A}}$, puis celle des profils–colonnes $\overline{\mathbf{B}}$.

2.2 ACP des profils–lignes

Les profils–lignes, provenant de $\overline{\mathbf{T}}$, sont associés aux n individus observés. Leur ACP conduit ainsi à une représentation graphique des individus, inconnue en AFC classique.

PROPOSITION 1. — *L'ACP des profils–lignes issue de l'AFC réalisée sur le tableau disjonctif complet associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_c^{-1}$–symétrique et positive :*

$$\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix}.$$

Les $r + c$ valeurs propres de $\overline{\mathbf{A}}\,\overline{\mathbf{B}}$ s'écrivent

$$\mu_k = \frac{1 \pm \sqrt{\lambda_k}}{2},$$

où les λ_k sont les valeurs propres de la matrice $\mathbf{AB}$ (donc celles de l'AFC classique de X^1 et X^2).

Les vecteurs propres $\overline{\mathbf{D}}_c^{-1}$–orthonormés associés se mettent sous la forme

$$\overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix};$$

la matrice $\mathbf{U}$ (resp. $\mathbf{V}$) contient les vecteurs propres $\mathbf{D}_r^{-1}$–orthonormés (resp. $\mathbf{D}_c^{-1}$-orthonormés) de la matrice $\mathbf{BA}$ (resp. $\mathbf{AB}$) ; autrement dit, les matrices $\mathbf{U}$ et $\mathbf{V}$ sont les matrices de vecteurs propres obtenues en faisant l'AFC classique de la table de contingence croisant X^1 et X^2.

La matrice des composantes principales s'écrit

$$\overline{\mathbf{C}}_r = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r + \mathbf{X}_2\mathbf{C}_c\right]\mathbf{\Lambda}^{-1/2},$$

où $\mathbf{C}_r$ et $\mathbf{C}_c$ sont encore les matrices de composantes principales de l'AFC classique.

Dans la pratique, on ne considère que les $d = \inf(r-1, c-1)$ plus grandes valeurs propres différentes de 1, ainsi que les vecteurs propres associés. Les valeurs propres sont rangées dans la matrice

$$\mathbf{M} = \text{ diag }(\mu_1, \ldots, \mu_d) = \frac{1}{2}\left[\mathbf{I}_d + \mathbf{\Lambda}^{1/2}\right].$$

Les autres valeurs propres non nulles sont dues à l'artifice de construction de la matrice à diagonaliser ; elles n'ont donc pas de signification statistique.

On notera que la matrice $\overline{\mathbf{C}}_r$, $n \times d$, fournit les coordonnées permettant la représentation graphique des individus sur les axes factoriels.

2.3 ACP des profils–colonnes

Les profils–colonnes sont associés aux $r + c$ modalités des variables. Leur ACP conduit donc à une représentation graphique de ces modalités dont on verra qu'elle est très voisine de celle fournie par une AFC classique.

PROPOSITION 2. — *L'ACP des profils–colonnes issue de l'AFC réalisée sur le tableau disjonctif complet associé à 2 variables conduit à l'analyse spectrale de la matrice* $\overline{\mathbf{D}}_r^{-1}$*–symétrique et positive :*

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{2n}\left[\mathbf{X}_1\mathbf{D}_r^{-1}\mathbf{X}_1' + \mathbf{X}_2\mathbf{D}_c^{-1}\mathbf{X}_2'\right].$$

Les $r + c$ valeurs propres non nulles de $\overline{\mathbf{B}}\,\overline{\mathbf{A}}$ sont les μ_k.

Les vecteurs propres $\overline{\mathbf{D}}_r^{-1}$–orthonormés associés se mettent sous la forme :

$$\overline{\mathbf{U}} = \frac{1}{n}\overline{\mathbf{C}}_r\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_r \\ \mathbf{C}_c \end{bmatrix} \mathbf{\Lambda}^{-1/2}\mathbf{M}^{1/2}.$$

Ainsi, l'AFC du tableau disjonctif complet permet, grâce aux coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}}_c$, une représentation simultanée des modalités des 2 variables. Cette représentation est très voisine de celle obtenue par l'AFC classique, définie au chapitre précédent. Une simple homothétie sur chaque axe factoriel, de rapport $\sqrt{\frac{1+\sqrt{\lambda_k}}{2\lambda_k}}$, permet de passer de l'une à l'autre.

De plus, cette approche permet aussi de réaliser une représentation graphique des individus avec les coordonnées contenues dans les lignes de la matrice $\overline{\mathbf{C}}_r$. À un facteur près, chaque individu apparaît comme le barycentre des 2 modalités qu'il a présentées. Dans le cas où n est grand, le graphique des individus a néanmoins peu d'intérêt ; seule sa forme générale peut en avoir un.

Remarque. — Si, dans l'AFC classique, on choisit d'utiliser, pour la représentation simultanée des modalités de X^1 et de X^2, les lignes des matrices

$$\mathbf{C}_r^* = \mathbf{D}_r^{-1}\mathbf{U} = \mathbf{C}_r\mathbf{\Lambda}^{-1/2} \text{ et } \mathbf{C}_c^* = \mathbf{D}_c^{-1}\mathbf{V} = \mathbf{C}_c\mathbf{\Lambda}^{-1/2}$$

(voir chapitre précédent, sous–section 4.4), alors on obtient par AFC du tableau disjonctif complet la matrice

$$\overline{\mathbf{C}}_c^* = \overline{\mathbf{C}}_c\mathbf{M}^{-1/2} = \begin{bmatrix} \mathbf{C}_r^* \\ \mathbf{C}_c^* \end{bmatrix};$$

il y a invariance de la représentation des modalités lorsque l'on passe d'une méthode à l'autre. Pour les individus, on obtient

$$\overline{\mathbf{C}}_r^* = \frac{1}{2}\left[\mathbf{X}_1\mathbf{C}_r^* + \mathbf{X}_2\mathbf{C}_c^*\right]\mathbf{M}^{-1/2}$$

(le commentaire est alors le même qu'avec $\overline{\mathbf{C}}_r$).

3 AFC du tableau de Burt relatif à 2 variables

Dans cette section, on s'intéresse aux résultats fournis par l'AFC réalisée sur le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, $(r + c) \times (r + c)$, relatif aux 2 variables X^1 et X^2 ; $\mathcal{B}$ est encore considéré comme une table de contingence. La matrice $\mathcal{B}$ étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; il suffit donc de considérer une seule ACP

Les notations des matrices usuelles de l'AFC sont maintenant réutilisées surmontées d'un tilde. On obtient ainsi :

$$\begin{aligned}
\widetilde{\mathbf{T}} &= \mathcal{B} = \begin{bmatrix} n\mathbf{D}_r & \mathbf{T} \\ \mathbf{T}' & n\mathbf{D}_c \end{bmatrix}; \\
\widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{2}\begin{bmatrix} \mathbf{D}_r & 0 \\ 0 & \mathbf{D}_c \end{bmatrix} = \frac{1}{2}\mathbf{\Delta} = \overline{\mathbf{D}}_c\,; \\
\widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{2}\begin{bmatrix} \mathbf{I}_r & \mathbf{B} \\ \mathbf{A} & \mathbf{I}_c \end{bmatrix} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.
\end{aligned}$$

On considère encore l'AFC comme l'ACP des profils–lignes $\widetilde{\mathbf{A}}$ (ou des profils–colonnes $\widetilde{\mathbf{B}}$).

PROPOSITION 3. — *L'ACP des profils–lignes (ou des profils–colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à 2 variables qualitatives conduit à l'analyse spectrale de la matrice* $\widetilde{\mathbf{D}_c}^{-1}$ *–symétrique et positive :*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2.$$

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$ *–orthonormés*

$$\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}} = \frac{1}{2}\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}.$$

Les valeurs propres associées vérifient : $\nu_k = \mu_k^2$.

La matrice des composantes principales s'écrit :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \begin{bmatrix}\mathbf{C}_r\\ \mathbf{C}_c\end{bmatrix}\mathbf{\Lambda}^{-1/2}\mathbf{M}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant une représentation simultanée des modalités des deux variables. À une homothétie près, cette représentation est identique à celle de l'AFC classique, réalisée sur la table de contingence $\mathbf{T}$ (mais le rapport d'homothétie, sur chaque axe, n'est plus le même qu'avec $\overline{\mathbf{C}}_c$).

Remarque. —

- En reprenant les notations de la remarque 2.3, on obtient ici :

$$\widetilde{\mathbf{C}_r^*}\ (=\widetilde{\mathbf{C}_c^*}) = \widetilde{\mathbf{C}_r}\mathbf{M}^{-1} = \overline{\mathbf{C}}_c^* = \begin{bmatrix}\mathbf{C}_r^*\\ \mathbf{C}_c^*\end{bmatrix}.$$

  Ainsi, si l'on utilise ce mode de représentation graphique, les trois approches de l'AFC que nous avons présentées conduisent à la même représentation simultanée des modalités des 2 variables : il y a donc invariance de cette représentation.

- Dans les deux cas d'AFC considérés dans ce chapitre (sur tableau disjonctif complet et sur tableau de Burt) on trouve, par construction, des valeurs propres non nulles sans signification statistique. En conséquence, les critères de qualité s'exprimant comme une "part d'inertie expliquée" n'ont plus de signification.
- L'AFC sur tableau de Burt ne prend en compte que l'information contenue dans $\mathcal{B}$ qui ne considère que les croisements de variables prises deux à deux. En conséquence, les interactions de niveau plus élevé sont ignorées par cette approche, à moins de procéder à des recodages de variables comme l'explique l'exemple présenté dans la section 5.

4 Analyse Factorielle des Correspondances Multiples

4.1 Définition

On considère maintenant p variables qualitatives ($p \geq 3$) notées $\{X^j\ ; j = 1,\ldots,p\}$, possédant respectivement c_j modalités, avec $c = \sum_{j=1}^{p} c_j$. On suppose que ces variables sont observées sur les mêmes n individus, chacun affecté du poids $1/n$.

Soit $\mathbf{X} = [\mathbf{X}_1|\cdots|\mathbf{X}_p]$ le tableau disjonctif complet des observations ($\mathbf{X}$ est $n \times c$) et $\boldsymbol{\mathcal{B}} = \mathbf{X}'\mathbf{X}$ le tableau de Burt correspondant ($\boldsymbol{\mathcal{B}}$ est carré d'ordre c, symétrique).

DÉFINITION 4. — *On appelle Analyse Factorielle des Correspondances Multiples (AFCM) des variables* $(X^1,\ldots,X^p)$ *relativement à l'échantillon considéré, l'AFC réalisée soit sur la matrice* $\mathbf{X}$ *soit sur la matrice* $\boldsymbol{\mathcal{B}}$.

On note n_k^j $(1 \leq j \leq p, 1 \leq k \leq c_j)$ l'effectif de la k–ième modalité de X^j, $\mathbf{D}_j = \frac{1}{n}\,\text{diag}\,(n_1^j,\ldots,n_{c_j}^j)$ et $\mathbf{\Delta} = \text{ diag }(\mathbf{D}_1\ldots\mathbf{D}_p)$ ($\mathbf{\Delta}$ est carrée d'ordre c et diagonale).

4.2 AFC du tableau disjonctif complet

Comme dans le cas $p = 2$, on reprend les notations de l'AFC classique en les surlignant. On obtient ainsi :

$$\begin{aligned}
\overline{\mathbf{T}} &= \mathbf{X}\,; \\
\overline{\mathbf{D}}_r &= \frac{1}{n}\mathbf{I}_n\,; \\
\overline{\mathbf{D}}_c &= \frac{1}{p}\boldsymbol{\Delta}\,; \\
\overline{\mathbf{A}} &= \frac{1}{p}\mathbf{X}'\,; \\
\overline{\mathbf{B}} &= \frac{1}{n}\mathbf{X}\boldsymbol{\Delta}^{-1}.
\end{aligned}$$

ACP des profils–lignes

PROPOSITION 5. — *L'ACP des profils–lignes issue de l'AFC réalisée sur le tableau disjonctif complet de p variables qualitatives conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_c^{-1}$–symétrique et positive :*

$$\overline{\mathbf{A}}\,\overline{\mathbf{B}} = \frac{1}{np}\boldsymbol{\mathcal{B}}\boldsymbol{\Delta}^{-1}.$$

Il y a m ($m \leq c-p$) valeurs propres notées μ_k, $(0 < \mu_k < 1)$ rangées dans la matrice diagonale $\mathbf{M}$.

La matrice des vecteurs propres $\overline{\mathbf{D}}_c^{-1}$–orthonormés associés se décompose en blocs de la façon suivante :

$$\overline{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_p \end{bmatrix};$$

chaque bloc $\mathbf{V}_j$ est de dimension $c_j \times m$.

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_r = \sum_{j=1}^{p} \mathbf{X}_j \mathbf{D}_j^{-1} \mathbf{V}_j.$$

Comme dans le cas $p = 2$, la matrice des composantes principales permet de réaliser une représentation graphique des individus dans laquelle chacun apparaît, à un facteur près, comme le barycentre des p modalités qu'il a présentées.

Remarque. — La généralisation au cas $p > 2$ restreint les propriétés. Ainsi, les vecteurs des blocs $\mathbf{V}_j$ ne sont pas les vecteurs propres $\mathbf{D}_j^{-1}$–orthonormés d'une matrice connue.

ACP des profils–colonnes

PROPOSITION 6. — *L'ACP des profils–colonnes issue de l'AFC réalisée sur le tableau disjonctif complet de p variables conduit à l'analyse spectrale de la matrice $\overline{\mathbf{D}}_r^{-1}$–symétrique et positive :*

$$\overline{\mathbf{B}}\,\overline{\mathbf{A}} = \frac{1}{np}\mathbf{X}\boldsymbol{\Delta}^{-1}\mathbf{X}' = \frac{1}{np}\sum_{j=1}^{p}\mathbf{X}_j\mathbf{D}_j^{-1}\mathbf{X}_j'.$$

La matrice des vecteurs propres $\overline{\mathbf{D}}_r^{-1}$–orthonormés vérifie :

$$\overline{\mathbf{U}} = \overline{\mathbf{B}}\overline{\mathbf{V}}\mathbf{M}^{-1/2}.$$

La matrice des composantes principales s'écrit :

$$\overline{\mathbf{C}}_c = p\boldsymbol{\Delta}^{-1}\overline{\mathbf{V}}\mathbf{M}^{1/2}\,;$$

elle se décompose en blocs sous la forme :

$$\overline{\mathbf{C}}_c = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_p \end{bmatrix}.$$

Chaque bloc $\mathbf{C}_j$, de dimension $c_j \times m$, fournit en lignes les coordonnées des modalités de la variable X^j permettant la représentation graphique simultanée.

4.3 AFC du tableau de Burt

Le tableau de Burt $\mathcal{B} = \mathbf{X}'\mathbf{X}$, carré d'ordre c, étant symétrique, les profils–lignes et les profils–colonnes sont identiques ; on ne considère donc ici qu'une seule ACP

En utilisant encore le tilde dans ce cas, les matrices usuelles de l'AFC deviennent :

$$\begin{aligned}\widetilde{\mathbf{T}} &= \mathcal{B}\,;\\ \widetilde{\mathbf{D}_r} &= \widetilde{\mathbf{D}_c} = \frac{1}{p}\mathbf{\Delta} = \overline{\mathbf{D}}_c\,;\\ \widetilde{\mathbf{A}} &= \widetilde{\mathbf{B}} = \frac{1}{np}\mathcal{B}\mathbf{\Delta}^{-1} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}.\end{aligned}$$

PROPOSITION 7. — *L'ACP des profils–lignes (ou des profils–colonnes) issue de l'AFC réalisée sur le tableau de Burt associé à p variables qualitatives conduit à l'analyse spectrale de la matrice $\widetilde{\mathbf{D}_c}^{-1}$–symétrique et positive :*

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} = \left[\overline{\mathbf{A}}\,\overline{\mathbf{B}}\right]^2.$$

Elle admet pour matrice de vecteurs propres $\widetilde{\mathbf{D}_c}^{-1}$–orthonormés $\widetilde{\mathbf{U}} = \widetilde{\mathbf{V}} = \overline{\mathbf{V}}$.

Les valeurs propres associées vérifient $\nu_k = \mu_k^2$.

La matrice des composantes principales s'écrit :

$$\widetilde{\mathbf{C}_r} = \widetilde{\mathbf{C}_c} = \overline{\mathbf{C}}_c\mathbf{M}^{1/2} = \begin{bmatrix}\mathbf{C}_1\\ \vdots\\ \mathbf{C}_p\end{bmatrix}\mathbf{M}^{1/2}.$$

La matrice $\widetilde{\mathbf{C}_r}$ fournit les coordonnées permettant la représentation simultanée des modalités de toutes les variables (on ne peut pas faire de représentation des individus si l'on fait l'AFC du tableau de Burt).

4.4 Variables illustratives

Soit X^0 une variable qualitative, à c_0 modalités, observée sur les mêmes n individus que les X^j et n'étant pas intervenue dans l'AFCM Soit $\mathbf{T}_{0j}$ la table de contingence $c_0 \times c_j$ croisant les variables X^0 en lignes et X^j en colonnes. L'objectif est maintenant de représenter les modalités de cette variable supplémentaire X^0 dans le graphique de l'AFCM réalisée sur $X^1, \ldots, X^p$. Pour cela, on considère les matrices :

$$\begin{aligned}\mathbf{B}_0 &= [\mathbf{T}_{01}|\ldots|\mathbf{T}_{0p}]\ ;\\ \mathbf{D}_0 &= \frac{1}{n}\,\mathrm{diag}\,(n_1^0, \ldots, n_{c_0}^0)\ ;\\ \mathbf{A}_0 &= \frac{1}{np}\mathbf{D}_0^{-1}\mathbf{B}_0.\end{aligned}$$

Les coordonnées des modalités de la variable supplémentaires X^0 sur les axes factoriels sont alors fournies dans les lignes de la matrice

$$\mathbf{C}_0 = \mathbf{A}_0\widetilde{\mathbf{D}_c}^{-1}\widetilde{\mathbf{V}} = p\mathbf{A}_0\mathbf{\Delta}^{-1}\overline{\mathbf{V}}.$$

4.5 Interprétation

Les représentations graphiques sont interprétées de manière analogue à ce qui est fait dans l'AFC de deux variables, bien que la représentation simultanée des modalités de toutes les variables ne soit pas, en toute rigueur, réellement justifiée.

