Brigitte Le Roux

Analyse géométrique des données multidimensionnelles

Édition revue et augmentée du livre
« Analyse des données multidimensionnelles.
Statistique en Sciences Humaines »
paru en 1993 et rédigé par Henry Rouanet et Brigitte Le Roux

DUNOD

Illustration de couverture
Franco Novati

ISBN 978-2-10-059820-5

Table des matières

Avant–propos

La réalité est multidimensionnelle.
(Les journaux)

Le précédent livre, paru chez Dunod il y a vingt ans, avait besoin d'une nouvelle version revue et augmentée. Ainsi, j'ai complété les chapitres sur l'analyse des correspondances et l'analyse des correspondances multiples en y incluant les nouveaux développements de ces dernières années et j'ai ajouté deux chapitres l'un traitant de l'analyse des données structurées et l'autre des méthodes de classification euclidiennes. J'ai aussi changé le titre[1] afin de mettre l'accent sur l'*analyse géométrique*, qui marque la spécificité des méthodes présentées dans ce livre, et sur les *données multidimensionnelles*.

Dire que la *réalité est multidimensionnelle* est un truisme. Cependant la pensée statistique est imprégnée de l'idéologie que faire un travail scientifique signifie quantifier ("tout ce qui existe doit exister dans une certaine quantité"). La «réduction à l'uni–dimensionnalité» est quelquefois si futile que cela conduit maint chercheurs à un rejet massif de toute analyse statistique, comme on l'entend souvent dire «l'intelligence est multidimensionnelle, elle ne peut donc pas être mesurée».

Entre le qualitatif et le quantitatif, il y a la *géométrie* dont les objets (points, droites, plans, figures géométriques) peuvent être décrits par des nombres mais ne peuvent pas être réduits à des nombres. La pensée géométrique en statistique —avec l'idée que, pour transmettre l'information, un bon graphique vaut mieux qu'une multitude de nombres— est probablement aussi ancienne que la statistique elle–même avec son cortège de diagrammes divers et variés. A l'ère informatique, plutôt que se replier sur une approche qualitative, une élégante manière de relever le défi de la multidimensionnalité a été apportée par l'*analyse des données* que J-P. Benzécri, le statisticien–géomètre, a commencé à développer dès les années 1960.

Face aux données multivariées, l'AGD procède à la modélisation des données sous forme de nuages de points et base l'interprétation sur ces nuages. Les nuages de points ne sont pas donnés d'emblée, ils sont construits à partir de tableaux de données. Cette construction

[1] L'appellation "analyse géométrique des données" (AGD) nous a été suggérée par Patrick Suppes à la suite d'un séminaire à l'Université de Stanford, en 1996.

est fondée sur les structures mathématiques de l'algèbre linéaire abstraite. La formalisation de ces structures, en premier lieu la dualité entre mesures et variables, fait partie intégrante de cette approche ; on peut dire que l'AGD est, à proprement parler, l'approche géométrico–formelle de l'analyse multivariée. Toutefois, les nuages de points ne sont pas de simples représentations graphiques, ils ont une échelle de distance bien définie comme les cartes géographiques.

Dans les exposés des méthodes d'AGD, nous avons mis l'accent sur les points suivants :
— la *formalisation* qui est un guide précieux pour l'étape cruciale de construction des nuages ;
— les *aides à l'interprétation* qui sont les éléments indispensables à toute interprétation ;
— l'*analyse des correspondances multiples* qui est un outil puissant pour l'analyse des questionnaires et que nous avons étendue dans deux directions ;
— l'*analyse des données structurées* qui est une synthèse de l'AGD et de l'analyse de variance ;
— les méthodes de *classification euclidienne* qui sont des outils de plus en plus utiles au regard des grands ensembles de données ;
— les études de *données réelles* qui détaillent la stratégie d'analyse.

Ce livre est le fruit d'une longue expérience d'enseignement, en premier lieu à l'université Paris Descartes, puis à Sciences-Po ainsi qu'aux *ateliers d'analyse géométrique des données* organisés en France et à l'étranger. Le public concerné est, en premier lieu, les étudiants de licence et de master, mais aussi les praticiens et les chercheurs de toutes disciplines (sciences sociales, psychologie, médecine, etc.).

Le lecteur trouvera dans ce livre un exposé des méthodes avec les principales formules, mais aussi des guides pour l'interprétation des données, le tout illustré par de nombreux exemples. Plusieurs lectures devraient être possibles selon les connaissances du lecteur en mathématiques et statistique : une *lecture pratique* d'utilisateur et une lecture plus *technique* pour ceux qui ont une formation en mathématique, sans parler de son rôle de «clefs pour Benzécri» que certains y chercheront peut–être.

La présentation matérielle du livre est conçue de la façon suivante : chaque chapitre contient un exposé de style cours magistral suivi d'exercices résolus et souvent commentés. La rubrique *Bases mathématiques*, répartie à la fin des premiers chapitres, rend le livre

autonome ; elle pourra être consultée à tout moment. On trouvera en fin d'ouvrage juste avant les index (des auteurs, des symboles et des matières) des références bibliographiques. L'étude systématique des exercices d'analyse des données et des études de cas constitue à elle seule un véritable «cours d'analyse des données par la pratique».

Pour plus de détails sur l'organisation des chapitres, voir le §1.3 du chapitre *Prologue*, en particulier le diagramme de la page 11.

MODE DE LECTURE DU LIVRE

En gros caractères : texte principal ;
en encadré, les principaux résultats et propriétés.

En moyens caractères : développements secondaires, pouvant être omis en première lecture, commentaires et *Bases Mathématiques*.

En moyens caractères précédés d'un ◇ : remarques

Table des matières : p. I.

Références : p. 395

Index : des auteurs p. 401, des symboles p. 402 ; terminologique p. 404.

Remerciements

Je voudrai en premier lieu remercier les laboratoires MAP5/CNRS (Université Paris Descartes) et CEVIPOF/CNRS (Sciences-po Paris) qui m'accueillent en tant que chercheur associé.

Je remercie chaleureusement Pierre Cazes (Université Paris Dauphine) dont la lecture approfondie et critique m'a été précieuse et permis des améliorations significatives. Je remercie également, pour leurs remarques, Philippe Bonnet (Université Paris descartes), J-L Durand (Université Paris Nord), Frédérik Cassor (Cevipof) et Solène Bienaise (Université Paris Dauphine).

Toute ma gratitude va aux experts grâce à qui j'ai pu progresser aussi bien dans la théorie que dans la pratique de l'analyse géométrique des données : Frédéric Lebaron (Université de Versailles–Saint-Quentin), Jean Chiche et Pascal Perrineau (Cevipof, Sciences-Po), Johs Hjellbrekke et Olav Korsnes (Université de Bergen), Donald Broady et Mikael Börjesson (Université d'Uppsala), Lennart Rosenlund (Université de Stavanger), Annick Prieur (Université d'Aalborg), Mike Savage (London School of Economics) et le CRESC *(Université de Manchester).*

Je dédie ce livre à la mémoire de Henry Rouanet avec qui j'ai approfondi mes connaissances en statistique, élaboré de nouvelles méthodes et pris mon bâton de pèlerin pour aller, de par le monde, diffuser ces méthodes géométriques, en particulier en relation avec la théorie sociologique de Pierre Bourdieu.

J'ai aussi une pensée émue pour mes petits-enfants Marco, Nathan et Émeline qui, durant les vacances d'été, m'ont vue passer de longues heures à écrire ce texte.

Enfin je remercie, pour sa bienveillante patience, Marie-Laure Davezac-Duhem, des Editions Dunod.

A Paris, le 28 Octobre 2013
Brigitte Le Roux

Chapitre 1
Prologue

Dans ce chapitre d'introduction, nous commentons des analyses de données effectuées sur deux exemples simples (§1.1). Puis nous précisons l'approche qui est développée dans l'ouvrage (§1.2) et présentons l'organisation des chapitres (§1.3). A la fin de ce chapitre, on trouvera une introduction au calcul matriciel, premier thème de la rubrique *bases mathématiques* (§M).

1.1 Exemples commentés

▷ **Exercice de mise en train : Portraits chinois**

> Voici une liste de six *hommes politiques* : Valéry Giscard d'Estaing, Michel Poniatowski, Jacques Chirac, Jean–Jacques Servan–Schreiber, François Mitterrand, Georges Marchais.
> Voici maintenant une liste de six *couleurs* : *blanc*, *noir*, *bleu*, *orange*, *jaune*, *vert*.
>
> Attribuez une couleur à chaque homme politique, chaque couleur ne peut être attribuée qu'à un seul homme politique.
>
> Maintenant, associez à chaque homme politique une *femme célèbre* : Brigitte Bardot, Mireille Mathieu, Jane Birkin, Michèle Morgan, Jackie Kennedy, Elizabeth II d'Angleterre.

Le lecteur aura reconnu une variante du jeu des *Portraits chinois* : si Giscard était une couleur, ce serait... etc. Les tableaux 1.1 et 1.2 (p. 2) présentent des données (fréquences exprimées en pourcentages) recueillies en 1975[1].

[1] *Source : Sondages* (1975), 3, 4, 31-47. *Référence* : Bourdieu (1979, p. 625-640) ; on trouvera dans ce texte les tableaux correspondant à dix autres questions (chapeaux, métiers, personnages de bandes dessinées, etc.).

	blanc	*noir*	*bleu*	*orange*	*jaune*	*vert*
Giscard d'Estaing	*35*	10	29	6	9	12
Poniatowski	16	22	14	16	18	13
Chirac	16	9	25	12	18	18
Servan–Schreiber	14	9	12	23	23	19
Mitterrand	13	10	13	23	18	*24*
Marchais	6	*40*	7	20	14	14

Tableau 1.1. Hommes politiques et *Couleurs*.

	Brigitte Bardot	*Mireille Mathieu*	*Jane Birkin*	*Michèle Morgan*	*Jackie Kennedy*	*Elizabeth d'Anglet.*
Giscard d'Estaing	14	7	10	26	15	27
Poniatowski	12	15	11	15	14	*32*
Chirac	21	14	15	15	21	15
Servan–Schreiber	24	10	19	8	*29*	11
Mitterrand	16	18	21	25	13	7
Marchais	13	*36*	24	11	8	8

Tableau 1.2. Hommes politiques et *Femmes célèbres*.

1.1.1 Des tableaux vers les nuages

L'examen des tableaux fait apparaître des attributions privilégiées correspondant aux cases dont le pourcentage est maximum en ligne et en colonne : *noir* pour Marchais, *blanc* pour Giscard ; *Mireille Mathieu* pour Marchais, la *reine Elizabeth* pour Poniatowski, etc.

Étude du tableau «Hommes politiques et Couleurs»

Le diagramme suivant, obtenu en effectuant l'*analyse des correspondances* (AC) du tableau 1.1, comporte douze points : six représentent les hommes politiques (•) et six représentent les couleurs (○).

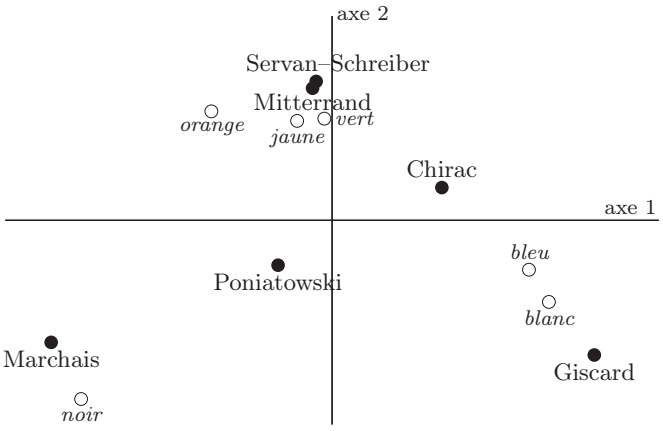


Figure 1.1. Représentation simultanée de l'analyse des correspondances du tableau 1.1 (Hommes politiques et Couleurs).

Ce diagramme est un premier exemple de données représentées sous forme de *nuages de points* avec, sur le même diagramme, deux nuages celui des hommes politiques et celui des couleurs.

Les proximités entre les points d'un même nuage correspondent à des similarités entre *profils* : les profils de Marchais et de Giscard sont très dissemblables, en ce sens que les couleurs les plus fréquentes pour Marchais sont les moins fréquentes pour Giscard et vice–versa ; en revanche, les profils de Mitterrand et de Servan–Schreiber sont très proches.

Examen à vue du diagramme. On voit que :

— les hommes politiques s'organisent autour de trois pôles : Marchais à gauche, Mitterrand et Servan–Schreiber au sommet, Giscard à droite ; le long d'une ligne à peu près triangulaire se succèdent Marchais, puis Mitterrand et Servan–Schreiber, puis Chirac, puis Giscard alors que Poniatowski occupe une position plutôt centrale ;

— les six couleurs s'organisent selon un schéma analogue : *noir* à gauche ; *orange*, *jaune* et *vert* au sommet ; *bleu* et *blanc* à droite ;

— le diagramme suggère les attributions suivantes : pour Giscard le *blanc*, pour Marchais le *noir* (attributions déjà mentionnées lors de l'examen du tableau), pour Mitterrand et Servan–Schreiber l'*orange*, le *jaune* et le *vert* ; il suggère que Chirac est entre *bleu* et *jaune–vert* ; quant à Poniatowski, on est tenté de lui attribuer «incolore» (répartition uniforme des couleurs).