Les "principes" suivants sont donc appliqués :

- on interprète globalement les proximités et les oppositions entre les modalités des différentes variables, comme en AFC, en privilégiant les modalités suffisamment éloignées du centre du graphique (attention aux modalités à faible effectif !) ;
- les rapports de valeurs propres ne sont pas interprétables comme indicateurs de qualité globale ; on peut néanmoins regarder la décroissance des premières valeurs propres pour choisir la dimension ;
- les coefficients de qualité de chaque modalité ne peuvent pas être interprétés ; seules les contributions des modalités à l'inertie selon les axes sont interprétées, selon le même principe qu'en AFC

5 Exemple

L'AFCM ne donne pas de résultats très intéressants sur les données bancaires à l'exception du graphe présenté dans le chapitre d'introduction qui est relativement plus sophistiqué car il fait préalablement appel à une classification. Il en est de même pour les données d'expression qui sont quantitatives.

TABLE 1 – *Données sous la forme d'une table de contingence complète*

			Histologie			
			Inflam. minime		Grande inflam.	
Centre	*Âge*	*Survie*	Maligne	Bénigne	Maligne	Bénigne
Tokyo	< 50	non	9	7	4	3
		oui	26	68	25	9
	$50 - 69$	non	9	9	11	2
		oui	20	46	18	5
	> 70	non	2	3	1	0
		oui	1	6	5	1
Boston	< 50	non	6	7	6	0
		oui	11	24	4	0
	$50 - 69$	non	8	20	3	2
		oui	18	58	10	3
	> 70	non	9	18	3	0
		oui	15	26	1	1
Glamorgan	< 50	non	16	7	3	0
		oui	16	20	8	1
	$50 - 69$	non	14	12	3	0
		oui	27	39	10	4
	> 70	non	3	7	3	0
		oui	12	11	4	1

En revanche, l'AFCM est très indiquée et très utilisée dans des enquêtes de nature épidémiologique.

5.1 Les données

La littérature anglo-américaine présente souvent des données relatives à plusieurs variables qualitatives sous la forme d'une table de contingence *complète* (5). C'est le cas de l'exemple ci–dessous qui décrit les résultats partiels d'une enquête réalisée dans trois centres hospitaliers (Boston, Glamorgan, Tokyo) sur des patientes atteintes d'un cancer du sein. On se propose d'étudier la survie de ces patientes, trois ans après le diagnostic. En plus de cette information, quatre autres variables sont connues pour chacune des patientes :

- le centre de diagnostic,
- la tranche d'âge,
- le degré d'inflammation chronique,
- l'apparence relative (bénigne ou maligne).

L'objectif de cette étude est une analyse descriptive de cette table en cherchant à mettre en évidence les facteurs de décès.

5.2 Analyse brute

On se reportera à la figure 5. La variable survie, qui joue en quelques sortes le rôle de variable à expliquer, est très proche de l'axe 2 et semble liée à chacune des autres variables.

5.3 Analyse des interactions

Pour essayer de mettre en évidence d'éventuelles interactions entre variables, les données sont reconsidérées de la façon suivante :

- les variables `centre` et `âge` sont croisées, pour construire une variable `c_x_âge`, à 9 modalités ;
- les variables `inflam` et `appar` sont également croisées pour définir la variable `histol`, à 4 modalités.

Une nouvelle analyse est alors réalisée en considérant comme actives les deux variables nouvellement créées, ainsi que la variable `survie`, et comme illustratives les variables initiales : `centre`, `âge`, `inflam`, `appar`. Les résultats sont donnés dans la figure 2.

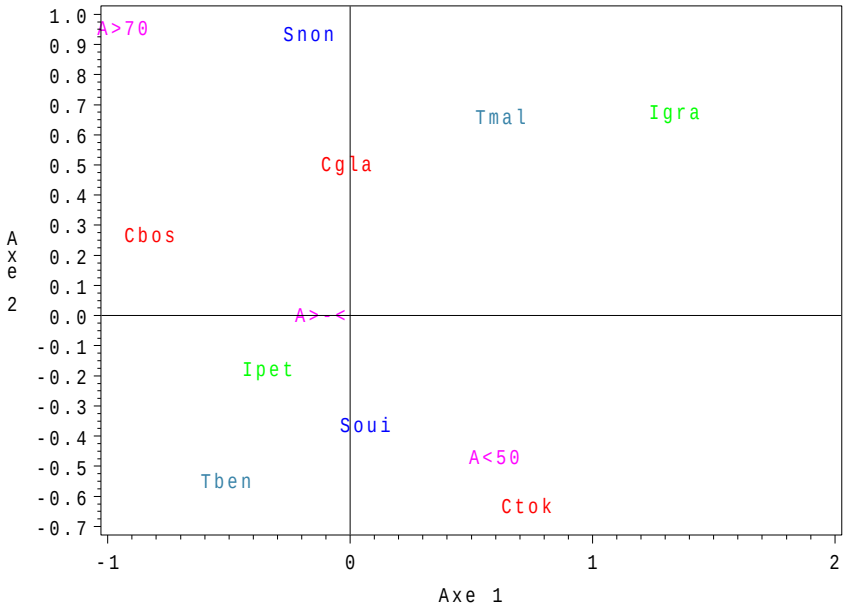

FIGURE 1 – *Cancer du sein : analyse des données brutes.*

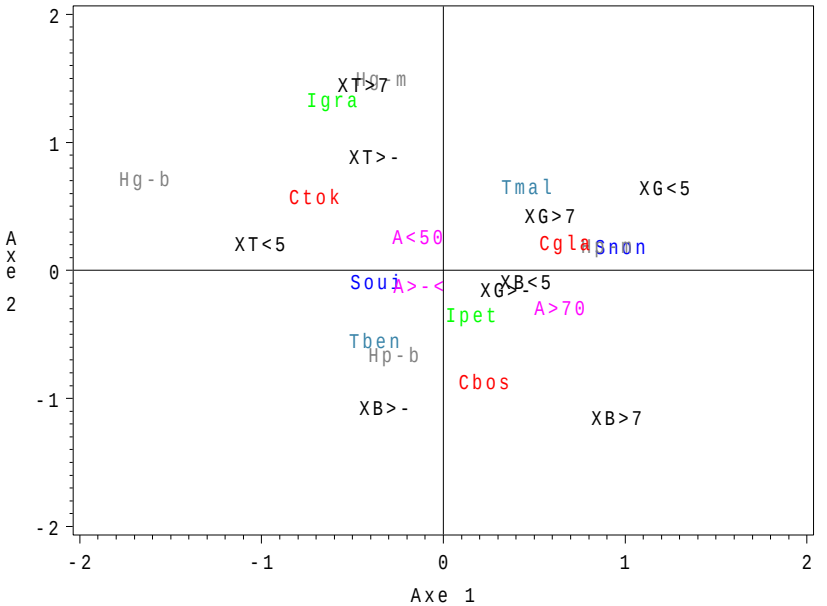

FIGURE 2 – *Cancer du sein : analyse des interactions.*

Analyse canonique des corrélations (ACC)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de deux ensembles de données quantitatives observées sur les mêmes individus. Représentations graphiques des individus, des variables et simultanée. Lien avec la régression multivariée et les tests associés.

Rerour au plan du cours.

1 Introduction

L'analyse canonique (A.C.) est une méthode de statistique descriptive multidimensionnelle qui présente des analogies à la fois avec l'analyse en composantes principales (A.C.P.), pour la construction et l'interprétation de graphiques, et avec la régression linéaire, pour la nature des données. L'objectif général de l'A.C. est d'explorer les relations pouvant exister entre deux groupes de variables quantitatives observées sur le même ensemble d'individus. L'étude des relations entre deux groupes de variables constitue la principale particularité de l'A.C. par rapport à l'A.C.P. De ce point de vue, l'A.C. est d'avantage proche de la régression linéaire multiple (explication d'une variable quantitative par un ensemble d'autres variables quantitatives), méthode dont elle constitue, d'ailleurs, une généralisation (on retrouve la régression lorsque l'un des deux groupes de l'A.C. ne comporte qu'une seule variable).

En fait, l'analyse canonique est, sur le plan théorique, la méthode centrale de la statistique descriptive multidimensionnelle, dans la mesure où elle généralise diverses autres méthodes. Outre la régression linéaire, l'A.C. redonne en effet l'analyse factorielle discriminante lorsque l'un des deux groupes de variables est remplacé par les indicatrices d'une variable qualitative. Elle redonne également l'analyse factorielle des correspondances lorsque chacun des deux groupes est remplacé par les indicatrices d'une variable qualitative. Signalons également qu'il existe certaines généralisations de l'A.C. à plus de deux groupes de variables quantitatives et qu'elles permettent de retrouver l'analyse des correspondances multiples (en remplaçant chaque groupe par les indicatrices d'une variable qualitative), ainsi que l'A.C.P. (en ne mettant qu'une seule variable quantitative dans chaque groupe). Nous ne nous intéresserons ici qu'à l'A.C. classique, entre deux groupes de variables quantitatives.

En dépit de sa place centrale au sein des méthodes de statistique multidimensionnelle, pendant longtemps, l'A.C. n'était pas (ou très peu) enseignée dans ces cursus, compte tenu du petit nombre d'applications auxquelles elle donnait lieu. Les choses ont changé, d'abord vers le milieu des années 1990, avec le développement de la régression P.L.S. (*partial least squares*), méthode assez voisine de l'A.C., ensuite, plus récemment, avec l'apparition des données de biopuces, dont certaines relèvent typiquement de l'A.C. quant à leur traitement.

Le logiciel statistique SAS dispose d'une procédure assez complète dédiée à l'A.C. : `CANCORR`. Divers développements de ce chapitre ont pour objectif de mieux saisir la signification de certaines sorties de cette procédure. Les commandes R permettant de mettre en œuvre l'A.C., telles qu'elles seront présentées dans les T.P., ont été quelque peu calquées sur le principe de la procédure `CANCORR`.

2 Approche élémentaire

2.1 Exemple : nutrition chez la souris

C'est encore l'exemple de la nutrition chez la souris qui sera utilisé pour illustrer l'A.C. Nous disposons donc des 40 souris sur lesquelles on s'intéresse maintenant à deux catégories de mesures (de variables) : les expressions des 120 gènes considérés et les proportions de 21 acides gras hépatiques. La question qui va être abordée ici est celle des relations entre ces deux ensembles de variables : certains acides gras sont-ils plus présents lorsque certains gènes sont surexprimés, ou le contraire... La réponse sera essentiellement fournie par les graphiques produits par l'A.C. et dans lesquels seront simultanément représentés gènes et acides gras : il s'agira donc de graphiques relatifs aux variables.

Notons tout de suite qu'il n'est pas très courant de représenter les individus en A.C. Toutefois, compte tenu des particularités de l'exemple considéré ici (petit nombre d'observations et structuration de ces observations selon les fac-

teurs “génotype” et “régime”), nous réaliserons ces graphiques et nous verrons quel est leur intérêt.

2.2 Notations

Dans toute la suite de ce chapitre, on notera n le nombre d'individus considérés (autrement dit, la taille de l'échantillon observé, ici 40), p le nombre de variables (quantitatives) du premier groupe (les gènes) et q le nombre de variables (également quantitatives) du second groupe (les acides gras). On désignera par $\mathbf{X}$ la matrice, de dimension $n \times p$, contenant les observations relatives au premier groupe de variables et par $\mathbf{Y}$ la matrice, de dimension $n \times q$, contenant celles relatives au second groupe. La j-ième colonne de $\mathbf{X}$ ($j = 1, \ldots, p$) contient donc les observations x_i^j de la j-ième variable du premier groupe (notée X^j, il s'agit de l'expression du j-ième gène retenu) sur les n individus considérés ($i = 1, \ldots, n$). De même, la k-ième colonne de $\mathbf{Y}$ ($k = 1, \ldots, q$) contient les observations y_i^k de la k-ième variable du second groupe (notée Y^k, il s'agit du pourcentage relatif au k-ième acide gras retenu).

En A.C., il est nécessaire d'avoir $p \leq n$, $q \leq n$, $\mathbf{X}$ de rang p et $\mathbf{Y}$ de rang q. Par conséquent, dans l'exemple considéré, il a été nécessaire de faire une sélection des gènes et de ne retenir que les plus importants (ceux dont le rôle prépondérant a préalablement été mis en évidence au moyen des techniques exploratoires). Bien que ce ne soit pas imposé par la théorie, nous avons également fait, pour être cohérents, une sélection des acides gras. Finalement, nous avons sélectionné 10 gènes et 11 acides gras hépatiques.

Les gènes sont les suivants :

```
PMDCI THIOL CYP3A11 CYP4A10 CYP4A14 Lpin Lpin1 GSTmu GSTp12 S14.
```

Les acides gras sont les suivants :

```
C16_0 C18_0 C18_1n_7 C18_1n_9 C18_2n_6 C18_3n_3
C20_4n_6 C20_5n_3 C22_5n_3 C22_5n_6 C22_6n_3.
```

Remarque. — On notera que la notation habituelle des acides gras est un peu différente de celle ci-dessus ; ainsi `C18_1n_7` correspond à `C18:1n-7` ; la notation adoptée est nécessaire pour la lecture par le logiciel SAS.

Enfin, sans perte de généralité, on suppose également $p \leq q$ (on désigne donc par premier groupe celui qui comporte le moins de variables). Finalement, nous avons ici : $n = 40$; $p = 10$; $q = 11$.

2.3 Principe général de la méthode

Chaque variable de chacun des deux groupes (les 10 gènes et les 11 acides gras) sont mesurées sur les n individus ($n = 40$). On peut donc associer à chacune un ensemble de 40 valeurs, autrement dit un vecteur de $\mathbb{R}^{40}$ (espace vectoriel que l'on a préalablement muni d'une base adéquate et d'une métrique appropriée). C'est dans cet espace ($\mathbb{R}^{40}$) que l'on peut définir la méthode : elle consiste à rechercher le couple de vecteurs, l'un lié aux gènes, l'autres aux acides, les plus corrélés possible. Ensuite, on recommence en cherchant un second couple de vecteurs non corrélés aux vecteurs du premier et le plus corrélés entre eux, et ainsi de suite. La démarche est donc similaire à celle utilisée en A.C.P. ou en analyse factorielle discriminante. La représentation graphique des variables se fait soit par rapport aux vecteurs liés aux gènes, soit par rapport à ceux liés aux acides (en général, les deux sont équivalentes, au moins pour ce qui est de leur interprétation). Ces vecteurs, obtenus dans chaque espace associé à chacun des deux groupes de variables, sont analogues aux facteurs de l'A.C.P. et sont ici appelés *variables canoniques*. Comme en A.C.P., on peut tracer le cercle des corrélations sur le graphique des variables, ce qui en facilite l'interprétation (dont le principe est le même que pour le graphique des variables en A.C.P.). Des considérations techniques permettent de faire également un graphique pour les individus.

Appelons d le nombre de couples de variables canoniques jugés intéressants, autrement dit la dimension retenue pour les représentations graphiques. On a nécessairement $1 \leq d \leq p$, et on choisit en général d entre 2 et 4. Nous noterons (V^s, W^s) ($s = 1, \ldots, d$) les couples de variables canoniques retenus ; on posera $\rho_s = \text{Cor}(V^s, W^s)$ et on appellera *corrélations canoniques* les coefficients ρ_s qui sont, par construction, décroissants.

3 Approche mathématique

Dans ce paragraphe, nous reprenons, plus en détail et avec plus de rigueur mathématique, les éléments présentés dans le paragraphe précédent. Le lecteur biologiste peu familiarisé avec ces notions de mathématiques pourra donc le parcourir très rapidement et se contenter d'aller y chercher quelques résultats, lorsque nécessaire.

3.1 Représentations vectorielles des données

Comme en A.C.P., on peut considérer plusieurs espaces vectoriels réels associés aux observations.

Tout d'abord, l'espace des variables ; c'est $F = \mathbb{R}^n$, muni de la base canonique et d'une certaine métrique, en général l'identité. À chaque variable X^j est associé un vecteur unique x^j de F dont les coordonnées sur la base canonique sont les x_i^j $(i = 1, \ldots, n)$. De même, à chaque variable Y^k est associé un vecteur unique y^k de F, de coordonnées les y_i^k. On peut ainsi définir dans F deux sous-espaces vectoriels : F_X, engendré par les vecteurs x^j $(j = 1, \ldots, p)$, en général de dimension p, et F_Y, engendré par les vecteurs y^k $(k = 1, \ldots, q)$, en général de dimension q.

Remarque. — Il est courant de munir l'espace vectoriel F de la métrique dite "des poids", définie, relativement à la base canonique, par la matrice diag $(p_1, \ldots, p_n)$, où les p_i $(i = 1, \ldots, n)$ sont des poids (positifs et de somme égale à 1) associés aux individus observés. Lorsque tous ces poids sont égaux, ils valent nécessairement $\frac{1}{n}$ et la matrice définissant la métrique des poids vaut $\frac{1}{n}\mathbf{I}_n$, où $\mathbf{I}_n$ est la matrice identité d'ordre n. Dans ce cas, il est équivalent d'utiliser la métrique identité, ce que nous ferons par la suite, dans la mesure où les individus seront systématiquement équipondérés.

On peut ensuite considérer deux espaces vectoriels pour les individus, $E_X = \mathbb{R}^p$ et $E_Y = \mathbb{R}^q$, eux aussi munis de leur base canonique et d'une certaine métrique. Dans E_X, chaque individu i est représenté par le vecteur x_i, de coordonnées x_i^j $(j = 1, \ldots, p)$ sur la base canonique. De même, dans E_Y, l'individu i est représenté par le vecteur y_i, de coordonnées les y_i^k.

En fait, c'est surtout l'espace F que nous considérerons par la suite, la définition de l'A.C. y étant plus naturelle.

3.2 Retour sur le principe de la méthode

Le principe général de l'A.C. est décrit ci-dessous, dans l'espace des variables F.

Dans un premier temps, on cherche un couple de variables (V^1, W^1), V^1 étant une combinaison linéaire des variables X^j (donc un élément de F_X), normée, et W^1 une combinaison linéaire des variables Y^k (donc un élément de F_Y), normée, telles que V^1 et W^1 soient le plus corrélées possible.

Ensuite, on cherche le couple normé (V^2, W^2), V^2 combinaison linéaire des X^j non corrélée à V^1 et W^2 combinaison linéaire des Y^k non corrélée à W^1, telles que V^2 et W^2 soient le plus corrélées possible. Et ainsi de suite...

Remarque. — Dans la mesure où l'A.C. consiste à maximiser des corrélations, quantités invariantes par translation et par homothétie de rapport positif sur les variables, on peut centrer et réduire les variables initiales X^j et Y^k sans modifier les résultats de l'analyse. Pour des raisons de commodité, on le fera systématiquement. Par conséquent, les matrices $\mathbf{X}$ et $\mathbf{Y}$ seront désormais supposées centrées et réduites (en colonnes).