Les points sont rapportés à deux axes qui sont les deux premiers *axes principaux* des nuages.

Interprétation des axes. L'axe 1 (horizontal) oppose Marchais (situé à gauche sur la figure 1.1) à Giscard (à droite), ainsi que *noir* (à gauche) à *blanc* et *bleu* (à droite) ; l'axe 2 (vertical) oppose Marchais et Giscard (en bas) à Mitterrand et Servan–Schreiber (en haut), ainsi que *noir* (en bas) à *orange*, *jaune* et *vert* (en haut).

L'analyse des correspondances de ce tableau est particulièrement simple, parce qu'elle conduit à des points qui se trouvent à peu près dans un *plan* ; le diagramme précédent fournit un bon résumé des données et suffit donc comme base pour l'interprétation.

Étude du tableau «Hommes politiques et Femmes célèbres»

Passons maintenant au tableau 1.2 (p. 2). L'analyse des correspondances conduit encore à deux nuages de six points qui sont à peu près dans un espace à trois dimensions avec donc trois axes principaux. Pour représenter graphiquement les données, on a effectué deux diagrammes en projetant les deux nuages d'une part sur le plan déterminé par les axes principaux 1 et 2 (premier *plan principal*), d'autre part sur la droite déterminée par le troisième axe principal.

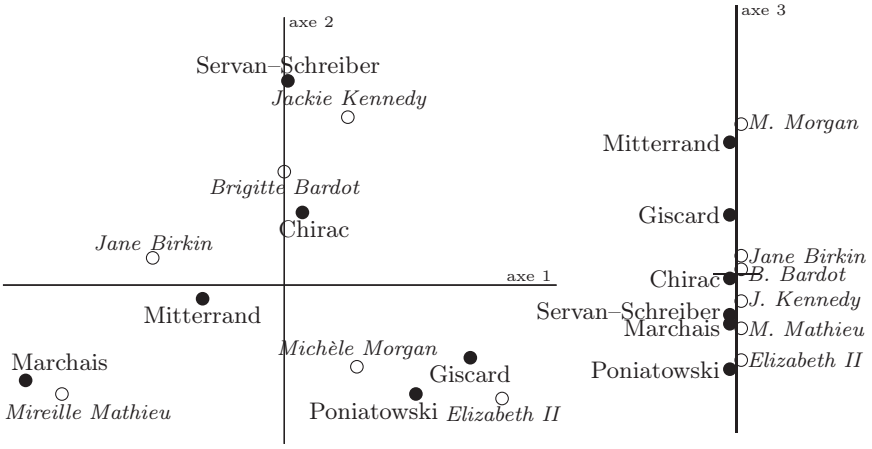


Figure 1.2. Diagramme simultané de l'AC du tableau 1.2 (plan 1-2 et axe 3).

Interprétation des axes. Pour les hommes politiques, le *premier axe* oppose Marchais à Poniatowski et Giscard ; le *deuxième axe* oppose Servan–Schreiber à Poniatowski et Marchais. Pour les femmes célèbres, on voit, dans le plan des axes 1 et 2, trois points extrêmes : Mireille Mathieu, Jackie Kennedy et Elizabeth II ; sur le deuxième axe, il y a opposition entre Brigitte Bardot, Jackie Kennedy et Jane Birkin d'une part, et les trois autres d'autre part. Le *troisième axe* est principalement celui de Michèle Morgan et de Mitterrand, tous deux extrêmes sur cet axe et éloignés des autres. L'examen du troisième axe permet d'affiner les interprétations tirées du plan des axes 1 et 2 ; il suggère un «rapprochement» entre Mitterrand et Michèle Morgan et montre également que Giscard et Poniatowski, proches dans le plan 1-2, s'opposent sur ce troisième axe : dans l'espace à trois dimensions, Giscard et Mitterrand se trouvent d'un côté du plan des axes 1 et 2, Poniatowski étant de l'autre.

1.1.2 Espace multidimensionnel

Ces exemples nous ont permis d'introduire l'objet fondamental des méthodes d'analyse géométrique des données (AGD) : un nuage de points dans un *espace multidimensionnel*. Avec l'exemple "couleurs", nous avons considéré des nuages plans (ou de dimension 2) ; avec celui des "femmes célèbres", des nuages de dimension 3. L'analyse d'autres tableaux conduirait à des nuages de dimension supérieure à 3.

La *dimension d'un nuage de points* est au plus égale au nombre de points moins un. Ainsi un nuage de deux points est sur une droite (unique si les deux points sont distincts) ; un nuage de trois points est dans un plan (unique si les trois points sont non–alignés) ; un nuage de quatre points est dans un espace de dimension 3 (unique si les quatre points sont non–coplanaires), etc. Les nuages des exemples précédents sont dans un espace de dimension $6 - 1 = 5$ au plus, on les a approchés par des nuages plans pour ceux issus du tableau 1.1 et par des nuages de dimension 3 pour ceux issus du tableau 1.2.

En pratique, on est presque toujours amené à étudier des nuages de dimension élevée que l'on cherche à réduire. Pour ce faire, on projette orthogonalement les nuages sur un sous–espace de dimension plus faible qui les contient *à peu près*. Ensuite, l'examen des données se fait sur des diagrammes de dimension 1 (droites) ou 2 (plans) qui conduisent à une interprétation synthétique des données.

> ◇ Les projections étant *orthogonales*, deux points éloignés en projection sont a fortiori éloignés dans l'espace, deux points proches dans l'espace le sont aussi en projection, alors que deux points proches en projection peuvent ne pas être proches dans l'espace. On gardera toujours cette propriété à l'esprit lors de l'interprétation à partir des nuages projetés.

1.1.3 Classification

Sur ces données, on peut aussi procéder à une *classification*. L'idée de classification est intuitive. Par exemple, chercher à classer les hommes politiques, c'est chercher à mettre ensemble ceux qui sont proches et à mettre dans des classes différentes ceux qui apparaissent éloignés. L'examen des nuages issus du tableau "couleurs" suggère de placer dans la même classe Mitterrand et Servan–Schreiber et dans des classes différentes Marchais et Giscard.

Les méthodes de classification sont souvent utilisées en complément des méthodes d'axes principaux. Nous présenterons celles dont

l'objet de base est un nuage euclidien. A titre d'exemple, la figure 1.3 présente le résultat d'une *classification ascendante hiérarchique* sous forme d'un *arbre* ; cet arbre représente une *hiérarchie de classes*, les classes correspondent aux nœuds de l'arbre. L'arbre, bien que construit de manière ascendante, s'interprète en descendant. En haut de l'arbre, on a une seule classe. En parcourant l'arbre en descendant, on obtient deux classes, l'une contenant Marchais et l'autre contenant les cinq autres hommes politiques ; celle–ci se subdivise en deux classes : celle de Giscard et celle des quatre autres, cette dernière se divisant à son tour en deux classes : celle de Chirac et Poniatowski et celle de Mitterrand et Servan–Schreiber. Ensuite, chacune de ces deux classes se divise en deux en commençant par celle de Chirac et Poniatowski. On arrive ainsi au bas de l'arbre avec six classes élémentaires (une par homme politique).

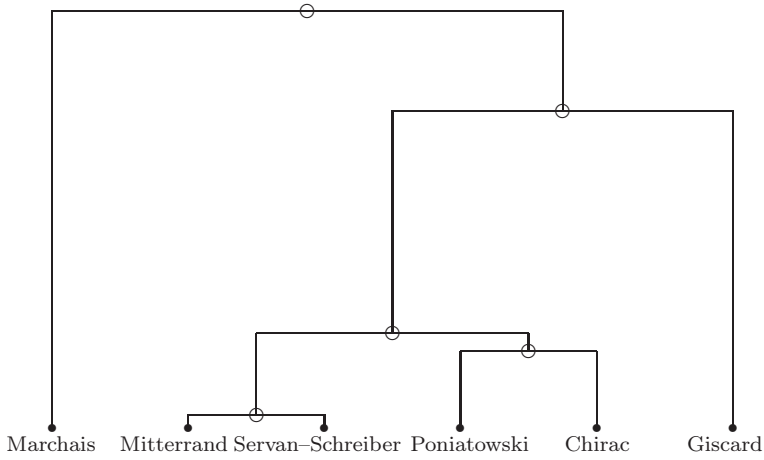


Figure 1.3. Arbre hiérarchique de la classification des hommes politiques (tab.1.1).

La classification hiérarchique d'un ensemble de six objets conduit à un système de six *partitions emboîtées*, depuis la partition grossière en une seule classe au sommet de l'arbre, jusqu'à la partition la plus fine en six classes à la base de l'arbre. On obtient ces partitions successives en parcourant l'arbre en descendant et en le coupant à divers niveaux. En coupant l'arbre près du sommet, on obtient la partition en deux classes (Marchais et les cinq autres), puis celle en trois classes (Marchais, Giscard et les quatre autres), etc. Si l'on recherche une «bonne partition», plutôt que de fixer a priori un nombre de classes, on prendra en compte les «chutes de niveau» : la plus importante

correspond ici à la partition en trois classes, ce qui est un argument en faveur de cette partition. Nous avons représenté les trois classes de cette partition dans le plan principal 1-2 (Figure 1.4).

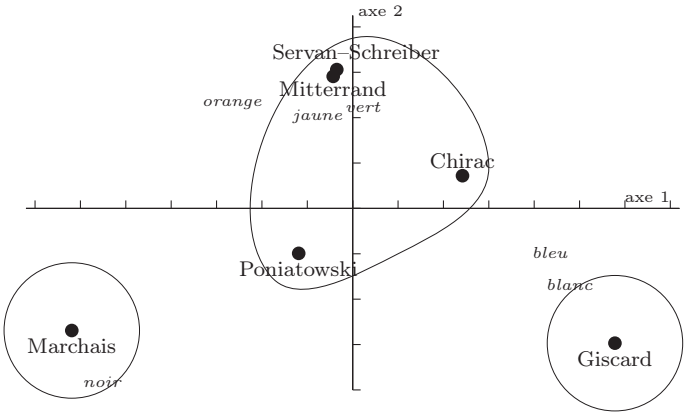


Figure 1.4. Diagramme de l'analyse des correspondances et les trois classes de la classification.

Dans ce texte introductif, nous avons anticipé[2] sur les chapitres à venir, en introduisant des termes techniques (axes principaux, arbre hiérarchique, etc.), qui seront explicités à partir du chapitre 5.

1.2 Analyse géométrique des données multidimensionnelles

Les *données multidimensionnelles* sont des données où les observations sont à valeurs dans plusieurs variables. Les procédures statistiques applicables à ces données «multivariées» constituent la statistique multidimensionnelle, appelée encore analyse des données multidimensionnelles pour marquer qu'on adopte la démarche consistant à aller *des données vers les modèles*.

1.2.1 Analyse géométrique des données

Devant le foisonnement des méthodes multidimensionnelles, la nécessité d'une approche unificatrice s'impose. Nous adopterons celle

[2]Dans l'exercice 7.2 (p. 215) du chapitre 7 [*Analyse des correspondances*], on trouvera les résultats et calculs statistiques qui étayent les interprétations présentées dans ce chapitre.

de l'*analyse géométrique des données* avec ses trois caractéristiques : géométrique, formelle et descriptive.

— *Approche descriptive.* Elle accorde aux procédures descriptives la première place. Les données multidimensionnelles vont parfois (littéralement) «dans tous les sens» ; il importe avant tout de les *résumer* : tel est l'objectif premier des méthodes que nous avons illustrées au §1.1 (analyse des correspondances et classification).

— *Approche géométrique.* La représentation des données sous forme de *nuages de points* est la phase cruciale de l'analyse de données ; l'interprétation des données se fait de manière privilégiée sur les nuages de points.

— *Approche formelle.* Elle consiste à prendre en compte les structures mathématiques sous–jacentes aux procédures statistiques. Pour interpréter conjointement plusieurs procédures appliquées aux mêmes données (par exemple, analyse des correspondances et classification), il faut prendre des structures en harmonie.

Les structures commandent les procédures !

◇ Cette approche est due, pour l'essentiel, à J-P. Benzécri qui l'a développée à propos de l'analyse des correspondances. En fait, elle permet d'organiser de façon cohérente l'ensemble des méthodes de la statistique multidimensionnelle.

1.2.2 Structures multidimensionnelles

En statistique élémentaire, il est de tradition de présenter les procédures en liaison avec les *structures d'échelles* : médiane et structure ordinale, moyenne et échelles d'intervalles, etc. Les structures géométriques et linéaires mises en œuvre en AGD, sont tout aussi éclairantes pour le choix des procédures, c'est pourquoi dans ce livre nous les mettrons en avant.

Les *structures géométriques* prolongent les structures familières de l'espace physique à trois dimensions : alignement et parallélisme (structures affines), distances et angles (structures euclidiennes). On peut les étayer sur l'intuition spatiale.

Les *structures linéaires*, plus abstraites, sont les structures profondes de la statistique. Nous les présenterons progressivement, en liaison avec les procédures statistiques. La *dualité* sera présentée au chapitre 2, avec la distinction entre les quantités «qui s'ajoutent» (les

mesures) et celles qui «se moyennent» (les variables). Ultérieurement, au chapitre 5 apparaîtra le *théorème spectral*, cœur de l'analyse géométrique des données.

1.2.3 Analyse descriptive et visée inductive

En analyse géométrique des données, la démarche en deux temps :

la description d'abord, l'inférence ensuite !

s'impose plus que jamais.