L'A.C. produit ainsi une suite de p couples de variables (V^s, W^s), $s = 1, \ldots, p$. Les variables V^s constituent une base orthonormée de F_X (les V^s, combinaisons linéaires de variables centrées, sont centrées ; comme elles sont non corrélées, elles sont donc orthogonales pour la métrique identité). Les variables W^s constituent, de même, un système orthonormé de F_Y (ils n'en constituent une base que si $q = p$). Les couples (V^s, W^s), et plus particulièrement les premiers d'entre eux, rendent compte des liaisons linéaires entre les deux groupes de variables initiales. Les variables V^s et W^s sont appelées les *variables canoniques*. Leurs corrélations successives (décroissantes) sont appelées les *coefficients de corrélation canonique* (ou *corrélations canoniques*) et notées ρ_s $(1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_p \geq 0)$.

Remarque. — Toute variable canonique V^{s_0} est, par construction, non corrélée (donc orthogonale) avec les autres variables canoniques V^s, $s \neq s_0$. On peut également montrer que V^{s_0} est non corrélée avec W^s, si $s \neq s_0$ (la même propriété est bien sûr vraie pour toute variable W^{s_0} avec les variables V^s, $s \neq s_0$).

Remarque. — Si nécessaire, on peut compléter le système des variables W^s $(s = 1, \ldots, p)$ pour obtenir une base orthonormée de F_Y dans laquelle les dernières variables W^s $(s = p+1, \ldots, q)$ sont associées à des coefficients de corrélation canonique nuls ($\rho_s = 0$, pour $s = p+1, \ldots, q$).

3.3 Propriété

La propriété donnée ici permet, dans la pratique, de déterminer les variables canoniques V^s et W^s en utilisant un algorithme standard de recherche des vecteurs propres d'une matrice.

Dans l'espace vectoriel F muni de la métrique identité, notons $\mathbf{P_X}$ et $\mathbf{P_Y}$ les matrices des projecteurs orthogonaux sur les sous-espaces F_X et F_Y définis plus haut. Les formules usuelles de définition des projecteurs permettent d'écrire ($\mathbf{X}'$ désignant la matrice transposée de $\mathbf{X}$) :

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \; ; \; \mathbf{P_Y} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'.$$

On peut alors montrer la propriété ci-dessous.

PROPOSITION 1. — *Les vecteurs V^s sont les vecteurs propres normés de la matrice $\mathbf{P_X P_Y}$ respectivement associés aux valeurs propres λ_s rangées par ordre décroissant (on peut vérifier que ces valeurs propres sont comprises entre 1 et 0). De même, les vecteurs W^s sont les vecteurs propres normés de la matrice $\mathbf{P_Y P_X}$ respectivement associés aux mêmes valeurs propres λ_s. De plus, les coefficients de corrélation canonique ρ_s sont les racines carrées positives de ces valeurs propres : $\rho_s = \sqrt{\lambda_s}$, $s = 1, \ldots, p$ (le logiciel SAS fournit les corrélations canoniques ρ_s ainsi que leurs carrés λ_s).*

3.4 Retour sur les représentations graphiques

Comme en A.C.P., les représentations graphiques des résultats d'une A.C. se font en dimension réduite (souvent 2 ou 3). Nous noterons d cette dimension, avec : $1 \leq d \leq p$. Plusieurs représentations sont envisageables, à la fois pour les variables et pour les individus.

Représentation des variables dans le sous-espace F_X

Désignons par v^s et w^s les vecteurs de F_X et F_Y respectivement associés aux variables canoniques V^s et W^s.

Dans F_X, on considère la base orthonormée $(v^1, \ldots, v^p)$ que l'on restreint à $(v^1, \ldots, v^d)$ pour les représentations graphiques.

On peut tout d'abord représenter chacune des variables initiales X^j au moyen de ses coordonnées sur les v^s. Ces coordonnées s'obtiennent en calculant les produits scalaires $< x^j, v^s >$, $j = 1, \ldots, p$, $s = 1, \ldots, d$. Les variables X^j étant centrées et réduites, les vecteurs x^j sont centrés et normés (et il en va de même pour les vecteurs v^s), de sorte que ces produits scalaires sont égaux aux corrélations entre variables initiales X^j et variables canonique V^s (au coefficient n près, puisqu'on a considéré la métrique identité).

Dans le même espace, on peut également représenter les variables de l'autre groupe, les Y^k, en projetant tout d'abord les vecteurs y^k dans F_X, au moyen de $\mathbf{P_X}$, puis en prenant le produit scalaire de ces projections avec les vecteurs v^s. On doit donc calculer pour cela les produits scalaires

$$< \mathbf{P_X}(y^k), v^s > = < y^k, \mathbf{P_X}(v^s) > = < y^k, v^s >,$$

encore égaux aux corrélations entre les variables initiales Y^k et les variables canoniques V^s.

Dans la mesure où le graphique ainsi obtenu est "bon" (sur ce point, voir plus loin), on peut l'utiliser pour interpréter les relations (proximités, oppositions, éloignements) entre les deux ensembles de variables. Par construction, ce graphique représente les corrélations entre les variables canoniques V^s et les variables initiales X^j et Y^k, corrélations à la base de son interprétation. On peut aussi conforter cette interprétation en utilisant les coefficients de corrélation linéaire entre variables X^j, entre variables Y^k, et entre variables X^j et Y^k. Tous ces coefficients sont en général fournis par les logiciels.

Représentation des variables dans le sous-espace F_Y

De façon symétrique, on restreint le système $(w^1, \ldots, w^p)$ de F_Y aux premières variables $(w^1, \ldots, w^d)$, par rapport auxquelles on représente aussi bien les variables initiales X^j que les Y^k, selon le même principe que celui décrit ci-dessus (les coordonnées sont les corrélations).

Là encore, dans la mesure où ce graphique est "bon", il permet d'interpréter les relations entre les deux ensembles de variables.

Les deux graphiques (dans F_X et dans F_Y) ayant la même qualité et conduisant aux mêmes interprétations, un seul suffit pour interpréter les résultats d'une analyse.

Représentation des individus

Dans chacun des espaces relatifs aux individus (E_X et E_Y), il est encore possible de faire une représentation graphique de ces individus en dimension d, ces deux représentations graphiques étant comparables (d'autant plus comparables que les corrélations canoniques sont élevées).

En fait, on peut vérifier que les coordonnées des individus sur les axes canoniques pour ces deux représentations sont respectivement données par les

lignes des matrices $\mathbf{V_d}$ (dans E_X) et $\mathbf{W_d}$ (dans E_Y), $\mathbf{V_d}$ et $\mathbf{W_d}$ désignant les matrices $n \times d$ dont les colonnes contiennent les coordonnées des d premières variables canoniques sur la base canonique de F.

Choix de la dimension

Comme dans toute méthode factorielle, différents éléments doivent être pris en compte pour le choix de la dimension d dans laquelle on réalise les graphiques (et dans laquelle on interprète les résultats).

- Tout d'abord, il est clair que d doit être choisi petit, l'objectif général de la méthode étant d'obtenir des résultats pertinents dans une dimension réduite ; ainsi, le plus souvent, on choisit d égal à 2, 3 ou 4.
- Plus l'indice de dimension s augmente, plus la corrélation canonique ρ_s diminue ; or, on ne s'intéresse pas aux corrélations canoniques faibles, puisqu'on cherche à expliciter les relations entre les deux groupes de variables ; par conséquent, les dimensions correspondant à des ρ_s faibles peuvent être négligées.
- Le pourcentage que chaque valeur propre λ_s représente par rapport à la somme de toutes les valeurs propres, c'est-à-dire par rapport à la trace de la matrice diagonalisée, facilitent également le choix de d (voir la remarque 5).

4 Compléments : analyse canonique et régression multivariée

L'objectif principal de ce paragraphe est de donner une idée, à l'utilisateur du logiciel SAS, du principe des tests figurant dans la procédure `CANCORR`, celle qui permet de réaliser l'analyse canonique. Accessoirement, ce paragraphe introduit la régression multivariée et fait le lien entre cette technique et l'analyse canonique.

On notera que les tests présentés ici sont des tests statistiques classiques dans le contexte de l'analyse multivariée, que ce soit l'analyse canonique, la régression multivariée, l'analyse de variance multivariée (la MANOVA), ou même l'analyse discriminante. Ils apparaissent ainsi dans toutes les procédures du logiciel SAS permettant de mettre en œuvre ces méthodes.

Le lecteur peu familiarisé avec les méthodes multivariées pourra néanmoins sauter ce paragraphe.

4.1 Introduction

Ouvrages et logiciels anglo-saxons de statistique présentent souvent l'analyse canonique parallèlement à la régression linéaire multivariée (régression d'un ensemble de variables Y^k, à expliquer, sur un autre ensemble de variables X^j, explicatives). Cette approche est, en fait, assez naturelle, dans la mesure où les données sont de même nature dans les deux méthodes et où l'on cherche, dans l'une comme dans l'autre, des relations linéaires entre variables.

Il convient toutefois de noter les deux différences fondamentales entre les deux méthodes : contrairement à ce qu'il se passe en A.C., les deux ensembles de variables X^j et Y^k ne sont pas symétriques en régression, puisqu'il s'agit d'expliquer les variables Y^k au moyen des variables X^j ; d'autre part, toujours en régression, on suppose la normalité des variables réponses Y^k, alors qu'aucune hypothèse de cette nature n'est nécessaire en A.C. L'avantage de cette hypothèse (lorsqu'elle est "raisonnable") est de permettre de réaliser des tests dans le modèle de régression.

4.2 Le modèle de régression multivariée

Le modèle de régression multivariée des variables Y^k sur les variables X^j s'écrit :

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} ;$$

les matrices $\mathbf{Y}$, $n \times q$, et $\mathbf{X}$, $n \times p$, sont celles introduites en A.C. ; $\mathbf{B}$ est la matrice $p \times q$ des paramètres inconnus, à estimer (les coefficients de régression) ; $\mathbf{U}$ est la matrice $n \times q$ des erreurs du modèle. Chaque ligne U_i de $\mathbf{U}$ est un vecteur aléatoire de $\mathbb{R}^q$ supposé $\mathcal{N}_q(0, \Sigma)$, les U_i étant indépendants (Σ est une matrice inconnue, à estimer, supposée constante en i).

L'estimation maximum de vraisemblance de $\mathbf{B}$ conduit à la solution :

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

On appelle alors *valeurs prédites* (de $\mathbf{Y}$ par le modèle) les quantités :

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{P_X}\mathbf{Y} ;$$

d'autre part, on appelle *résidus* les quantités :

$$\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y}$$

(dans l'écriture ci-dessus, $\mathbf{P}_{\mathbf{X}}^{\perp}$ désigne, dans $\mathbb{R}^n$, le projecteur orthogonal sur le sous-espace supplémentaire orthogonal à F_X dans $\mathbb{R}^n$; on sait que ce projecteur s'écrit : $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$).

4.3 Matrices intervenant dans les tests

Dans le cadre du modèle gaussien, on peut tester la significativité du modèle en généralisant le test de Fisher, bien connu dans le cas unidimensionnel. Au numérateur de la statistique de Fisher figure la norme carrée du vecteur $\hat{y} - \overline{y}$, ici remplacée par $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ (cette matrice est centrée). Au dénominateur figure la norme carrée des résidus, ici remplacée par $\hat{\mathbf{U}}'\hat{\mathbf{U}}$ (on néglige, pour l'instant, les degrés de liberté de ces quantités). La statistique de Fisher est donc remplacée par le produit matriciel $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}$. Comme on a $\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$, il vient : $\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{H}$ (la notation $\mathbf{H}$ est standard, car cette quantité est liée à l'hypothèse nulle testée). D'autre part, $\hat{\mathbf{U}} = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y}$ entraîne : $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y} = \mathbf{E}$ (il s'agit encore d'une notation standard, cette matrice représentant les erreurs du modèle). Les tests multidimensionnels de significativité du modèle sont ainsi basés sur l'étude des valeurs propres soit du produit matriciel

$$\mathbf{H}\mathbf{E}^{-1} = (\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y})(\mathbf{Y}'\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y})^{-1},$$

soit encore du produit $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$, les valeurs propres de ces deux matrices se déduisant les unes des autres. Développons le second produit matriciel :

$$\mathbf{H} + \mathbf{E} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} + \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y} = \mathbf{Y}'\mathbf{Y};$$

d'où :

$$\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1},$$

matrice ayant les mêmes valeurs propres que

$$\mathbf{P}_{\mathbf{X}}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}' = \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}},$$

c'est-à-dire les λ_s ($s = 1, \ldots, p$), carrés des corrélations canoniques.

Remarque. — On peut vérifier (le résultat est classique) que les valeurs propres de la matrice $\mathbf{H}\mathbf{E}^{-1}$ valent $\dfrac{\lambda_s}{1-\lambda_s}$. Ces valeurs propres sont fournies par le logiciel SAS, ainsi que les pourcentages (et les pourcentages cumulés) qu'elles représentent par rapport à leur somme, trace de la matrice $\mathbf{H}\mathbf{E}^{-1}$.

En interprétant ces pourcentages comme la part d'inertie globale du nuage des individus restituée par les différents axes canoniques (ce qu'elles sont, par exemple, en analyse factorielle discriminante), ces quantités facilitent le choix de la dimension d retenue pour les graphiques et les interprétations.

4.4 Tests

Il existe plusieurs tests de significativité du modèle de régression multivariée, en général équivalents (au moins au niveau des décisions qu'ils entraînent). Ces tests sont les généralisations classiques du test de Fisher au cas multivarié (on les retrouve, par exemple, en analyse de variance multivariée) et sont des tests asymptotiques. Le logiciel SAS fournit les trois premiers ci-dessous, mais pas le quatrième. Il fournit également le test de Roy, basé sur la plus grande valeurs propre de la matrice $\mathbf{H}\mathbf{E}^{-1}$, soit $\dfrac{\lambda_1}{1-\lambda_1}$, mais ce test est à déconseiller.

- Le test de Wilks, adaptation du test du rapport des vraisemblances, est basé sur la statistique

$$\Lambda = \prod_{s=1}^{p}(1-\lambda_s) = \prod_{s=1}^{p}(1-\rho_s^2).$$

- Le test de la trace de Pillai est basé sur la statistique

$$Z = \text{trace}\ \mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \sum_{s=1}^{p}\lambda_s.$$

- Le test de la trace de Lawley-Hotelling est basé sur la statistique

$$T^2 = \text{trace}\ \mathbf{H}\mathbf{E}^{-1} = \sum_{s=1}^{p}\frac{\lambda_s}{1-\lambda_s}.$$

- Le test du khi-deux est basé sur la statistique

$$K = -[(n-1) - \frac{1}{2}(p+q+1)]\ln\prod_{s=1}^{p}(1-\lambda_s).$$

Le test du khi-deux présente l'avantage d'être directement utilisable, puisqu'on compare la statistique K à une loi de khi-deux à pq degrés de libertés (il s'agit d'un test approché).

Dans les trois autres tests ci-dessus, on doit transformer la statistique (Λ, Z ou T^2) pour obtenir un test de Fisher approché, les transformations étant assez compliquées à expliciter (toutefois, SAS les réalise automatiquement).

Remarque. — Dans un article de 1951, Rao a montré que, dans la plupart des cas, l'approximation de Fisher du test de Wilks est la meilleure. C'est donc le test que nous conseillerons.

Si le modèle de régression est significatif (il en va alors de même pour l'analyse canonique), on peut tester la significativité d'une dimension et de l'ensemble des suivantes, en particulier pour guider le choix de la dimension en A.C. Ainsi, supposons que les corrélations canoniques soient significatives depuis la première jusqu'à la k-ième ($1 \leq k \leq p$). On peut alors tester l'hypothèse nulle

$$\{H_0 : \rho_{k+1} = \cdots = \rho_p = 0\} \ (\Longleftrightarrow \{H_0 : d = k\})$$

contre l'alternative

$$\{H_1 : \rho_{k+1} > 0\} \ (\Longleftrightarrow \{H_1 : d > k\}).$$

Pour cela, il faut adapter soit le test de Wilks, soit le test du khi-deux.

Pour le test de Wilks, il suffit de faire le produit des quantités $(1 - \lambda_s)$ de l'indice $k + 1$ à l'indice p et d'adapter la transformation en fonction des nouvelles dimensions. SAS le fait automatiquement. Pour le test du khi-deux, il faut considérer la statistique

$$K_k = -[(n - 1 - k) - \frac{1}{2}(p + q + 1) + \sum_{s=1}^{k} \frac{1}{\lambda_s}] \ln \prod_{s=k+1}^{p} (1 - \lambda_s)$$

et la comparer à une loi de khi-deux à $(p - k)(q - k)$ degrés de liberté.

Remarque. — Dans l'utilisation de ces tests, il convient de ne pas perdre de vue d'une part qu'il s'agit de tests asymptotiques (d'autant meilleurs que la taille de l'échantillon, n, est grande), d'autre part qu'ils ne sont valables que sous l'hypothèse de normalité des variables Y^k.

5 Exemple : nutrition chez la souris

5.1 Traitements préliminaires

Nous donnons ci-dessous les statistiques élémentaires relatives aux deux groupes de variables. Les corrélations entre gènes se trouvent en Annexe A, celles entre acides en Annexe B.

Variable	N	Mean	Std Dev	Minimum	Maximum
PMDCI	40	-0.7673	0.1861	-1.07	-0.44
THIOL	40	-0.4110	0.2125	-0.90	-0.03
CYP3A11	40	-0.5083	0.2556	-1.02	0.06
CYP4A10	40	-0.9798	0.2237	-1.33	-0.48
CYP4A14	40	-0.9930	0.2460	-1.29	-0.15
Lpin	40	-0.7533	0.1735	-1.13	-0.48
Lpin1	40	-0.7648	0.1638	-1.10	-0.49
GSTmu	40	-0.1190	0.1504	-0.44	0.23
GSTpi2	40	0.2298	0.1422	0	0.55
S14	40	-0.8068	0.2008	-1.05	-0.25

Variable	N	Mean	Std Dev	Minimum	Maximum
C16_0	40	23.03	3.57	14.65	29.72
C18_0	40	6.75	2.64	1.68	10.97
C18_1n_7	40	4.43	3.38	1.53	15.03
C18_1n_9	40	25.27	7.34	14.69	41.23
C18_2n_6	40	15.28	8.76	2.31	40.02
C18_3n_3	40	2.89	5.83	0	21.62
C20_4n_6	40	5.28	4.46	0.75	15.76
C20_5n_3	40	1.79	2.59	0	9.48
C22_5n_3	40	0.87	0.86	0	2.58
C22_5n_6	40	0.44	0.66	0	2.52
C22_6n_3	40	5.91	5.33	0.28	17.35

Remarque. — Les valeurs ci-dessus sont relatives aux variables brutes (aux données initiales). Comme indiqué dans la remarque 3, ces variables ont ensuite été centrées et réduites avant la réalisation de l'A.C.

5.2 Analyse canonique

Généralités

Les premiers résultats fournis par une A.C. sont les corrélations croisées entre les deux groupes de variables. Nous donnons ces corrélations dans l'an-

nexe C.

Ensuite sont données les corrélations canoniques reproduites ci-dessous.

	Canonical Correlation
1	0.96
2	0.93
3	0.91
4	0.86
5	0.79
6	0.72
7	0.61
8	0.41
9	0.25
10	0.04

On notera que “le plus petit” groupe ne comportant que 10 variables, on ne peut déterminer que 10 corrélations canoniques. L’objectif principal de l’A.C. étant d’étudier les relations entre variables des deux groupes, on peut noter ici qu’il existe effectivement des relations fortes entre ces deux groupes, puisque les premiers coefficients canoniques sont très élevés. Compte tenu des valeurs importantes des premiers coefficients, on peut raisonnablement se contenter de deux ou trois dimensions pour étudier les résultats fournis par la méthode et nous avons choisi ici seulement deux dimensions, compte tenu qu’il s’agit essentiellement d’une illustration.

Remarque. — Les valeurs propres de la matrice $\mathbf{HE}^{-1}$ et les pourcentages d’inertie restitués par les différentes dimensions sont les suivants :

	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
	Eigenvalue	Difference	Proportion	Cumulative
1	12.7583	6.1471	0.4167	0.4167
2	6.6111	1.7001	0.2159	0.6326
3	4.9111	2.1433	0.1604	0.7930
4	2.7678	1.1107	0.0904	0.8833
5	1.6571	0.5892	0.0541	0.9375
6	1.0679	0.4877	0.0349	0.9723
7	0.5802	0.3792	0.0189	0.9913
8	0.2010	0.1369	0.0066	0.9978
9	0.0641	0.0623	0.0021	0.9999
10	0.0018		0.0001	1.0000

Par ailleurs, les tests de Wilks, de significativité de chaque dimension, sont les suivants :

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.00003857	4.08	110	155.53	<.0001
2	0.00053068	3.31	90	145.91	<.0001
3	0.00403909	2.77	72	135.32	<.0001
4	0.02387531	2.21	56	123.78	0.0001
5	0.08995724	1.78	42	111.33	0.0090
6	0.23902627	1.41	30	98	0.1087
7	0.49427788	0.99	20	83.865	0.4795
8	0.78104952	0.56	12	69.081	0.8636
9	0.93806320	0.29	6	54	0.9380
10	0.99819295	0.03	2	28	0.9750

On voit que le choix optimal de la dimension serait probablement $d = 4$ (ne pas oublier que ces tests sont asymptotiques et que nous avons $n = 40$). Pour simplifier, nous ne présentons, par la suite, que les graphiques selon les deux premières dimensions.

Graphique des individus

Dans un premier temps, nous avons réalisé le graphique des individus (les 40 souris) relativement aux deux premiers axes canoniques de l’espace des gènes E_X (Fig. 1). En général, dans une A.C., ce graphique sert seulement à contrôler l’homogénéité de l’ensemble des individus (absence d’individus atypiques par exemple). Ici, dans la mesure où les individus proviennent d’un plan d’expériences à deux facteurs croisés (le génotype et le régime), il est intéressant de regarder si l’on retrouve la structure de ce plan. On notera que cela est très net en ce qui concerne le génotype et encore assez net pour ce qui est du régime (en fait, la sélection des gènes a été réalisée de telle sorte que ceux retenus soient le plus structurant possible pour ces deux facteurs ; le résultat, s’il est rassurant, n’a donc rien d’extraordinaire).

Signalons pour terminer qu’on a également réalisé le graphique des individus relativement aux deux premiers axes de l’autre espace (espace des acides gras, E_Y) et qu’il est très semblable à celui-ci.

Graphique des variables

Pour la représentation des variables, nous avons considéré le sous-espace F_X, engendré par les 10 gènes, et nous avons représenté à la fois les gènes et les acides gras relativement aux deux premières variables canoniques, V^1 et V^2 (Fig. 2). Comme indiqué en 3.4, les coordonnées des variables initiales

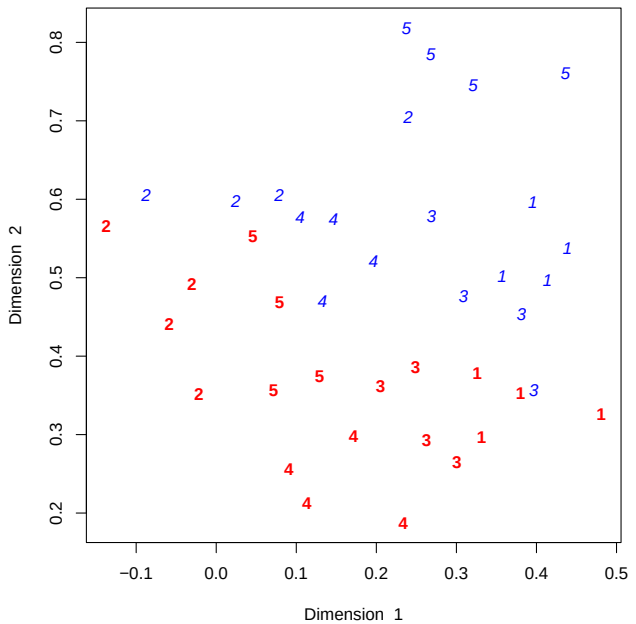

FIGURE 1 – *Souris : représentation des individus (souris) dans l'espace des gènes. Les WT sont en rouge-gras et les PPAR en bleu-italique ; les numéros correspondent aux régimes.*

sont fournies par leur corrélations avec les variables canoniques.

Certaines associations entre gènes et acides gras, en particulier celles correspondant à des points éloignés de l'origine, sont intéressantes à noter. Ainsi peut-on observer que la séparation des génotypes est principalement liée d'une part à l'accumulation préférentielle de l'acide gras `C18_2n_6` chez les souris `PPAR`, au détriment de `C16_0`, de `C18_0` et des acides gras longs polyinsaturés `C20_5n_3` et `C22_6n_3` (les *oméga 3*), d'autre part à la plus forte expression des gènes `THIOL`, `PMDCI`, `CYP3A11` et `GSTpi2` chez les souris `WT` par rapport aux souris `PPAR`. On peut également noter les proximités entre le `C16_0` et le gène `THIOL`, ainsi que les proximités entre `CYP3A11` et `GSTpi2` et les acides gras `C18_0` et `C22_6n_3`. Par ailleurs, l'opposition entre le régime `2-efad` et les régimes `1-dha` et `3-lin` est liée, sous régime `efad`, à l'accumulation d'acides gras monoinsaturés (`C18_1n_9` et `C18_1n_7`) chez les souris des deux génotypes (mais plus marquée chez les souris `PPAR`), accompagnée de la sur-expression du gène `S14` presque exclusivement chez les souris `WT`. Sous régime riche en *Oméga 3* (`1-dha` et `3-lin`), on observe une accumulation préférentielle des acides gras `C20_5n_3` (surtout pour le régime `lin`), `C22_6n_3` (surtout pour le régime `dha`) et `C18_0` accompagnée de régulations positives des gènes `GSTpi2`, `CYP3A11` et des `CYP4A` qui, cependant, se révèlent moins marquées, voire absentes, chez les souris `PPAR`. Enfin, remarquons que la position particulière du régime `5-tsol` chez les souris `PPAR` est liée à l'accumulation extrêmement marquée de `C18_2n_6` dans le foie de ces souris sous le régime `tsol` (sous ce régime, la proportion de `C18_2n_6` est presque deux fois plus importante chez les souris `PPAR` que chez les souris `WT`), soulignant ainsi le rôle primordial de PPARα dans la prise en charge de cet acide gras, que ce soit pour sa dégradation ou pour son utilisation pour la biosynthèse des acides gras longs polyinsaturés de la famille *Oméga 6*.

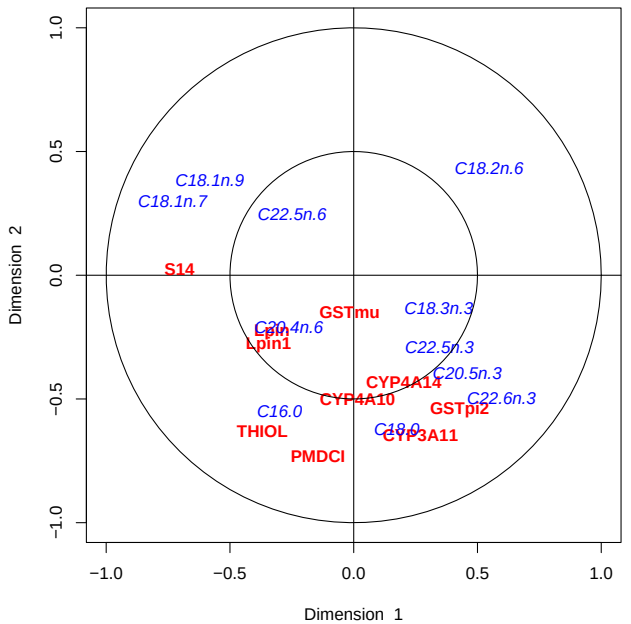

FIGURE 2 – *Souris : représentation des gènes (en rouge-gras) et des acides (en bleu-italique) dans le sous-espace des gènes.*

Positionnement multidimensionnel (MDS)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique d'une matrice de distances ou dissemblances entre individus. ACP d'un tableau de distances ou multidimensional scaling.

Travaux pratiques *avec étude de données élémentaires.*

Retour au plan du cours.

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 $(n \times n)$ $\mathcal{D}$. 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.

Exemple élémentaire : Les données sont constituées d'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 la décroissance des valeurs propres aide à ce choix.

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 expressions transcriptomiques et un exemple pluls qualitatif :

- 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.
- Considérons un tableau avec, en ligne, les individus d'un groupe et en colonne les pays de la C.E. La valeur 1 est mise dans une case lorsque l'individu de la ligne a passé au moins une nuit dans le pays concerné. Il est alors facile de construire une matrice de similarité avec un indice qui compte le nombre de 1 apparaissant dans les mêmes colonnes de tous les couples d'individus. L'objectif est ensuite d'obtenir une représentation graphique rapprochant les individus ayant visité les mêmes pays.

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)[1].

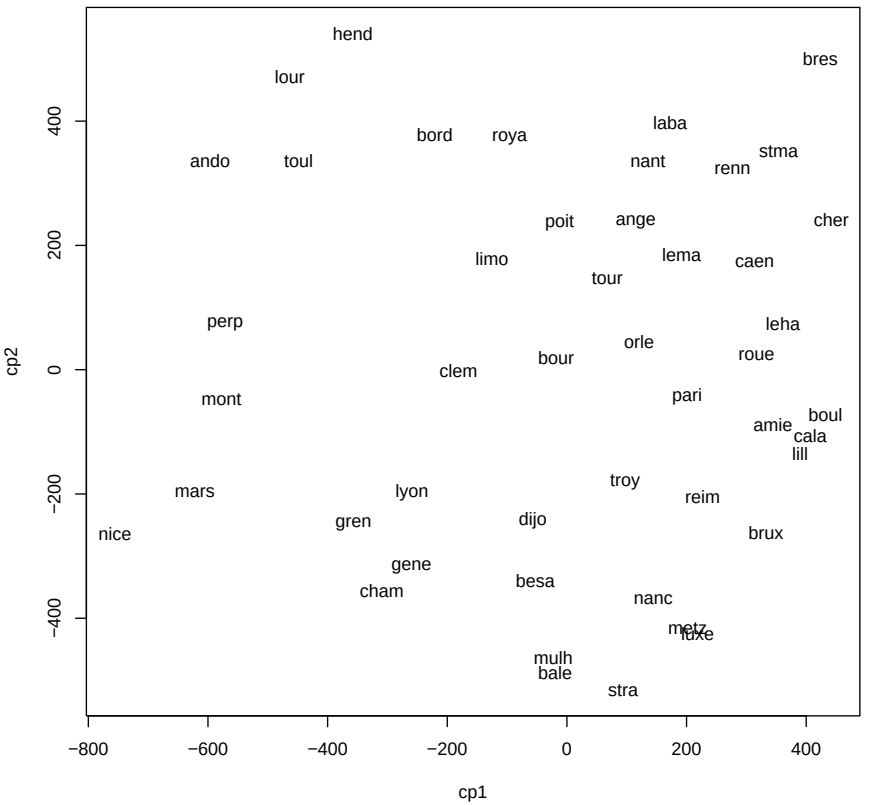

FIGURE 1 – *Villes : Positionnement de 47 villes à partir de la matrice de leurs distances kilométriques.*

2 Distance, similarités

Rappelons quelques propriétés et définitions élémentaires à propos de la notion de dissemblance ou similarité. Ces poins sont reprécisés dans la vignette sur la Classification non-supervisée.

2.1 Définitions

DÉFINITION 1. —

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

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

Pour que cette définition corresponde formellement à celle d'une "distance" il faudrait ajouter l'axiome d'inégalité triangulaire.

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 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 $\mathcal{D}$ 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 3. —

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

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

alors $\mathcal{D}$ 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 $\mathcal{C}$ est positive alors la matrice de distance $\mathcal{D}$ 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 :

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

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 :

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

PROPOSITION 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 $\mathcal{D}$ 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. — *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* $\mathbf{B} = \mathbf{U}\mathbf{\Delta}\mathbf{U}'$*. Ce sont les lignes de la matrice centrée* $\mathbf{X} = \mathbf{U}\mathbf{\Delta}^{1/2}$ *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 :

1. construction de la matrice $\mathbf{A}$ de terme général $-1/2d_j^{k\,2}$,
2. calcul de la matrice des produits scalaires par double centrage $\mathbf{B} = \mathbf{HAH}'$,
3. diagonalisation de $\mathbf{B} = \mathbf{U}\mathbf{\Delta}\mathbf{U}'$;
4. les coordonnées d'une configuration, appelées *coordonnées principales*, sont les lignes de la matrice $\mathbf{X} = \mathbf{U}\mathbf{\Delta}^{1/2}$.

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

PROPOSITION 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 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 Données génomiques

Une analyse en composantes principales 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. 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.

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.

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 2 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

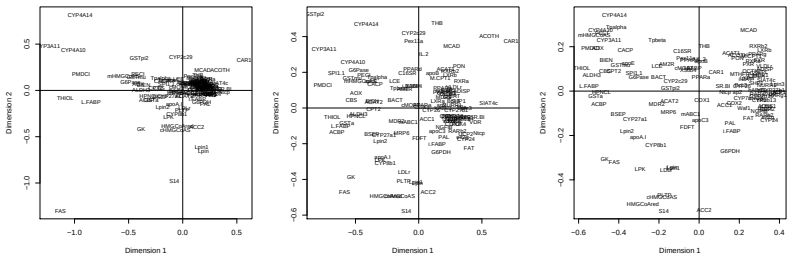

FIGURE 2 – *Souris : positionnement multidimensionnel des gènes sur les axes 1 et 2 selon 3 distances différentes : distance euclidienne (d_1 à gauche), corrélation (d_3 au centre), corrélation carrée (d_2 à droite).*

hépatique. On observera qu'aucun des trois graphiques de la figure 2, 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, il a été 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.

Références

[1] K.V. Mardia, J.T. Kent et J.M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

Classification non supervisée

Résumé

Méthodes de classification non supervisée (ou clustering*). Notions de distance, classification ascendante hiérarchique et choix de distances entre classes, construction du dendrogramme, chjoix du nombre de classes Classification par ré-allocation dynamique (*k*-means,* partitionning around medoïds*), méthode mixte pour les grands tableaux.*

Travaux pratiques avec SAS et R pour la recherche de classes et leurs représentations.

Retour au plan du cours.

1 Introduction

1.1 Les données

Comme dans le cas du thème 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.

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 ; le nombre de partitions de n éléments en k classes est le nombre de Stirling, le nombre total de partition est celui de Bell : $P_n = \frac{1}{e}\sum k = 1^{\infty}\frac{k^n}{k!}$.

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 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)[1] 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 ;

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

- 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 3.5 qui reprend les données élémentaires de la vignette sur le MDS). 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 ré-allocation dynamique

Dans ce cas, le nombre de classes, k, est fixé *a priori*. Ayant initialisé k centres de classes par tirage aléatoire (ou autre procédure), 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.

2 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.

2.1 Indice de ressemblance, ou similarité

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

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

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 $[0,1]$.

2.2 Indice de dissemblance, ou dissimilarité

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

$$\begin{array}{rcll} \forall (i,j) & \in & \Omega \times \Omega & \\ d(i,j) & = & d(j,i), : & \text{symétrie ;} \\ d(i,j) = 0 & \Leftrightarrow & i = j. & \end{array}$$

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.

2.3 Distance

Une distance sur Ω est, par définition, une dissimilarité 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.

2.4 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.

2.5 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, absence de caractéristiques), de nombreux indices de ressemblances ont été proposés dans la littérature. Ils sont basés sur les quantités suivantes définis pour deux individus i et j distincts :

- a_{ij} = nombre de caractères communs à i et j sur les p considérés,
- b_{ij} = nombre de caractères possédés par i mais pas par j,
- c_{ij} = nombre de caractères possédés par j mais pas par i,
- c_{ij} = nombre de caractères que ne possèdent ni i ni j.
- bien sûr, $a_{ij} + b_{ij} + b_{ij} + d_{ij} = p$.

Les indices de ressemblance les plus courants sont :

- Concordance : $\frac{a_{ij}+d_{ij}}{p}$,
- Jaccard : $\frac{a_{ij}}{a_{ij}+b_{ij}+b_{ij}}$,
- Dice : $\frac{2a_{ij}}{2a_{ij}+b_{ij}+b_{ij}}$.

Puis, il est facile de construire un indice de dissemblance.

Dans le cas plus général de p variables qualitatives, la distance la plus utilisée est celle, euclidienne, dite du χ^2 entre profils-lignes du tableau disjonctif complet (cf. chapitre 6 AFCM). La distance entre deux individus i et k est alors définie par :

$$d^2_{\chi^2} = \frac{n}{p} \sum_{j=1} p \sum_{\ell=1}^{m_j} \delta_{ik}^{j\ell} \frac{1}{n_\ell^j}.$$

où m_j est le nombre de modalités de la variable qualitative Y^j, n_ℓ^j est l'effectif de la ℓ-ième modalité de Y^j et $\delta_{ik}^{j\ell}$ vaut 1 si les individus i et k présentent une discordance pour la ℓ-ième modalité de la variables Y^j et 0 sinon. L'importance donnée à une discordance est d'autant plus importante que les modalités considérées sont rares. Le coefficient n/p peut être omis.

Mélange quantitatif, qualitatif

Différentes stratégies sont envisageables dépendant de l'importance relative des nombres de variables qualitatives et quantitatives.