Analyse descriptive des données

Toutes les procédures présentées dans ce livre sont *descriptives*, au sens du critère opérationnel suivant : elles peuvent être effectuées à partir des distributions de fréquences. En d'autres termes, dire qu'*une statistique est descriptive*, c'est dire que sa valeur est inchangée si l'on multiplie tous les effectifs par un même nombre : la *taille des données n'intervient pas*. Une *procédure descriptive* est une procédure conduisant à une statistique descriptive.

Une analyse descriptive ne présuppose pas un cadre d'hypothèses techniques contraignant, mais pour qu'elle soit *utile*, elle doit satisfaire certaines exigences ; J-P. Benzécri, à propos de l'analyse des correspondances (mais la portée est générale), formule les deux suivantes :

— *Homogénéité* : «Toutes les grandeurs recensées dans le tableau sont des quantités de même nature» (Benzécri, 1973, tome 2, p. 21-22). Nous dirons qu'on doit disposer d'un *espace d'observables*.

— *Exhaustivité* : «Les marges du tableau représentent un inventaire complet d'un dossier réel dont le cadre n'est guère discutable». Mais comme en pratique, ajoute J-P. Benzécri, l'exhaustivité n'est souvent qu'approchée par échantillonnage, ce qui importe, c'est de «faire du réel une coupe bien choisie». Nous dirons que, à défaut d'exhaustivité, on a l'exigence de *représentativité*.

Analyse inductive des données

Lorsqu'on songe à prolonger des conclusions descriptives à une population plus vaste, on adopte une *visée inductive*. Si les exigences qui viennent d'être mentionnées sont satisfaites, on peut déjà, à partir des conclusions descriptives, apporter une première réponse à la visée inductive : les statistiques descriptives (fréquence, moyenne,

etc.) sont des *estimations* des paramètres correspondants de la population. On peut ensuite mettre en œuvre des *procédures inductives*, ou d'*inférence statistique*. Ces procédures sont communément utilisées en régression, moins en analyse en composantes principales, moins encore en analyse des correspondances ou en classification. Cette diversité est surtout due aux traditions : modèle aléatoire de régression chez les économistes, monographies descriptives chez les sociologues, combinaisons variées de procédures descriptives et inductives chez les psychologues (Reuchlin, 1990, chap. 3).

Pour donner sens aux procédures inductives, il faut les placer dans un *cadre d'interprétation* ; nous distinguerons les trois cadres suivants : combinatoire, fréquentiste et bayésien[3].

— Dans le cadre *combinatoire*, prolongement direct de la méthodologie descriptive, la probabilité est purement formelle ; c'est un simple calcul de proportions d'échantillons qui permet d'évaluer le *potentiel de généralisabilité* des données. Des procédures comme les *valeurs–tests* (cf. Lebart *et al.*, 2000) se rattachent à ce cadre.

— Le cadre *fréquentiste*, en introduisant des hypothèses (plus ou moins réalistes), permet de pratiquer les *tests de signification*. Ce cadre reste le plus usité mais ses limitations sont flagrantes, surtout en statistique multidimensionnelle, et il prête à l'illusion de significativité : on interprète indûment un effet significatif comme un effet important et un effet non significatif comme un effet négligeable.

— Le cadre *bayésien*, plus souple et qui restitue à la probabilité son interprétation naturelle (aller du connu vers l'inconnu), est mieux adapté à une véritable méthodologie d'*analyse inductive des données*.

Une présentation argumentée des procédures inductives multidimensionnelles entraînerait au delà du champ du présent ouvrage ; nous ne les aborderons pas dans ce livre.

1.3 Organisation de l'ouvrage

Les *préalables statistiques*[4] nécessaires à le lecture de ce livre sont les procédures descriptives élémentaires : moyenne, variance, diagramme de corrélation, corrélation, tableau de contingence.

[3]Voir Rouanet *et al.* (1990) ; Rouanet *et al.* (1998) ; Le Roux et Rouanet (2004, chap. 8 et 9).

[4]Voir par exemple Rouanet *et al.* (1987), Rouanet et Le Roux (1995), Lebaron (2006).

Organisation des chapitres

Après ce chapitre introductif viennent deux chapitres consacrés aux notions de base. Au chapitre 2 [*Mesures et variables*], on introduit les mesures et les variables avec le concept central de *dualité*. Puis, au chapitre 3 [*Protocoles multivariés*], on aborde les protocoles multivariés, avec leur représentation sous forme de nuages de points dans un espace multidimensionnel. Les procédures statistiques seront étudiées d'une part dans l'espace d'observables (espace géométrique), d'autre part dans l'espace des variables (espace vectoriel), les deux études se complétant l'une l'autre.

Deux chapitres sont ensuite consacrés aux méthodes de la statistique multidimensionnelle. Au chapitre 4 [*Régression linéaire*], on présente brièvement la régression linéaire, qui met en jeu les seules structures affines de l'espace d'observables (points moyens, écarts). On poursuit au chapitre 5 [*Nuage euclidien*] en introduisant la structure euclidienne (distances, angles) et en prenant pour objet d'étude un nuage euclidien.

Les trois chapitres suivants sont consacrés aux principales procédures d'*analyse géométrique des données*, qui à partir d'un tableau de données conduisent à construire et étudier un nuage euclidien. Les procédures diffèrent selon le type de données : analyse en composantes principales pour un tableau de variables (chap. 6), analyse des correspondances pour un tableau de contingence (chap. 7), analyse des correspondances multiples pour un tableau de variables catégorisées (questionnaire)(chap. 8).

Ensuite, au chapitre 9 [*Analyse des données structurées*], on in-

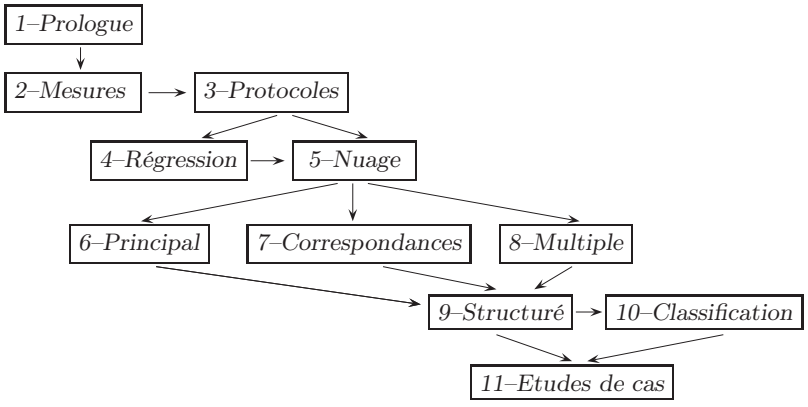


Figure 1.5. Organisation des chapitres

troduit les méthodes d'analyse des données structurées appliquées à un nuage euclidien en tenant compte des facteurs qui ont présidé au recueil des données. Au chapitre 10 [*Classification*], on présente les méthodes de classification d'un nuage euclidien avec, en particulier, la classification ascendante hiérarchique suivant la variance.