Rendre tout qualitatif . Les variables quantitatives sont rendues qualitatives par découpage en classes. Les classes d'une même variable sont généralement recherchées d'effectifs sensiblement égaux : bornes des classes égales à des quantiles. La métrique à utiliser est alors celle du χ^2 décrite ci-dessus.

Rendre tout quantitatif à l'aide d'une AFCM. Une AFCM est calculée sur les seules variables qualitatives ou sur l'ensemble des variables après découpage en classes des variables quantitatives. L'AFCM calculée par AFC du tableau disjonctif complet produit des *scores* (cf. chapitre 6) qui sont les composantes principales de l'ACP des profils-lignes. Dans le cas d'une AFCM partielle des seules variables qualitatives, les variables quantitatives restantes doivent être nécessairement réduites. Ces scores sont ensuite utilisés comme coordonnées quantitatives des individus en vue d'une classification.

Métrique de Gower permet de mixer les types de variables mais celle-ci reste très peu utilisée.

2.6 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.

2.7 Accord entre partitions

Une partition de n individus définit une variable qualitative dont les catégories sont les classes de la partition. Une comparaison de deux partitions est obtenue an construisant la table de contingence croisant ces deux variables. Cependant, les numéros des classes étant arbitraires, l'appréciation de cet accord est difficile aussi un indice quantitatif a été proposé en considérant toutes les paires d'individus, selon qu'ils appartiennent à la même classe dans les deux partitions, qu'ils sont dans la même classe pour l'une mais pas pour l'autre, et enfin qu'ils sont séparés dans les deux partitions.

En notant n_{kl} le terme général de la table de contingence croisant les deux partitions, l'indice dit de Rand s'écrit :

$$R = \frac{\sum_k \sum_l n_{kl} - \sum_k n_{k+}^2 \sum_l n_{+l}^2}{+} n^2 n^2.$$

Cet indice prend ses valeurs entre 0 et 1, il est égal à 1 lorsque les deux partitions sont identiques. D'autres variantes ont été proposées.

3 Classification ascendante hiérarchique

3.1 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*.

3.2 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'accommodent 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 toutes les propriétés.

$$d(A,B) = \min_{i\in A, j\in B}(d_{ij}) \quad \text{(saut minimum, } \textit{single linkage}\text{)},$$

$$d(A,B) = \sup_{i\in A, j\in B}(d_{ij}) \quad \text{(saut maximum ou diamètre, } \textit{complete linkage}\text{)},$$

$$d(A,B) = \frac{1}{\text{card}(A)\text{card}(B)} \sum_{i\in A, j\in B} d_{ij} \quad \text{(saut moyen, } \textit{group average linkage}\text{)}.$$

Cas d'une distance euclidienne

Considérons que les données sont sous la forme d'une matrice $n \times p$ de variables quantitatives associée à une métrique euclidienne dans $\mathbb{R}^p$ ou directement sous la forme d'une matrice de distances euclidiennes $(n \times n)$ des individus 2 à 2. Dans le premier cas, il est facile de calculer les barycentres des classes et donc de considérer les distances suivantes entre deux groupes.

$$d(A,B) = d(g_A, g_B) \quad \text{(distance des barycentres, centroïd)},$$

$$d(A,B) = \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{(saut de Ward)}.$$

Dans le 2ème cas, le carré de la distance entre 2 barycentres se calcule à partir de la matrice des distances 2 à 2.

Remarque : Le saut de Ward joue un rôle particulier et est la stratégie la plus courante ; c'est même souvent 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. De plus, même si la distance entre individus n'est pas euclidienne, la même expression est utilisée pour faire du "saut de Ward" dans le cas non-euclidien.

3.3 Algorithme

ALGORITHME 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 :*
  1. *regrouper les deux classes les plus proches au sens de la "distance" entre classes choisie,*
  2. *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.*

3.4 Résultats

Graphes

Les graphes obtenus à l'issue d'une CAH sont présentés et illustrés dans la section suivante. Il s'agit du graphique d'aide au choix du nombre de classes et du dendrogramme, regroupant hiérarchiquement les observations et groupes par des branches dont la longueur est la distance entre les objets regroupés. Attention, la représentation du dendrogramme n'est pas unique, celui-ci est invariant par rotation d'une branche. L'ordre des observations sur l'axe horizontal est donc artificiel, il peut amener à rapprocher des observations qui sont de fait très éloignées l'une de l'autre car regroupées par de longues branches.

"Qualité" et choix du nombre de classes

La *corrélation cophénétique* est un indicateur de qualité d'une classification hiérarchique ou d'un dendrogramme est obtenue à partir de la notion de *distance cophénétique*. Cette distance est définie entre deux observations représentées dans un arbre par la hauteur des branches qui finissent par les réunir dans un même groupe. C'est également la distance entre les deux groupes contenant ces observations avant qu'ils ne soient réunis en un même groupe ; c'est par exemple la distance entre deux espèces dans un arbre phylogénétique. Toutes les distances ainsi définies entre les objets deux à deux sont rangées dans une matrice triangulaire de distances cophénétiques.

La "qualité" d'un arbre de classification peut se résumer par un coefficient de *corrélation cophénétique* entre les valeurs de la matrice de distances initiales, par exemple eucliennes, et celle des distances cophénétiques. Évidemment,

plus proche est cette valeur de 1, meilleure est la classification.

La *silhouette* (Rousseeuw, 1987)[8] d'une classification est un graphe montrant comment chaque observation appartient plus ou moins à sa classe. Supposons que n observations aient été réparties en k classes par un quelconque algorithme. Soit $a(i)$ la moyenne des dissimilarités (ou distances) de l'observation i avec toutes les autres observations au sein d'une même classe. Plus $a(i)$ est petit meilleur est l'assignation de i à sa classe ; $a(i)$ est la dissimilarité moyenne de i à cette classe.

Soit $b(i)$ la plus faible moyenne des dissimilarités (ou distances) de l'observation i à chaque autre classe dont i ne fait pas partie. La classe avec cette plus faible dissimilarité moyenne est appelé classe *voisine* de i car c'est la meilleure classe suivante pour l'observation i.

La silhouette de la $i$ème observations est alors donnée par

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}.$$

Plus ces valeurs sont proches de 1 et meilleure est la classification. La moyenne de toutes ces valeurs est un autre indicateur global de qualité.

Le choix du nombre de classes k est, comme le choix de la dimension en ACP, délicat à opérer. Plusieurs heuristiques on été proposées selon les critères précédents ou encore suivant le graphe de décroissance de la distance interclasses qui est aussi la décroissance de la variance inter-classe dans le cas du saut de Ward. La recherche d'un "coude" dans ce graphe est une indication heuristique du choix de k ; voir l'application dans la section suivante.

La statistique du "*gap*" est une proposition de (Tibshirani et al.)[9] pour tenter de rationaliser cette démarche. Soit D_r la somme de toutes les distances prises entre les observations deux à deux au sein d'une même classe $r = 1, k$; W_k est la moyenne pondérée (par al taille de la classe) de ces sommes de distances. Si la distance initiale est euclidienne, W est (à un facteur 2 près) la norme carrée de la matrice de variance intra-classe. L'idée est alors de comparée le graphe de $\log(W_k)$ par rapport à celui d'une distribution de référence obtenue par simulation (Monte Carlo) selon une loi uniforme et de rechercher le plus grand écart ou *gap*. La fonction `clusGap` qui implémente ce critère dans la librairie `cluster` propose 5 méthodes ou critères ! pour rechercher le plus grand *gap*. Attention, cette fonction n'accepte que des données sous la forme d'une matrice de variables quantitatives, pas celle d'une matrice de distances ou dissimilarités.

Enfin , dans le contexte de mélanges supposés gaussiens, c'est-à-dire si l'hypothèse d'une situation gaussienne multidimensionnelle, le choix du nombre de classes s'apparente à une sélection de modèle par des critères AIC, BIC, spécifiques. Il n'est pas abordé dans ce cours aperçu des méthodes de classification non-supervisée.

3.5 Illustration

Les données sont celles déjà représentées à l'aide du MDS : 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 qui n'est pas euclidienne à cause du relief.

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 3.5). 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.
  La fonction `clusGap` ne permet pas de calculer la statistique de *gap* sur une matrice de distances. La corrélation cophénétique de l'arbre est de $0,64$ mais cela est guère utile dans l'absolu tandis que les silhouettes

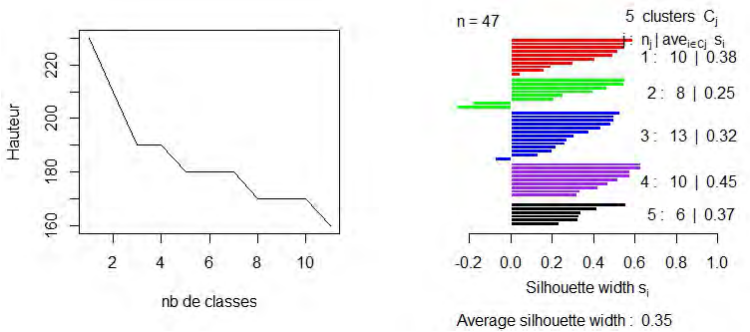

FIGURE 1 – *Villes : Décroissance de la variance interclasses à chaque regroupement dans le cas du saut de Ward (à gauche) et à droite silhouettes des observations dans leur classe respective.*

sont représentées dans la figure 3.5.

- Le *dendrogramme* (cf. figure 3.5) 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 par l'un ou l'autre des critères proposés, 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 :

- une ACP si la classification a été opérée sur des variables quantitatives assorties d'une métrique euclidienne,
- une AFCM si la classification a été opérée sur les composantes d'une AFCM de variables qualitatives,
- un MDS dans le cas de l'exemple (figure 3.5) car la classification est directement calculée sur un tableau de distance.

Signalons qu'il est courant, dans la pratique, de mettre en œuvre, à l'issue d'une CAH, une méthode de ré-allocation dynamique avec pour nombre de

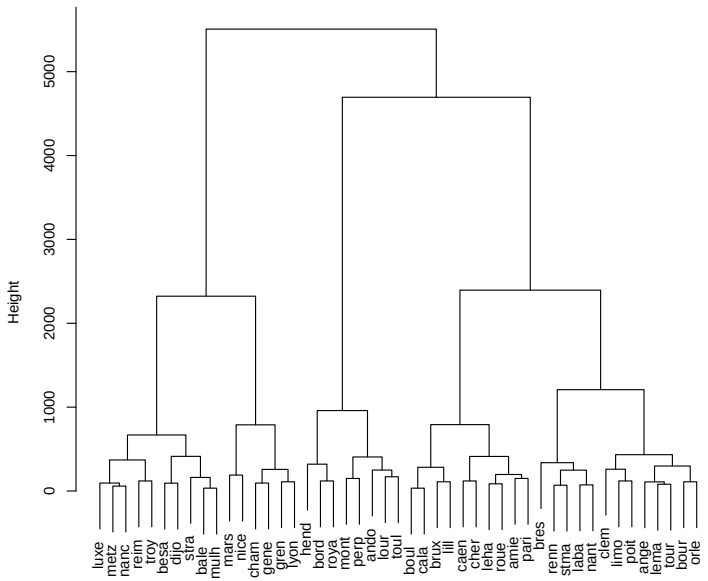

FIGURE 2 – *Villes : Exemple d'un dendrogramme issu de la classification des données par CAH et saut de Ward.*

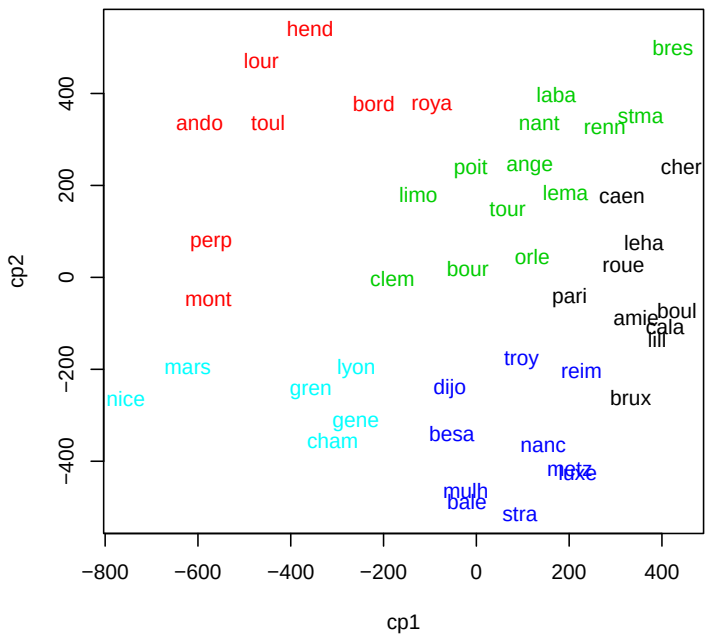

FIGURE 3 – *Villes : Représentation des classes (couleurs) obtenues par CAH dans les coordonnées du MDS.*

classes celui choisi par CAH et pour 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" ; il est recommandé de systématiquement compléter une classification par une représentation factorielle.

4 Agrégation autour de centres mobiles

4.1 Principes

Différents types d'algorithmes ont été définis autour du même principe de *ré-allocation 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. 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, par exemple, à tirer aléatoirement k individus, l'algorithme répète deux opérations jusqu'à la convergence d'un critère :

1. Chaque individu est affecté à la *classe* dont le centre est le plus proche au sens d'une métrique.
2. Calcul des k *centres* des classes ainsi constituées.

4.2 Principale méthode

Il s'agit de la version proposé par Forgy (1965)[3] des algorithmes de type *kmeans*.

ALGORITHME 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.*
  1. *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.*
  2. *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.*

4.3 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.

4.4 Variantes

kmeans

Toujours sous la même appellation (une option de la commande `kmeans` de R) Mac Queen (1967)[7] a proposé une modification de l'algorithme précédent. 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 la solution dépend de l'ordre des individus dans le fichier.

Nuées dynamiques

La variante proposée par Diday (1973)[2] et parallèlement par Hartigan et Wong (1979)[4] 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. Diday (1973) a également proposé la recherche de formes fortes communes à plusieurs partitions issues d'initialisations différentes.

Partitionning Around Medoïds

Cet algorithme (PAM), proposé par Kaufman & Rousseeuw (1990)[6], permet de classifier des données de façon plus robuste, c'est-à-dire moins sensible à des valeurs atypiques. Le noyau d'une classe est alors un médoïd c'est-à-dire l'observations d'une classe qui minimise la moyenne des distances ou dissimilarités aux autres observations de la classes. Une différence majeur avec l'algorithme *kmeans* est qu'un médoïd fait partie des données et permet donc de partitionner des matrices de dissimilarités. En contre-partie, il est limité par le nombre d'observations (matrice de dissimilarités à stocker) et en temps de calcul (algorithme en $\mathcal{O}(n^2)$). Il fonctionne de manière analogue à celui de Mac Queen. À chaque itération, un médoïd est mis en concurrence avec un autre individu aléatoire. Si l'échange améliore le critère, cet individu devient le nouveau médoïd.

D'autres algorithmes ont été proposés pour des types de données spécifiques : k-modes (Huang, 1998)[5] pour des variables qualitatives et k-prototypes pour des variables mixtes. Ils sont disponibles dans R.

La classification des villes par partitionnement autour de médoïds est fournie dans la figure 4.4 ; le nombre de classes est fixé *a priori* à 5 comme le suggère la CAH alors que les classes obtenues sont sensiblement différentes.

4.5 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 ré-allocation dynamique nécessite de fixer *a priori* le nombre de classes. La stratégie suivante, adaptée aux grands ensembles de données, permet de contourner ces difficultés.

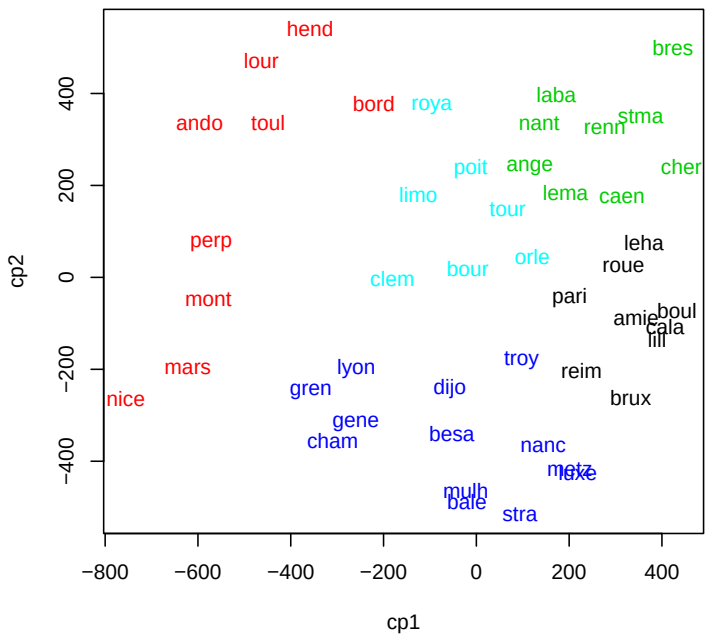

FIGURE 4 – *Villes : Représentation des classes (couleurs) obtenues par PAM dans les coordonnées du MDS.*

1. Exécuter une méthode de ré-allocation dynamique en demandant un grand nombre de classes, de l'ordre de 10% de n.
2. Sur les barycentres des classes précédentes, exécuter une classification hiérarchique puis déterminer un nombre "optimal" k de classes.
3. Exécuter une méthode de ré-allocation 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.

5 Données génomiques

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, n'est pas très aisée (figure 5).

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. 5) 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

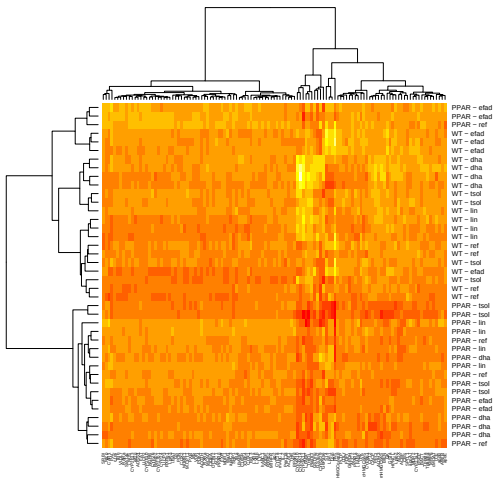

FIGURE 5 – *Souris : double classification ascendante hiérarchique des individus-souris et des variables-gènes selon la méthode de Ward, avec la distance euclidienne.*

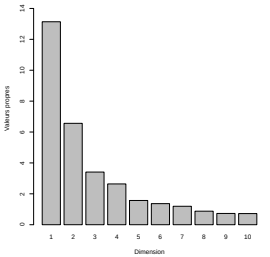

FIGURE 6 – *Souris : éboulis des valeurs propres pour le MDS de la matrice de distance euclidienne inter-gènes.*

en deux groupes, à l'exception de trois souris de type `PPAR` ayant suivi les régimes `efad` (pour deux d'entre elles) et `ref`. 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`.

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) nous oriente vers une représentation du MDS en deux dimensions.

La représentation de la figure 7 est analogue à celle déjà présentée en application du MDS. Elle est simplement complétée par un codage en couleurs

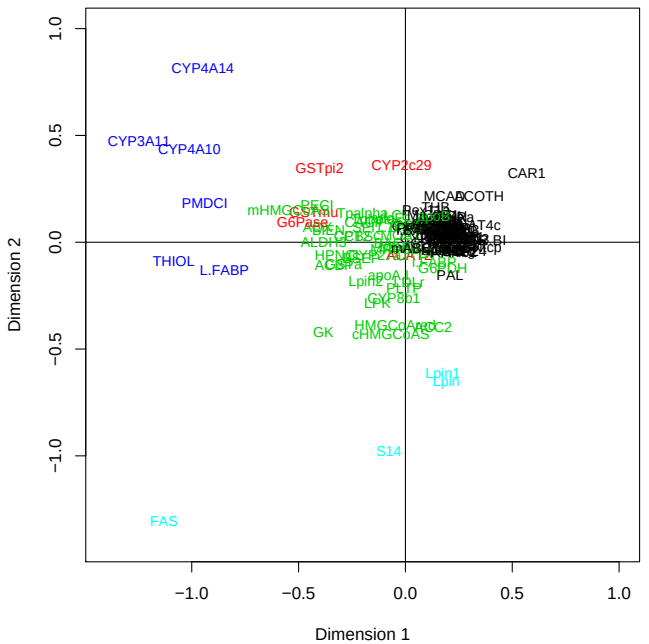

FIGURE 7 – *Souris : représentation par positionnement multidimensionnel (distance euclidienne) des 5 groupes issues de la classification hiérarchique des gènes.*

des gènes, selon leur appartenance à une classe issue 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 (sur-exprimé chez les souris `PPAR`) et un groupe de gènes (`CYP3A10`, `CYP4A10`, `CYP4A14`, `PMDCI`, `THIOL` et `L-FABP`) qui est sur-exprimé 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`.

Nous laissons au lecteur l'appréciation sur le nombre de combinaisons d'options possibles qui sont offertes par l'ensemble de ces outils : centrage, réduction, distance, critère de saut, projection, classification !

6 En guise de conclusion

Quelle méthode ou combinaison de méthodes associées à quelles options faut-il choisir et sur quel critère ? Réponse, celle et ceux qui fournissent des résultats les plus "utiles". Les techniques mises en œuvre sont fondamentalement "exploratoires", pas confirmatoires ni décisionnelles. Un résultat "utile" fournit de nouveaux éclairages, un point de vue fructueux sur des données complexes et ainsi une meilleure compréhension des interactions en jeu. C'est une étape préalable à la construction de futures modélisations et d'inférences qui seront à confirmer. Ainsi, associer au sein d'une même classe, des gènes de fonctions inconnues à d'autres de fonctions connues est une stratégie fréquente du biologiste pour poser des hypothèses sur l'annotation de ces gènes, hypothèses à infirmer ou confirmer par de nouvelles expériences.

Attention, la pertinence des résultats et la fiabilité des interprétations reposent sur une juste connaissance des méthodes, des options, des hypothèses sous-jacentes à ces méthodes : comment interpréter une proximité, au sens de

quelle distance avec quelle confiance ? sinon, l'orientation du travail pour poser de nouvelles hypothèses risque de se fourvoyer ou au mieux adopter une marche aléatoire (cf. devise Schadok) :

> *...en essayant continuellement on finit par réussir donc, plus ça rate, plus on a de chance que ça marche...*

Références

[1] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier et H. Ralambondrainy, *Classification automatique des données*, Dunod, 1989.

[2] E. Diday, *The dynamic clusters method in nonhierarchical clustering*, International Journal of Computer & Information Sciences **2** (1973), n° 1, 61–88.

[3] R. Forgy, *Cluster Analysis of Multivariate Data : Efficiency versus Interpretability of Classification*, Biometrics (1965), n° 21, 768–769.

[4] J. A. Hartigan et M. A. Wong, *Algorithm AS 136 : a k-means clustering algorithm*, Applied Statistics **28** (1979), 100–108.

[5] Zhexue Huang, *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*, Data Min. Knowl. Discov. **2** (1998), n° 3, 283–304.

[6] Leonard Kaufman et Peter J. Rousseeuw, *Finding Groups in Data – An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[7] J. Macqueen, *Some methods for classification and analysis of multivariate observations*, In 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967, p. 281–297.

[8] Peter J. Rousseeuw, *Silhouettes : A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics **20** (1987), n° 0, 53 – 65.

[9] Robert Tibshirani, Guenther Walther et Trevor Hastie, *Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society : Series B (Statistical Methodology) **63** (2001), n° 2, 411–423.

NMF Factorisation par matrices non négatives

Résumé

Réduction de dimension par factorisation d'une matrice creuse sous contrainte de non négativité des facteurs. Contrairement à l'ACP les facteurs ne sont pas orthogonaux et ne permettent pas de représentation mais, réduisant la dimension ils permettent classifications non supervisées et modèles de prévision. Description sommaire des nombreuses options de ces algorithmes principalement conçues pour l'analyse des très grandes matrices du e-commerce, text mining...

Retour au plan du cours.

1 Introduction

Le principe de la factorisation $\mathbf{X} = \mathbf{UV}'$ d'une matrice est largement utilisé en analyse en composantes principales qui utilise la décomposition en valeurs singulières de la matrice $\mathbf{X}$ (SVD) pour construire des facteurs orthogonaux deux à deux. Paatero et Tapper (1994)[5] puis Lee et Seung (1999)[4] ont proposé une autre décomposition sans contrainte d'orthogonalité mais avec celle de non négativité des matrices des facteurs afin d'en simplifier l'interprétation et sur la base d'une motivation "neuronale" : les neurones ne fonctionnent que de façon additive, pas soustractive. Cette technique a depuis été depuis largement utilisé dans de très nombreux domaines : imagerie, reconnaissance de formes, fouille de textes, systèmes de recommandations, génomique, avec pour objectif d'étudier la structure des très grandes matrices creuses. La bibliographie s'est donc largement développée autour de ce thème en proposant différentes versions de l'algorithme avec différentes initialisations et contraintes, par exemple de parcimonie, dont certaines parallélisables, et tout un ensemble d'applications.

La NMF est donc une technique de réduction de dimension adaptée aux matrices creuses contenant des données positives, par exemple des occurrences ou dénombrements de mots, de pannes... La méthode est donc plus adaptée à certaines situations que la SVD mais cela a un prix ; la complexité algorithmique de la SVD est polynomiale de l'ordre du produit $n \times p$ des dimensions de la matrice. La complexité de la NMF est un problème *nondeterministic polynomial - NP* ; l'existence d'un algortihme de complexité polynomiale est inconnue. En revanche, il existe des approches itératives efficaces mais convergeant vers une solution locale sauf dans des cas très spécifiques (Donoho et Stodden, 2003)[2] ; contrairement à la SVD qui conduit à une solution unique (vecteurs propres et valeurs propres d'une matrice).

Par ailleurs, les facteurs non orthogonaux ne permettent pas de représentation comme en ACP mais sont utilisés comme base d'une classification non supervisée ou préalable à une modélisation pour de l'apprentissage supervisée.

Lee et Seung (1999) illustre cette méthode sur la classification d'un corpus de 30991 articles de l'encyclopédie Grolier. Plutôt que de classer ces articles par thèmes choisis *a priori*, ils sont classés sur la base d'un vocabulaire de 15276 mots. chaque article se décompose (coefficients positifs), en principe parcimonieusement, sur des "facteurs" ou thèmes, eux-mêmes définis chacun par un sous-ensemble petit, jugé "pertinent", de ces mots. En traitement d'images, un corpus se classifie à partir de facteurs ou motifs élémentaires d'images, en génomique par rapport à des "métagènes". L'approche non supervisée est ainsi susceptible de révéler des structures cachées ou des tendances sans *a priori*. Par ailleurs, les facteurs de décomposition n'étant pas orthogonaux, des superpositions apparaissent : des même mots participants à plusieurs thèmes, des gènes à plusieurs fonctions...

2 NMF : méthode et implémentations

La description présente de la méthode de NMF ne se veut pas exhaustive ; elle est axée sur l'implémentation réalisée dans le package éponyme par Gaujoux et Seoighe (2010)[3] afin d'en préciser les options et critères mis en œuvre.

2.1 Principes

Soit $\mathbf{X}$ une matrice $(n \times p)$ ne contenant que des valeurs non négatives et sans ligne ou colonne ne comportant que des 0 ; r un entier choisi relativement petit devant n et p.

La factorisation non-négative de la matrice $\mathbf{X}$ est la recherche de deux matrices $\mathbf{W}_{n\times r}$ et $\mathbf{H}_{r\times p}$ ne contenant que des valeurs positives ou nulles et dont le produit approche $\mathbf{X}$.

$$\mathbf{X} \approx \mathbf{WH}.$$

Le choix du *rang* de factorisation $r << min(n,p)$ assure une réduction drastique de dimension et donc des représentations parcimonieuses. Évidemment, la qualité d'approximation dépend de la parcimonie de la matrice initiale.

La factorisation est résolue par la recherche d'un optimum local du problème d'optimisation :

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \left[L(\mathbf{X},\mathbf{WH}) + P(\mathbf{W},\mathbf{H})\right].$$

L est une fonction perte mesurant la qualité d'approximation et P une fonction de pénalisation optionnelle ; L est généralement soit un critère de moindres carrés (LS ou norme de Frobenius des matrices ou "norme trace"), soit la divergence de Kullback-Leibler (KL) ; P est une pénalisation optionnelle de régularisation utilisée pour forcer les propriétés recherchées des matrices $\mathbf{W}$ et $\mathbf{H}$, par exemple, la parcimonie des matrices ou la régularité des solutions dans le cas de données spectrales.

$$\begin{aligned} LS : L(\mathbf{A},\mathbf{B}) &= \text{tr}\left((\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})'\right) = \sum_{i,j}(a_{i,j}-b_{i,j})^2, \\ KL : L(\mathbf{A},\mathbf{B}) &= KL(\mathbf{A}||\mathbf{B}) = \sum_{i,j} a_{i,j}\log(\frac{a_{i,j}}{b_{i,j}}) - a_{i,j} + b_{i,j}. \end{aligned}$$

Dans la librairie NMF de R, construite surtout pour des applications en génomiques, les variables (*features*) sont en ligne et les individus / échantillons (*samples*) sont en colonnes. Ceci n'a pas d'importance lorsque le critère des moindres carrés est utilisé (LS), la résolution est invariante par transposition mais a du sens avec la divergence de Kullback-Leibler qui introduit une dissymétrie entre lignes et colonnes.

N.B. Non seulement la solution est locale car la fonction objectif n'est pas convexe en $\mathbf{W}$ et $\mathbf{H}$ mais en plus la solution n'est pas unique. Toute matrice $\mathbf{D}_{r\times r}$ non négative et inversible fournit des solutions équivalentes en terme d'ajustement :

$$\mathbf{X} \approx \mathbf{WDD^{-1}H}.$$

Une fois la factorisation construite il est ensuite facile d'utiliser ces matrices $\mathbf{W}$ et $\mathbf{H}$ pour construire des classifications (CAH, k-means), représentations (ACP, MDS), et prévisions à l'aide d'une des nombreuses méthodes d'apprentissage.

2.2 Algorithmes

De nombreuses variantes algorithmiques ou sur la forme des pénalisations ont été publiées et implémentées généralement en Matlab, parfois en C, quelques unes spécifiques en R ; Berry et al. (2007)[1] proposent un tour d'horizon de certaines tandis que Gaujoux et Seoighe (2010) en ont implémentées dans R pour rendre facilement possible la comparaison des résultats. Trois familles d'algorithmes sont généralement citées :

- Standard NMF algorithm with multiplicative update,
- Alternate Least Square (ALS) algorithm,
- Descente du gradient.

Chacun de ces algorithmes peut par ailleurs être initialisé de différentes façons :

- plusieurs initialisations aléatoires de $\mathbf{W}$ et $\mathbf{H}$, le meilleur ajustement est conservé,
- non-negative double singular value decomposition (NNSVD),
- une classification (k-means) des lignes ou des colonnes,
- parts positives de matrices issues d'une analyse en composantes indépendantes (ACI),
- ...

Entre le choix de la fonction objectif : fonction perte (LS ou KL) et l'éventuelle pénalisation (L^1, L^2, régularité), le choix de l'algorithme ou d'une de ses variantes, le choix de l'initialisation... cela fait beaucoup d'options à comparer, tester. Comme toujours avec une nouvelle méthode et la pression de publication, de très nombreuses variants apparaissent avant qu'une sélection "naturelle" n'opère pour aboutir à des choix plus efficaces et consensuels d'options en fonction du type de données traitées.

Berry et al. (2007)[1] décrivent très brièvement les principes de ces différents algorithmes et commentent leurs propriétés : convergence, complexité.

L'algorithme initial de Lee et Seung (1999)[4] (*Multiplicative update algorithms*) peut converger vers un point stationnaire pas nécessairement minima

local, voire un point de la frontière même pas point stationnaire. Ces cas sont heureusement rares en pratique mais la convergence est considérée comme lente, demandant plus d'itérations que ses concurrents alors que chaque itération nécessite de nombreux calculs ($O(n^3)$). Les algorithmes de descente du gradient posent des questions délicates concernant le choix des deux pas de descente. La dernière famille d'algorithme : moindres carrés alternés (ALS), exploite le fait que si le problème n'est pas convexe en à la fois $\mathbf{W}$ et $\mathbf{H}$, il l'est soit en $\mathbf{W}$ soit en $\mathbf{H}$. Il suit le principe ci-dessous et possède de bonnes propriétés (convergence, complexité).

ALGORITHME 1 : ALS

$\mathbf{W}$ =*random*(n, r)
for $i = 1$ à *Maxiter* ***do***
Résoudre en $\mathbf{H}$ *:* $\mathbf{W}'\mathbf{W}\mathbf{H} = \mathbf{W}'\mathbf{X}$
Mettre à 0 *les termes négatifs de* $\mathbf{H}$
Résoudre en $\mathbf{W}$ *:* $\mathbf{H}\mathbf{H}'\mathbf{W}' = \mathbf{H}\mathbf{X}'$
Mettre à 0 *les termes négatifs de* $\mathbf{W}$
end for

L'un des inconvénients du (*Multiplicative update algorithms*) originel est que si un élément des matrices $\mathbf{W}$ ou $\mathbf{H}$ prend la valeur 0, il reste à cette valeur, n'explorant ainsi pas de solutions alternatives. L'ALS est lui plus souple en permettant d'échapper à de mauvaises solutions locales.

La librairie `NMF` de R implémente 11 méthodes ; 9 sont basées sur l'algorithme initial de Lee et Seung (1999)[4] (*Multiplicative update algorithms*) avec différentes options de perte (LS, KL) et de pénalisation ou d'arrêt, deux sont basées sur les moindres carrés alternés (ALS) avec contrainte de parcimonie sur les lignes ou les colonnes. Systématiquement, l'option est offerte, et encouragée, de lancer plusieurs exécutions à partir de plusieurs initialisations aléatoires pour sélectionner les options "optimales" puis, une fois les choix opérés, pour retenir la meilleure parmi un ensemble d'exécutions.

2.3 Critères de choix

Les auteurs proposent différents critères pour aider aux choix des méthodes, algorithmes et paramètres, notamment celui du rang r de factorisation, pouvant intervenir au cours d'une étude. Ceux-ci sont illustrés dans la section suivante sur un jeu de données publiques. Un premier tableau (1) fournit des :

- résidus, part de variance expliquée, indice de parcimonie (*sparseness*), pour évaluer la qualité de l'ajustement,
- coefficient de corrélation cophénétique, pureté, entropie ou silhouette pour évaluer la "stabilité" sur plusieurs exécutions.

L'évaluation de la "stabilité" de plusieurs exécutions de NMF repose sur des critères (silhouette, consensus, corrélation cophénétique) issues des méthodes de classification non supervisée. Pour adapter ces critères à la NMF, la notion de classe d'une observation (resp. d'une variable) est remplacée par la recherche du facteur, ou élément de la base (colonne de $\mathbf{W}$ resp. de $\mathbf{H}$), pour laquelle l'observation (resp. la variable) a obtenu la plus forte contribution.

Comme pour le choix d'une dimension, d'un nombre de classes, seules des heuristiques sont proposées dans la littérature pour le difficile choix de r pour lequel il n'y a pas de critère nettement tranché. C'est finalement l'interprétation, biologique ou autre, qui oriente le choix en sous main, ou encore ci-dessous la relative stabilité d'une classification non-supervisée.

2.4 Graphiques

La librairie `NMF` propose tout un ensemble de graphiques intégrant chacun une pléthore d'options qu'il serait fastidieux de décrire exhaustivement ; se reporter à la documentation en ligne et à l'article de référence (Gaujoux et Seoighe, 2010)[3]. Leur présentation est largement inspirée des habitudes de la bioinformatique qui mettent en avant des graphes de type *heatmap*.

Des premiers graphiques, non représentés sur l'exemple mais qu'il est facile d'obtenir en exécutant le scénario, visualisent les valeurs (*heatmap*) positives ou nulles des matrices $\mathbf{W}$ (resp. $\mathbf{H}$) de la décomposition. Par défaut une classification ascendante hiérarchique (métrique euclidienne, *average linkage*) des lignes (resp. colonnes) est associée. Il est évidemment possible de modifier ces options par défaut ou d'intégrer les résultats d'une classification exécutée par ailleurs.

Des *consensus maps* sont proposées pour aider au choix de la méthode (figure 1) et au choix de la dimension (figure 3). Attention, ces graphiques sont construits sur les colonnes (variables ou *features*) de la matrice $\mathbf{X}$ et dépendent du choix initial de décomposer la matrice ou de sa transposée. Ces graphiques

montrent si, au cours de plusieurs exécutions de l'algorithme pour différentes méthodes ou pour différentes valeurs du rang r, les mêmes variables sont au mieux représentées par le même facteur. C'est donc une information sur la stabilité de l'optimisation obtenue par différentes initialisations.

Les mêmes indicateurs, que ceux présentés dans un tableau (1) pour le choix de la méthode, sont déclinés dans des graphiques (figure 2) avec le rang r des matrices en abscisse.