Enfin, le dernier chapitre présente des études cas qui mettent en œuvre les procédures sur des *données réelles* et concrétisent la finalité de l'ouvrage en détaillant la démarche méthodologique.

A la *fin de ce livre*, le lecteur aura acquis les bases théoriques et méthodologiques des principales méthodes d'analyse géométrique des données. Confronté à des données, il pourra construire les tableaux à analyser, disposer éléments actifs et supplémentaires, passer des tableaux aux nuages, interpréter un diagramme, etc. Confronté à des travaux faisant état d'analyses de données, il pourra en prendre connaissance en faisant preuve de sens critique.

Présentation matérielle des chapitres. Chaque chapitre comporte un exposé de style «*cours magistral*», illustré par des exemples volontairement schématiques. Le texte principal est écrit en *gros caractères* ; les démonstrations (qui commencent par «*Preuve.*») ainsi que les remarques (qui sont, en général, précédées du signe ⋄) sont en *moyens caractères* (cf. *Mode de lecture*, p. VI), elles peuvent être sautées lors d'une première lecture.

Les *exercices* sont placés sous l'une des trois rubriques suivantes : *application du cours* (pour s'assurer de l'acquisition des techniques), *théorique* (pour approfondir les notions), *analyse de données* (pour ébaucher la méthodologie sur des exemples réels) ; ils sont accompagnés de solutions et commentaires. Parmi les exemples traités, certains sont historiques («Scotland Yard»...), ils fourniront l'occasion d'évoquer des pionniers, de jeter des ponts entre des méthodes... Les commentaires des exercices font partie intégrante du cours, au même titre que les développements secondaires en moyens caractères.

Les préalables mathématiques consistent en des connaissances sur l'algèbre linéaire qui figurent dans le bagage de tout lecteur mathématicien. Afin de permettre une *lecture autonome* de ce livre, nous avons, sans viser à l'exhaustivité, inclus à la fin des chapitres 1, 2, 3 et 5 quelques jalons qui fixeront vocabulaire et notations. Ces jalons, rassemblés sous la rubrique *Bases mathématiques* (§M), sont en moyens caractères.

◇ L'étude systématique des exercices d'analyse des données et du chapitre 11 [*Études de cas*] constitue à elle seule un véritable *cours d'analyse des données par la pratique*.

Aspects informatiques. La mise en œuvre des méthodes d'analyse géométrique des données passe par l'informatique[5].

Pour les méthodes d'AGD, il paraît naturel d'utiliser les logiciels français, qui ont accompagné leurs développements théoriques et appliqués. Pour les études de cas présentées dans ce livre, nous avons utilisé le logiciel SPAD[6].

Bases mathématiques

M-1 Calcul matriciel

Les *bases mathématiques* de ce chapitre sont consacrées au *calcul matriciel*. En analyse géométrique des données, les calculs statistiques sont souvent des opérations effectuées «en bloc» sur des tableaux numériques à double entrée. Le calcul matriciel permet d'exprimer ces calculs de façon compacte, les tableaux numériques sont alors représentés par des *matrices* et traités par les opérations du calcul matriciel.

Dans ce livre, nous utilisons le calcul matriciel, non comme un outil de démonstration, mais comme une *sténographie*, pour résumer les opérations effectuées sur les tableaux numériques. Dans ce but, il suffit de connaître les trois opérations matricielles de base : transposition, addition, multiplication, que nous présentons dans cette section[7]. Les *conventions d'écriture* que nous suivons faciliteront la lecture des textes statistiques rédigés dans la mouvance anglo–saxonne : matrices notées par des lettres grasses, et pour la transposition le symbole «^{t}» précédant la matrice (exemple ${}^t\mathbf{X}$).

M-1.1 Matrices

Dans les écritures matricielles, nous spécifions les ensembles indexants (I et J). Une matrice à I lignes et J colonnes est dite matrice de type $I \times J$, ou en bref matrice $I \times J$.

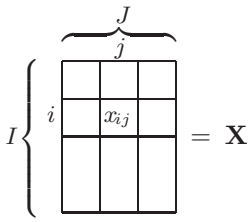
[5] Pour les exercices d'analyse des données, on donne le tableau des données de base, ce qui permettra au lecteur de retrouver les résultats des analyses. Les données de base des études de cas sont disponibles sur ma page personnelle.

[6] Locigiel distribué par Coheris-Spad : www.spad.eu

[7] Nous suivons ici la présentation du calcul matriciel faite par G. Th. Guilbaud à l'E.H.E.S.S. (http ://www.ehess.fr/revue-msh/video_gb.php ?auteur=1095), voir la revue Mathématiques et sciences Humaines.

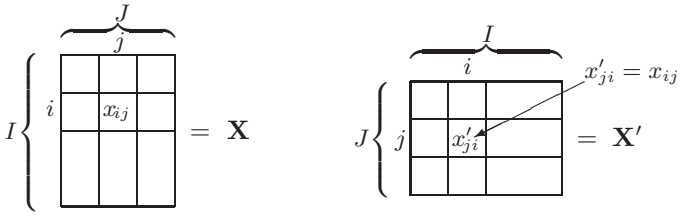
Une famille numérique $(x_{ij})_{i\in I, j\in J}$ est représentée par une matrice de type $I\times J$, notée $\mathbf{X} = [x_{ij}]$, l'ensemble I indexant les lignes, et l'ensemble J indexant les colonnes.

Dans les diagrammes, nous prendrons, en général, $I = 4$ et $J = 3$.



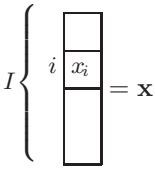
M-1.2 Transposition d'une matrice

Transposer une matrice consiste à échanger les lignes et les colonnes : la transposée de la matrice $\mathbf{X} = [x_{ij}]$ de type $I \times J$ (à I lignes et J colonnes) est la matrice $J \times I$ (à J lignes et I colonnes), notée ${}^t\mathbf{X}$, de terme général $x'_{ji} = x_{ij}$.

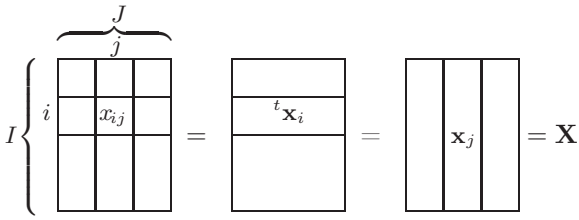


M-1.3 Vecteur–colonne, vecteur–ligne

Un *vecteur–colonne* I, ou I–colonne, est une matrice à une colonne et I lignes notée par une minuscule grasse. On définit de même une *vecteur–ligne*.
Une famille numérique à simple indice peut s'écrire matriciellement soit comme une colonne soit comme une ligne, le choix étant arbitraire.