Enfin, un dernier graphique (figure 4) trace une *heatmap* représentant les valeurs de la matrice initiale $\mathbf{X}$ dans laquelle les lignes et colonnes sont réorganisées par double classification ascendante hiérarchique. Ces classifications sont construites sur les matrices en utilisant par défaut la distance euclidienne et le critère de saut moyen.

TABLE 1 – Critères pour chacune des méthodes testées.

Méthode	brunet	lee	snmf/l	snmf/r
sparseness basis	0.42	0.38	0.39	0.38
sparseness coef	0.87	0.74	0.69	0.74
silhouette coef	0.88	0.73	0.78	0.82
silhouette basis	0.57	0.62	0.51	0.39
residuals	23.k	5.4k	5.6k	5.6k
niter	510	2000	380	460
cophenetic	0.90	0.97	1.00	1.00
dispersion	0.67	0.82	1.00	0.97
silhouette consensus	0.50	0.84	0.98	0.95

3 Exemple

3.1 Les données

L'illustration de la factorisation non négative d'une matrice utilise les données décrites dans le scénario explorant les spécificités d'un corpus de pourriels. Elles se présentent sous une forme classique en fouille de texte d'un tableau avec en lignes des messages et en colonnes des nombres ou taux d'occurrences de mots ou caractères spécifiques. La nature des données : matrice très creuse pouvant présenter des valeurs très disparates rend les techniques factorielles habituelles (ACP, AFCM) peu adaptées. Le principal objectif sur ces données est de prévoir le statut *spam* ou non *spam* d'un message en fonction de son contenu et c'est l'objet d'un autre scénario. Il s'agit, dans un premier temps de les décrire, par exemple, en représentant et classifiant les principaux mots clefs.

3.2 Choix de méthode, de rang

La méthode (critère et algorithme) "optimale" est choisie en consultant le tableau 1 et les graphiques de la figure method. Sur ces données, il n'est pas difficile de se déterminer pour une méthode de moindres carrés (`snmf/l`) convergeant plus rapidement et présentant des valeurs optimales (cophenetic, residuals,...) ainsi que la meilleure stabilité sur plusieurs exécutions.

Ce choix étant arrêté, les figures 2 et 3 conduisent de façon consensuelle au choix de $r = 5$: corrélation cophénétique de 1 avant décroissance et meilleur graphique de consensus.

La dernière représentation nécessite évidemment d'être agrandie pour être mieux interprétée. Il est néanmoins facile d'identifier les principaux critères (nombre de lettre capitales ;...) regroupés dans une même classe et correspondant simultanément à une classes de pourriels. En revanche le mot-clef "georges", qui est le prénom du destinataire, est isolé et caractérise des courriels correctes. Une analyse plus fine permettrait d'identifier le rôle d'autres classes de mots.

Références

[1] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca et Robert J. Plemmons, *Algorithms and applications for approximate nonnegative matrix factorization*, Computational Statistics & Data Analysis **52** (2007), n° 1, 155 – 173.

[2] David Donoho et Victoria Stodden, *When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts ?*, Advances in Neural Information Processing Systems 16 (S. Thrun, L.K. Saul et B. Schölkopf, réds.), MIT Press, 2004, p. 1141–1148.

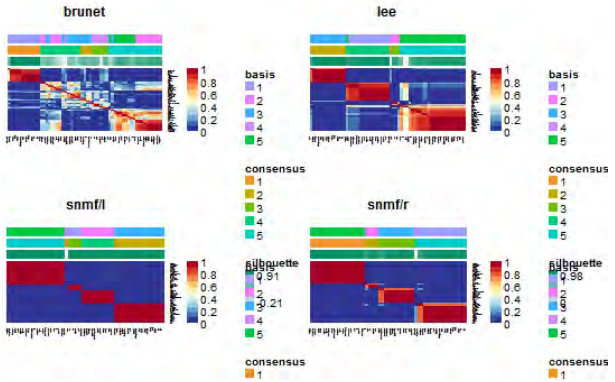

FIGURE 1 – *Spam : Matrice de "confusion" pour chaque méthode de la factorisation par NMF.*

[3] Renaud Gaujoux et Cathal Seoighe, *A flexible R package for nonnegativ matrix factorization*, BMC Bioinformatics **11** (2010), n° 1, 367, http //www.biomedcentral.com/1471-2105/11/367.

[4] D. Lee et S. Seung, *Learning the parts of objects by non-negative matri factorization*, Nature (1999).

[5] Pentti Paatero et Unto Tapper, *Positive matrix factorization : A non negative factor model with optimal utilization of error estimates of dat values*, Environmetrics **5** (1994), n° 2, 111–126.

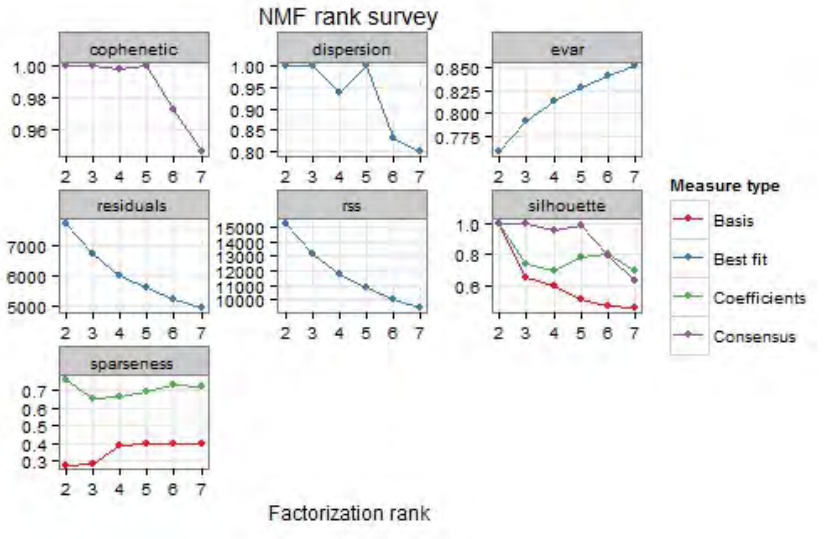

FIGURE 2 – *Spam : Évolution des différents critères en fonction du rang des matrices de la factorisation par NMF.*

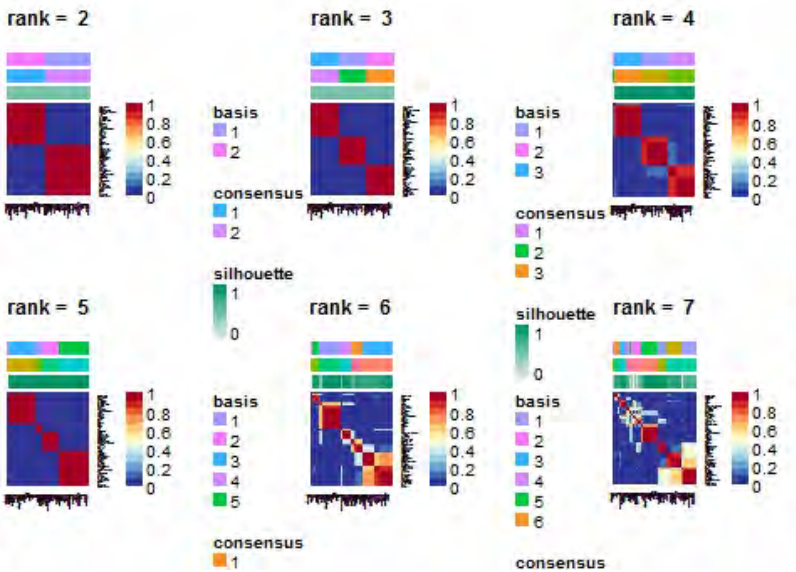

FIGURE 3 – *Spam : Matrice de "confusion" pour chaque valeur de rang des matrices de la factorisation par NMF.*

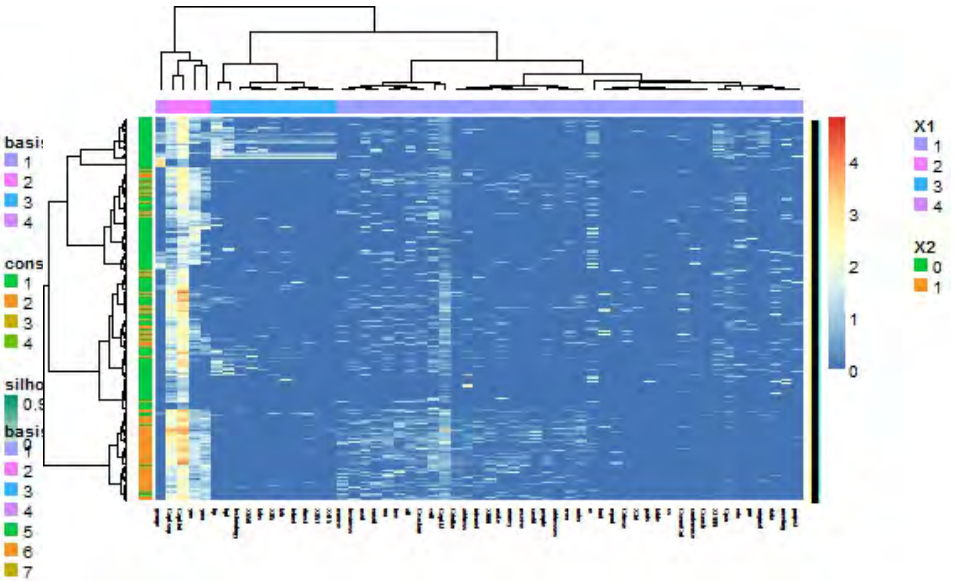

FIGURE 4 – *Spam : Double classification selon les facteurs de la factorisation par NMF et représentation de la matrice creuse initiale ; les message sont en lignes, les mots clefs en colonnes.*

Rappels et compléments d'algèbre linéaire

Résumé

Cette vignette rassemble des notations et rappels d'algèbre linéaire de niveau L. Il introduit les principaux théorèmes d'approximation matricielle par décomposition en valeurs singulières qui sont à la base des méthodes statistique factorielles.

Retour au plan du cours.

1 Notations

Dans tout ce qui suit, E et F sont deux espaces vectoriels réels munis respectivement des bases canoniques $\mathcal{E} = \{\mathbf{e}_j \ ; \ j = 1, \ldots, p\}$ et $\mathcal{F} = \{\mathbf{f}_i \ ; \ i = 1, \ldots, n\}$. On note indifféremment soit un vecteur de E ou de F, un endomorphisme de E, ou une application linéaire de E dans F, soit leurs représentations matricielles dans les bases définies ci-dessus.

2 Matrices

2.1 Notations

La matrice d'ordre $(n \times p)$ associée à une application linéaire de E dans F est décrite par un tableau :

$$\mathbf{A} = \begin{bmatrix} a_1^1 & \ldots & a_1^j & \ldots & a_1^p \\ \vdots & & \vdots & & \vdots \\ a_i^1 & \ldots & a_i^j & \ldots & a_i^p \\ \vdots & & \vdots & & \vdots \\ a_n^1 & \ldots & a_n^j & \ldots & a_n^p \end{bmatrix}.$$

On note par la suite :

$$\begin{aligned} a_i^j &= [\mathbf{A}]_i^j \text{ le terme général de la matrice,} \\ \mathbf{a}_i &= [a_i^1, \ldots, a_i^p]' \text{ un vecteur-ligne mis en colonne,} \\ \mathbf{a}^j &= [a_1^j, \ldots, a_n^j]' \text{ un vecteur-colonne.} \end{aligned}$$

2.1.1 Types de matrices

Une matrice est dite :

- *vecteur-ligne (colonne)* si $n = 1$ ($p = 1$),
- *vecteur-unité* d'ordre p si elle vaut $\mathbf{1}_p = [1, \ldots, 1]'$,
- *scalaire* si $n = 1$ et $p = 1$,
- *carrée* si $n = p$.

Une matrice carrée est dite :

- *identité* $(\mathbf{I}_p)$ si $a_i^j = \delta_i^j = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$,
- *diagonale* si $a_i^j = 0$ lorsque $i \neq j$,
- *symétrique* si $a_i^j = a_j^i, \forall (i, j)$,
- *triangulaire* supérieure (inférieure) si $a_i^j = 0$ lorsque $i > j$ $(i < j)$.

2.1.2 Matrice partitionnée en blocs

Matrices dont les éléments sont eux-mêmes des matrices. Exemple :

$$\mathbf{A}(n \times p) = \begin{bmatrix} \mathbf{A}_1^1(r \times s) & \mathbf{A}_1^2(r \times (p-s)) \\ \mathbf{A}_2^1((n-r) \times s) & \mathbf{A}_2^2((n-r) \times (p-s)) \end{bmatrix}.$$

2.2 Opérations sur les matrices

Somme : $[\mathbf{A} + \mathbf{B}]_i^j = a_i^j + b_i^j$ pour $\mathbf{A}$ et $\mathbf{B}$ de même ordre $(n \times p)$.

Multiplication par un scalaire : $[\alpha \mathbf{A}]_i^j = \alpha a_i^j$ pour $\alpha \in \mathbf{R}$.

Transposition : $[\mathbf{A}']_i^j = a_j^i$, $\mathbf{A}'$ est d'ordre $(p \times n)$.

$$(\mathbf{A}')' = \mathbf{A} \ ; (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}' \ ; (\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \ ;$$

$$\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix}' = \begin{bmatrix} \mathbf{A}_1^{1'} & \mathbf{A}_2^{1'} \\ \mathbf{A}_1^{2'} & \mathbf{A}_2^{2'} \end{bmatrix}.$$

Produit scalaire élémentaire : $a'b = \sum_{i=1}^{n} a_i b_i$ où a et b sont des vecteurs-colonnes.

Produit : $[\mathbf{AB}]_i^j = a_i' b^j$ avec $\mathbf{A}_{(n\times p)}, \mathbf{B}_{(p\times q)}$ et $\mathbf{AB}_{(n\times q)}$, et pour des matrices par blocs :

$$\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1^1 & \mathbf{B}_1^2 \\ \mathbf{B}_2^1 & \mathbf{B}_2^2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1^1\mathbf{B}_1^1 + \mathbf{A}_1^2\mathbf{B}_2^1 & \mathbf{A}_1^1\mathbf{B}_1^2 + \mathbf{A}_1^2\mathbf{B}_2^2 \\ \mathbf{A}_2^1\mathbf{B}_1^1 + \mathbf{A}_2^2\mathbf{B}_2^1 & \mathbf{A}_2^1\mathbf{B}_1^2 + \mathbf{A}_2^2\mathbf{B}_2^2 \end{bmatrix}$$

sous réserve de compatibilité des dimensions.

2.3 Propriétés des matrices carrées

La *trace* et le *déterminant* sont des notions intrinsèques, qui ne dépendent pas des bases de représentation choisies, mais uniquement de l'application linéaire sous-jacente.

2.3.1 Trace

Par définition, si $\mathbf{A}$ est une matrice $(p \times p)$,

$$\mathrm{tr}\mathbf{A} = \sum_{j=1}^{p} a_j^j,$$

et il est facile de montrer :

$$\begin{aligned} \mathrm{tr}\alpha &= \alpha, \\ \mathrm{tr}\alpha\mathbf{A} &= \alpha\mathrm{tr}\mathbf{A}, \\ \mathrm{tr}(\mathbf{A}+\mathbf{B}) &= \mathrm{tr}\mathbf{A} + \mathrm{tr}\mathbf{B}, \\ \mathrm{tr}\mathbf{AB} &= \mathrm{tr}\mathbf{BA}, \\ &\quad \text{reste vrai si } \mathbf{A} \text{ est } (n\times p) \text{ et si } \mathbf{B} \text{ est } (p\times n) \\ \mathrm{tr}\mathbf{CC}' &= \mathrm{tr}\mathbf{C}'\mathbf{C} = \sum_{i=1}^{n}\sum_{j=1}^{p}(c_i^j)^2 \\ &\quad \text{dans ce cas, } \mathbf{C} \text{ est } (n\times p). \end{aligned}$$

2.3.2 Déterminant

On note $|\mathbf{A}|$ le *déterminant* de la matrice carrée $\mathbf{A}$ $(p \times p)$. Il vérifie :

$$\begin{aligned} |\mathbf{A}| &= \prod_{j=1}^{p} a_j^j, \text{ si } \mathbf{A} \text{ est triangulaire ou diagonale,} \\ |\alpha\mathbf{A}| &= \alpha^p|\mathbf{A}|, \\ |\mathbf{AB}| &= |\mathbf{A}||\mathbf{B}|, \\ \begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{vmatrix} &= |\mathbf{A}||\mathbf{C}|, \\ \begin{vmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{vmatrix} &= |\mathbf{A}_1^1||\mathbf{A}_2^2 - \mathbf{A}_2^1(\mathbf{A}_1^1)^{-1}\mathbf{A}_1^2| \quad (1) \\ &= |\mathbf{A}_2^2||\mathbf{A}_1^1 - \mathbf{A}_1^2(\mathbf{A}_2^2)^{-1}\mathbf{A}_2^1|, \quad (2) \\ &\quad \text{sous réserve de la régularité de } \mathbf{A}_1^1 \text{ et } \mathbf{A}_2^2. \end{aligned}$$

Cette dernière propriété se montre en considérant les matrices :

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} & -\mathbf{A}_1^2(\mathbf{A}_2^2)^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \text{ et } \mathbf{BAB}',$$

puis en comparant les déterminants $|\mathbf{BAB}'|$ et $|\mathbf{A}|$.

2.3.3 Inverse

L'*inverse* de $\mathbf{A}$, lorsqu'elle existe, est la matrice unique notée $\mathbf{A}^{-1}$ telle que :

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}\,;$$

elle existe si et seulement si $|\mathbf{A}| \neq 0$. Quelques propriétés :

$$(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}, \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}.$$

2.3.4 Définitions

Une matrice carrée $\mathbf{A}$ est dite :

symétrique si $\mathbf{A}' = \mathbf{A}$,

singulière si $|\mathbf{A}| = 0$,
régulière si $|\mathbf{A}| \neq 0$,
idempotente si $\mathbf{AA} = \mathbf{A}$,
définie-positive si, $\forall \mathbf{x} \in \mathbb{R}^p, \mathbf{x}'\mathbf{Ax} \geq 0$, et si $x'\mathbf{A}x = 0 \Rightarrow x = 0$,
positive, ou semi-définie-positive, si, $\forall \mathbf{x} \in \mathbb{R}^p, \mathbf{x}'\mathbf{Ax} \geq 0$,
orthogonale si $\mathbf{AA}' = \mathbf{A}'\mathbf{A} = \mathbf{I}$ $(\mathbf{A}' = \mathbf{A}^{-1})$.

3 Espaces euclidiens

E est un espace vectoriel réel de dimension p isomorphe à $\mathbb{R}^p$.

3.1 Sous-espaces

- Un sous-ensemble E_q de E est un *sous-espace vectoriel* (s.e.v.) de E s'il est non vide et stable :

$$\forall (\mathbf{x}, \mathbf{y}) \in E_q^2, \forall \alpha \in \mathbb{R}, \alpha(\mathbf{x} + \mathbf{y}) \in E_q.$$

- Le q-uple $\{\mathbf{x}_1, \ldots, \mathbf{x}_q\}$ de E constitue un système *linéairement indépendant* si et seulement si :

$$\sum_{i=1}^{q} \alpha_i \mathbf{x}_i = 0 \Rightarrow \alpha_1 = \cdots = \alpha_q = 0.$$

- Un système linéairement indépendant $\mathcal{E}_q = \{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$ qui engendre dans E un s.e.v. $E_q = \text{vec}\{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$ en constitue une *base* et $\dim(E_q) = \text{card}(\mathcal{E}_q) = q$.

3.2 Rang d'une matrice

Dans ce sous-paragraphe, $\mathbf{A}$ est la matrice d'une application linéaire de $E = \mathbb{R}^p$ dans $F = \mathbb{R}^n$.

$$\begin{aligned}
\text{Im}(\mathbf{A}) &= \text{vect}\{\mathbf{a}^1, \ldots, \mathbf{a}^p\} \text{ est le s.e.v. de } F \textit{ image} \text{ de } \mathbf{A} ; \\
\text{Ker}(\mathbf{A}) &= \{x \in E ;\ \mathbf{A}x = 0\} \text{ est le s.e.v. de } E \textit{ noyau} \text{ de } \mathbf{A} ; \\
E &= \text{Im}(\mathbf{A}) \oplus \text{Ker}(\mathbf{A}) \text{ si } \mathbf{A} \text{ est carrée associée à un endomorphisme de } E. \\
\text{et } p &= \dim(\text{Im}(\mathbf{A})) + \dim(\text{Ker}(\mathbf{A})).
\end{aligned}$$

$$\begin{aligned}
\text{rang}(\mathbf{A}) &= \dim(\text{Im}(\mathbf{A})), \\
0 \leq \text{rang}(\mathbf{A}) &\leq \min(n, p), \\
\text{rang}(\mathbf{A}) &= \text{rang}(\mathbf{A}'), \\
\text{rang}(\mathbf{A} + \mathbf{B}) &\leq \text{rang}(\mathbf{A}) + \text{rang}(\mathbf{B}), \\
\text{rang}(\mathbf{AB}) &\leq \min(\text{rang}(\mathbf{A}), \text{rang}(\mathbf{B})), \\
\text{rang}(\mathbf{BAC}) &= \text{rang}(\mathbf{A}), \text{ si } \mathbf{B} \text{ et } \mathbf{C} \text{ sont régulières}, \\
\text{rang}(\mathbf{A}) &= \text{rang}(\mathbf{AA}') = \text{rang}(\mathbf{A}'\mathbf{A}).
\end{aligned}$$

Enfin, si $\mathbf{B}$ $(p \times q)$ est de rang $q (q < p)$ et $\mathbf{A}$ est carrée $(p \times p)$ de rang p, alors la matrice $\mathbf{B}'\mathbf{AB}$ est de rang q.

3.3 Métrique euclidienne

Soit $\mathbf{M}$ une matrice carrée $(p \times p)$, symétrique, définie-positive ; $\mathbf{M}$ définit sur l'espace E :

- un *produit scalaire* : $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \mathbf{x}'\mathbf{My}$,
- une *norme* : $\|\mathbf{x}\|_{\mathbf{M}} = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{M}}^{1/2}$,
- une *distance* : $\text{d}_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathbf{M}}$,
- des *angles* : $\cos\theta_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}}}{\|\mathbf{x}\|_{\mathbf{M}} \|\mathbf{y}\|_{\mathbf{M}}}$.

La matrice $\mathbf{M}$ étant donnée, on dit que :

- une matrice $\mathbf{A}$ est $\mathbf{M}$-*symétrique* si $(\mathbf{MA})' = \mathbf{MA}$,
- deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ sont $\mathbf{M}$-*orthogonaux* si $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = 0$,
- un vecteur $\mathbf{x}$ est $\mathbf{M}$-*normé* si $\|\mathbf{x}\|_{\mathbf{M}} = 1$,
- une base $\mathcal{E}_q = \{\mathbf{e}_1, \ldots, \mathbf{e}_q\}$ est $\mathbf{M}$-*orthonormée* si

$$\forall (i, j), \langle \mathbf{e}_i, \mathbf{e}_j \rangle_{\mathbf{M}} = \delta_i^j.$$

3.4 Projection

Soit W un sous-espace de E et $\mathcal{B} = \{\mathbf{b}^1, \ldots, \mathbf{b}^q\}$ une base de W ; $\mathbf{P}(p \times p)$ est une matrice de projection $\mathbf{M}$-orthogonale sur W si et seulement si :

$$\forall \mathbf{y} \in E, \mathbf{Py} \in W \text{ et } \langle \mathbf{Py}, \mathbf{y} - \mathbf{Py} \rangle_{\mathbf{M}} = 0.$$

Toute matrice idempotente $(\mathbf{P}^2 = \mathbf{P})$ et $\mathbf{M}$-symétrique $(\mathbf{P}'\mathbf{M} = \mathbf{MP})$ est une matrice de projection $\mathbf{M}$-orthogonale et réciproquement.

3.4.1 Propriétés

- Les valeurs propres de $\mathbf{P}$ sont 0 ou 1 (voir § 4) :

$$\begin{array}{lll} \mathbf{u} \in W, & \mathbf{Pu} = \mathbf{u}, & \lambda = 1, \text{ de multiplicité } \dim(W), \\ \mathbf{v} \perp W, (\text{on note } \mathbf{v} \in W^{\perp}) & \mathbf{Pv} = 0, & \lambda = 0, \text{ de multiplicité } \dim(W^{\perp}). \end{array}$$

- $\mathrm{tr}\mathbf{P} = \dim(W)$.
- $\mathbf{P} = \mathbf{B}(\mathbf{B}'\mathbf{M}\mathbf{B})^{-1}\mathbf{B}'\mathbf{M}$, où $\mathbf{B} = \left[\mathbf{b}^1, \ldots, \mathbf{b}^q\right]$.
- Dans le cas particulier où les $\mathbf{b}^j$ sont $\mathbf{M}$-orthonormés :

$$\mathbf{P} = \mathbf{B}\mathbf{B}'\mathbf{M} = \sum_{i=1}^{q} \mathbf{b}^j \mathbf{b}^{j'} \mathbf{M}.$$

- Dans le cas particulier où $q = 1$ alors :

$$\mathbf{P} = \frac{\mathbf{b}\mathbf{b}'}{\mathbf{b}'\mathbf{M}\mathbf{b}}\mathbf{M} = \frac{1}{\|\mathbf{b}\|_{\mathbf{M}}}\mathbf{b}\mathbf{b}'\mathbf{M}.$$

- Si $\mathbf{P}_1, \ldots, \mathbf{P}_q$ sont des matrices de projection $\mathbf{M}$-orthogonales alors la somme $\mathbf{P}_1 + \cdots + \mathbf{P}_q$ est une matrice de projection $\mathbf{M}$-orthogonale si et seulement si : $\mathbf{P}_k\mathbf{P}_j = \delta_k^j \mathbf{P}_j$.
- La matrice $\mathbf{I} - \mathbf{P}$ est la matrice de projection $\mathbf{M}$-orthogonale sur $W^{\perp}$.

4 Eléments propres

Soit $\mathbf{A}$ une matrice carrée $(p \times p)$.

4.1 Définitions

- Par définition, un vecteur $\mathbf{v}$ définit une *direction propre* associée à une *valeur propre* λ si l'on a :

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

- Si λ est une valeur propre de $\mathbf{A}$, le noyau $\mathrm{Ker}(\mathbf{A} - \lambda\mathbf{I})$ est un s.e.v. de E, appelé sous-espace propre, dont la dimension est majoré par l'ordre de multiplicité de λ. Comme cas particulier, $\mathrm{Ker}(\mathbf{A})$ est le sous-espace propre associé, si elle existe, à la valeur propre nulle.
- Les valeurs propres d'une matrice $\mathbf{A}$ sont les racines, avec leur multiplicité, du *polynôme caractéristique* :

$$|\mathbf{A} - \lambda\mathbf{I}| = 0.$$

THÉORÈME 1. — *Soit deux matrices $\mathbf{A}(n \times p)$ et $\mathbf{B}(p \times n)$; les valeurs propres non nulles de $\mathbf{AB}$ et $\mathbf{BA}$ sont identiques avec le même degré de multiplicité. Si $\mathbf{u}$ est vecteur propre de $\mathbf{BA}$ associé à la valeur propre λ différente de zéro, alors $\mathbf{v} = \mathbf{Au}$ est vecteur propre de la matrice $\mathbf{AB}$ associé à la même valeur propre.*

Les applications statistiques envisagées dans ce cours ne s'intéressent qu'à des types particuliers de matrices.

THÉORÈME 2. — *Une matrice $\mathbf{A}$ réelle symétrique admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base orthonormée de E ; $\mathbf{A}$ se décompose en :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum_{k=1}^{p} \lambda_k \mathbf{v}^k \mathbf{v}^{k'}$$

où $\mathbf{V}$ est une matrice orthogonale $[\mathbf{v}^1, \ldots, \mathbf{v}^p]$ des vecteurs propres orthonormés associés aux valeurs propres λ_k, rangées par ordre décroissant dans la matrice diagonale $\mathbf{\Lambda}$.

THÉORÈME 3. — *Une matrice $\mathbf{A}$* réelle $\mathbf{M}$-symétrique *admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base $\mathbf{M}$-orthonormée de E ; $\mathbf{A}$ se décompose en :*

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'\mathbf{M} = \sum_{k=1}^{p} \lambda_k \mathbf{v}^k \mathbf{v}^{k'}\mathbf{M}$$

où $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^p]$ est une matrice $\mathbf{M}$-orthogonale ($\mathbf{V}'\mathbf{M}\mathbf{V} = \mathbf{I}_p$ et $\mathbf{V}\mathbf{V}' = \mathbf{M}^{-1}$) des vecteurs propres associés aux valeurs propres λ_k, rangées par ordre décroissant dans la matrice diagonale $\mathbf{\Lambda}$.

Les décompositions ne sont pas uniques : pour une valeur propre simple (de multiplicité 1) le vecteur propre normé est défini à un signe près, tandis que pour une valeur propre multiple, une infinité de bases $\mathbf{M}$-orthonormées peuvent être extraites du sous-espace propre unique associé.

Le rang de $\mathbf{A}$ est aussi le rang de la matrice $\mathbf{\Lambda}$ associée et donc le nombre (répétées avec leurs multiplicités) de valeurs propres non nulles.

Par définition, si $\mathbf{A}$ est positive, on note la racine carrée de $\mathbf{A}$:

$$\mathbf{A}^{1/2} = \sum_{k=1}^{p} \sqrt{\lambda_k} \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M} = \mathbf{V} \mathbf{\Lambda}^{1/2} \mathbf{V}' \mathbf{M}.$$

4.2 Propriétés

Si $\lambda_k \neq \lambda_j$,	$\mathbf{v}^k \perp_{\mathbf{M}} \mathbf{v}^j$;
$\mathrm{tr}\mathbf{A} = \sum_{k=1}^{p} \lambda_k$;	$\lvert\mathbf{A}\rvert = \prod_{k=1}^{p} \lambda_k$;
si $\mathbf{A}$ est régulière,	$\forall k, \lambda_k \neq 0$;
si $\mathbf{A}$ est positive,	$\lambda_p \geq 0$;
si $\mathbf{A}$ est définie-positive,	$\lambda_p > 0$;

4.3 Décomposition en Valeurs Singulières (DVS)

Il s'agit, cette fois, de construire la décomposition d'une matrice $\mathbf{X}(n \times p)$ rectangulaire relativement à deux matrices symétriques et positives $\mathbf{D}(n \times n)$ et $\mathbf{M}(p \times p)$.

THÉORÈME 4. — *Une matrice* $\mathbf{X}$ $(n \times p)$ *de rang* r *peut s'écrire :*

$$\mathbf{X} = \mathbf{U} \mathbf{\Lambda}^{1/2} \mathbf{V}' = \sum_{k=1}^{r} \sqrt{\lambda_k} \mathbf{u}^k \mathbf{v}^{k'}; \tag{3}$$

$\mathbf{U}$ $(n \times r)$ *contient les vecteurs propres* $\mathbf{D}$*-orthonormés* $(\mathbf{U}'\mathbf{D}\mathbf{U} = \mathbf{I}_r)$ *de la matrice* $\mathbf{D}$*-symétrique positive* $\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}$ *associés aux* r *valeurs propres non nulles* λ_k *rangées par ordre décroissant dans la matrice diagonale* $\mathbf{\Lambda}(r \times r)$ *;* $\mathbf{V}$ $(p \times r)$ *contient les vecteurs propres* $\mathbf{M}$*-orthonormés* $(\mathbf{V}'\mathbf{M}\mathbf{V} = \mathbf{I}_r)$ *de la matrice* $\mathbf{M}$*-symétrique positive* $\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{M}$ *associés aux mêmes valeurs propres. De plus,*

$$\mathbf{U} = \mathbf{X}\mathbf{M}\mathbf{V}\mathbf{\Lambda}^{-1/2} \text{ et } \mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{U}\mathbf{\Lambda}^{-1/2}.$$

5 Optimisation

5.1 Norme d'une matrice

L'espace vectoriel E de dimension p (resp. F de dimension n) est muni de sa base canonique et d'une métrique de matrice $\mathbf{M}$ (resp. $\mathbf{D}$). Soit $\mathbf{X}$ une matrice $(n \times p)$. L'ensemble $\mathcal{M}_{n,p}$ des matrices $(n \times p)$ est un espace vectoriel de dimension np ; on le munit du *produit scalaire* :

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{M},\mathbf{D}} = \mathrm{tr}\mathbf{X}\mathbf{M}\mathbf{Y}'\mathbf{D}. \tag{4}$$

Dans le cas particulier où $\mathbf{M} = \mathbf{I}_p$ et $\mathbf{D} = \mathbf{I}_n$, et en notant $\mathrm{vec}(\mathbf{X}) = \left[\mathbf{x}^{1'}, \ldots, \mathbf{x}^{p'}\right]'$ la matrice "vectorisée", ce produit scalaire devient :

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{I_p},\mathbf{I_n}} = \mathrm{tr}\mathbf{X}\mathbf{Y}' = \sum_{i=1}^{n} \sum_{j=1}^{p} x_i^j y_i^j = \mathrm{vec}(\mathbf{X})'\mathrm{vec}(\mathbf{Y}).$$

La *norme* associée à ce produit scalaire (4) est appelée norme trace :

$$\begin{aligned} \|\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} &= \mathrm{tr}\mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{D}, \\ \|\mathbf{X}\|^2_{\mathbf{I_p},\mathbf{I_n}} &= \mathrm{tr}\mathbf{X}\mathbf{X}' = \mathrm{SSQ}(\mathbf{X}) = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_i^j)^2 \\ &\quad \text{(SSQ signifie "sum of squares").} \end{aligned}$$

La *distance* associée à cette norme devient, dans le cas où $\mathbf{D}$ est une matrice diagonale ($\mathbf{D} = \mathrm{diag}(w_1, \ldots, w_n)$), le critère usuel des *moindres carrés* :

$$\mathrm{d}^2(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{M},\mathbf{D}} = \sum_{i=1}^{n} w_i \|\mathbf{x}_i - \mathbf{y}_i\|^2_{\mathbf{M}}.$$

5.2 Approximation d'une matrice

Les matrices $\mathbf{X}$, $\mathbf{M}$ et $\mathbf{D}$ sont définies comme ci-dessus ; $\mathbf{X}$ est supposée de rang r. On cherche la matrice $\mathbf{Z}_q$, de rang q inférieur à r, qui soit la plus proche possible de $\mathbf{X}$.

THÉORÈME 5. — *La solution du problème :*

$$\min_{\mathbf{Z}} \left\{ \|\mathbf{X} - \mathbf{Z}\|^2_{\mathbf{M},\mathbf{D}} ; \mathbf{Z} \in \mathcal{M}_{n,p},\ \mathit{rang}(\mathbf{Z}) = q < r \right\} \tag{5}$$

est donnée par la somme des q premiers termes de la décomposition en valeurs singulières (3) de $\mathbf{X}$ *:*

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

Le minimum atteint est :

$$\|\mathbf{X} - \mathbf{Z}_q\|^2_{\mathbf{M},\mathbf{D}} = \sum_{k=q+1}^{r} \lambda_k.$$

Les matrices $\mathbf{U}_q, \mathbf{\Lambda}_q$ et $\mathbf{V}_q$ contiennent les q premiers vecteurs et valeurs propres donnés par la DVS de $\mathbf{X}$; $\mathbf{Z}_q$ est appelée approximation de rang q de $\mathbf{X}$.

Ce théorème peut se reformuler d'une manière équivalente. On note $\widehat{\mathbf{P}_q}$ (resp. $\widehat{\mathbf{Q}_q}$) la projection $\mathbf{M}$-orthogonale sur $E_q = \text{Im}(\mathbf{V}_q)$ (resp. $\mathbf{D}$-orthogonale sur $F_q = \text{Im}(\mathbf{U}_q)$) :

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \sum_{k=1}^{q} \mathbf{v}^k \mathbf{v}^{k'} \mathbf{M} = \mathbf{V}_q \mathbf{V}'_q \mathbf{M} \\
\widehat{\mathbf{Q}_q} &= \sum_{k=1}^{q} \mathbf{u}^k \mathbf{u}^{k'} \mathbf{D} = \mathbf{U}_q \mathbf{U}'_q \mathbf{D}, \\
\mathbf{Z}_q &= \widehat{\mathbf{Q}_q} \mathbf{X} = \mathbf{X} \widehat{\mathbf{P}_q}'.
\end{aligned}$$

PROPOSITION 6. — *Avec les notations précédentes :*

$$\begin{aligned}
\widehat{\mathbf{P}_q} &= \textit{arg} \max_{\mathbf{P}_q} \left\{ \|\mathbf{X}\mathbf{P}'_q\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\quad \left. \mathbf{P}_q \textit{ projection } \mathbf{M}\textit{-orthogonale de rang } q < r \right\}, \\
\widehat{\mathbf{Q}_q} &= \textit{arg} \max_{\mathbf{Q}_q} \left\{ \|\mathbf{Q}_q\mathbf{X}\|^2_{\mathbf{M},\mathbf{D}} ; \right. \\
&\quad \left. \mathbf{Q}_q \textit{ projection } \mathbf{D}\textit{-orthogonale de rang } q < r \right\}.
\end{aligned}$$