Une matrice $\mathbf{X} = [x_{ij}]$ peut ainsi être regardée comme une famille de lignes ${}^t\mathbf{x}_i$, ou comme une famille de colonnes $\mathbf{x}_j$.

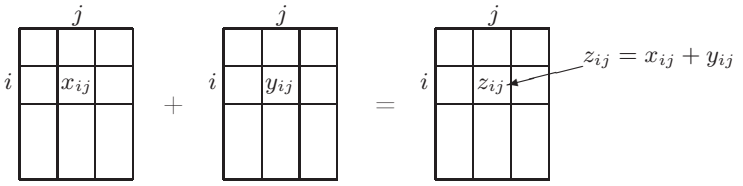


En règle générale, nous privilégions l'écriture selon une colonne $\mathbf{x}$, la transposée est alors la ligne notée ${}^t\mathbf{x}$.

Un nombre peut être représenté par une matrice à une ligne et une colonne qui est notée par une lettre ordinaire comme le nombre lui-même.

M-1.4 Addition matricielle

L'addition matricielle permet d'écrire la somme terme à terme de deux familles numériques à double indice. Si $\mathbf{X} = [x_{ij}]$ et $\mathbf{Y} = [y_{ij}]$ sont deux matrices $I \times J$, la matrice somme $\mathbf{Z}$ est la matrice $I \times J$ de terme général z_{ij} telle que $z_{ij} = x_{ij} + y_{ij}$. D'où l'écriture de l'addition matricielle : $\mathbf{X} + \mathbf{Y} = \mathbf{Z}$.



Propriété : la transposée de la somme de matrices est la somme des tranposées : ${}^t\mathbf{Z} = {}^t\mathbf{X} + {}^t\mathbf{Y}$.

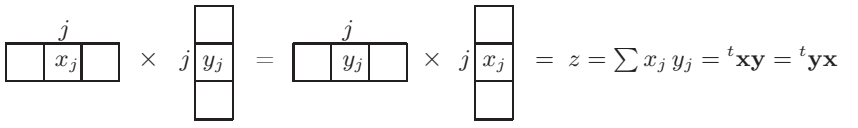
On appelle *matrice nulle*, notée $\mathbf{0}$, une matrice dont tous les termes sont nuls. Si $\mathbf{0}$ désigne la matrice $I \times J$ nulle, on a : $\forall\, \mathbf{X}\ :\ \mathbf{X} + \mathbf{0} = \mathbf{0} + \mathbf{X} = \mathbf{X}$.

Soit a un nombre entier positif, l'addition de a matrices $\mathbf{X} = [x_{ij}]$ donne la matrice de terme général $a\, x_{ij}$, ce qui conduit à définir le produit (commutatif) d'une matrice par un nombre réel a, que l'on écrit $\mathbf{X}a = a\mathbf{X} = [a\, x_{ij}]$.

M-1.5 Multiplication matricielle

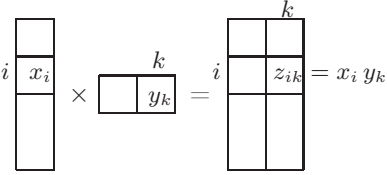
Le produit $\mathbf{XY}$ de deux matrices $\mathbf{X}$ et$\mathbf{Y}$ est défini si et seulement si l'ensemble indexant les colonnes de $\mathbf{X}$ coïncide avec celui indexant les lignes de $\mathbf{Y}$.

Produit d'une ligne par une colonne. La multiplication d'un vecteur–ligne par un vecteur–colonne permet d'écrire la somme des produits de deux familles numériques définies sur un même ensemble d'indices. Si $\mathbf{x}$ et $\mathbf{y}$ sont deux I–colonnes, le produit de la I–ligne ${}^t\mathbf{x}$ (transposée de $\mathbf{x}$) par la I–colonne $\mathbf{y}$, noté ${}^t\mathbf{xy}$, est le nombre $z = \sum x_i y_i$; il est égal au produit de la ligne ${}^t\mathbf{y}$ par la colonne $\mathbf{x}$.



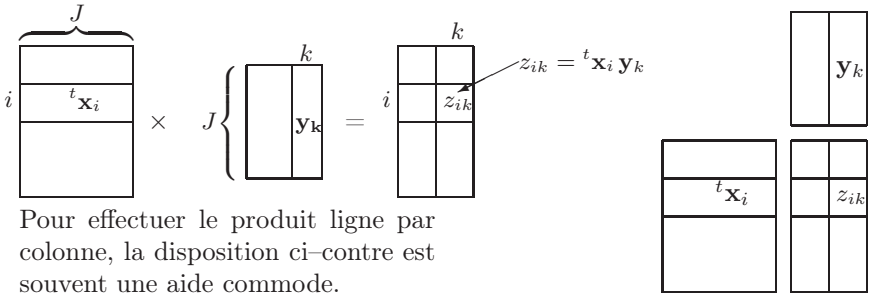
Produit d'une colonne par une ligne. La multiplication d'un vecteur–colonne par une vecteur–ligne permet d'écrire les produits terme à terme des deux familles numériques $x_I = (x_i)_{i \in I}$ et $y_K = (y_k)_{k \in K}$.

Soient $\mathbf{x}$ est une I–colonne et $\mathbf{y}$ une K–colonne, le produit de la colonne $\mathbf{x}$ par la ligne ${}^t\mathbf{y}$ est la matrice $\mathbf{Z} = \mathbf{x}\ {}^t\mathbf{y}$ de type $I \times K$ ayant pour terme général $z_{ik} = x_i\, y_k$.



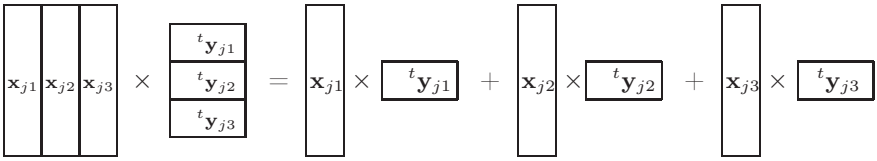
Produit de deux matrices : cas général. Si $\mathbf{X} = [x_{ij}]$ est une matrice $I \times J$ et $\mathbf{Y} = [y_{jk}]$ une matrice $J \times K$, le produit $\mathbf{X}\,\mathbf{Y}$ est la matrice $\mathbf{Z} = [z_{ik}]$ de type $I \times K$ dont le terme général est $z_{ik} = \sum_{j \in J} x_{ij} y_{jk}$. Matriciellement, ce produit peut s'écrire de deux façons en termes des lignes et colonnes des matrices $\mathbf{X}$ et $\mathbf{Y}$.

— *Produits lignes par colonnes* : la case z_{ik} du produit $\mathbf{Z} = \mathbf{XY}$ est le produit de J–ligne ${}^t\mathbf{x}_i$ ($i^{\text{ième}}$ ligne de $\mathbf{X}$) par la J–colonne $\mathbf{y}_k$ ($k^{\text{ième}}$ colonne de $\mathbf{Y}$) :



Pour effectuer le produit ligne par colonne, la disposition ci–contre est souvent une aide commode.

— *Produits colonnes par lignes* : si $\mathbf{Z}_j = \mathbf{x}_j\,{}^t\mathbf{y}_j$ est le produit de la I–colonne $\mathbf{x}_j$ ($j^{\text{ième}}$ colonne de $\mathbf{X}$) par la K–ligne ${}^t\mathbf{y}_j$ ($j^{\text{ième}}$ ligne de $\mathbf{Y}$), la matrice produit $\mathbf{Z} = \mathbf{XY}$ est la somme des J matrices $\mathbf{Z}_j$ de type $I \times K$.



($\mathbf{XY} = \sum \mathbf{x}_j\,{}^t\mathbf{y}_j$: diagramme pour $J = 3$)

Cette expression du produit matriciel fait apparaître qu'un produit de matrices est également une somme de matrices de rang 1.

Propriétés du produit matriciel.

— La *transposée du produit* de deux matrices est le produit de leurs transposées : ${}^t(\mathbf{X}\,\mathbf{Y}) = {}^t\mathbf{Y}\,{}^t\mathbf{X}$.

— Le produit matriciel est *associatif* : si les produits $\mathbf{XY}$ et $\mathbf{YZ}$ sont définis, on a : $(\mathbf{X}\,\mathbf{Y})\mathbf{Z} = \mathbf{X}(\mathbf{Y}\,\mathbf{Z})$, ce que l'on écrit $\mathbf{X}\,\mathbf{Y}\,\mathbf{Z}$.

— Le produit matriciel est *distributif* par rapport à l'addition :

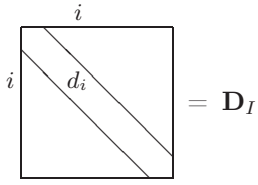
$$(\mathbf{X} + \mathbf{Y})\mathbf{Z} = \mathbf{X}\,\mathbf{Z} + \mathbf{Y}\,\mathbf{Z}$$

—Le produit *n'est pas commutatif* (en général). Si $\mathbf{X}$ est de type $I \times J$ et $\mathbf{Y}$ de type $J \times K$, le produit $\mathbf{Y}\,\mathbf{X}$ n'est pas défini si $I \neq K$. Les produits $\mathbf{X}\,{}^t\mathbf{X}$ et ${}^t\mathbf{X}\,\mathbf{X}$ existent toujours, le premier étant de type $I \times I$ et le second de type $J \times J$.

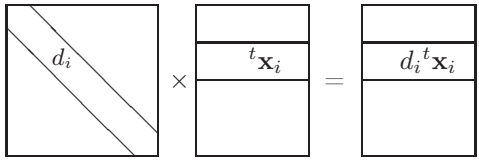
M-1.6 Matrices particulières

Lorsque les ensembles indexant les lignes et les colonnes d'une matrice sont en bijection naturelle avec un même ensemble, on dit que la matrice est une *matrice carrée*. Si $\mathbf{A}$ est une matrice carrée $I \times I$, les termes diagonaux sont les I termes $(a_{ii})_{i \in I}$; leur somme est appelée *trace de la matrice*, notée tr $\mathbf{A}$, avec tr $\mathbf{A} = \sum a_{ii}$.

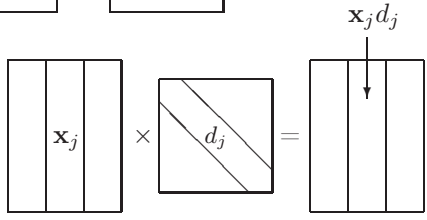
On dit qu'une matrice carrée $\mathbf{Q}$ est *symétrique* si elle est égale à sa transposée : $\forall\, i, i' \ :\ q_{ii'} = q_{i'i}$.
Une matrice carrée $\mathbf{D}$ dont les termes non–diagonaux sont nuls est appelée *matrice diagonale*. Une matrice diagonale fournit une nouvelle écriture matricielle d'une famille numérique à un indice $(d_i)_{i \in I}$ et sera notée $\mathbf{D}_I$.



Prémultiplier une matrice $\mathbf{X} = [x_{ij}]$ de type $I \times J$ par la matrice diagonale $\mathbf{D}_I$ c'est multiplier chaque ligne de la matrice $\mathbf{X}$ par le nombre d_i ; on obtient une matrice $I \times J$ de terme général $d_i\, x_{ij}$.



Post–multiplier une matrice $\mathbf{X} = [x_{ij}]$ par une matrice $\mathbf{D}_J$ diagonale c'est multiplier chaque colonne de $\mathbf{X}$ par le nombre d_j ; on obtient une matrice $I \times J$ de terme général $x_{ij}\, d_j$.



Une matrice diagonale dont les termes diagonaux sont égaux est appelée *matrice scalaire*. Les matrices scalaires commutent avec toute matrice pour laquelle les deux produits sont définis. Enfin, la *matrice–unité* $\mathbf{I}_I$ est une matrice scalaire de type $I \times I$ dont les termes diagonaux sont égaux à 1. Quelle que soit la matrice $\mathbf{X}$ de type $I \times J$, on a : $\mathbf{I}_I\, \mathbf{X} = \mathbf{X}\, \mathbf{I}_J = \mathbf{X}$.

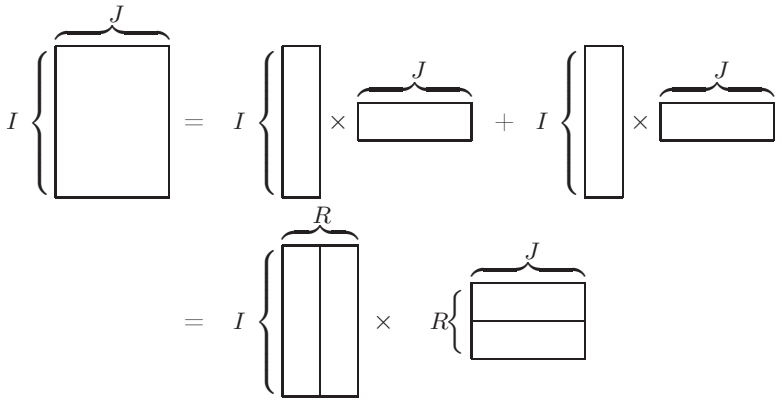
Soit $\mathbf{X}$ une matrice carrée, s'il existe une matrice carrée $\mathbf{Y}$ telle que $\mathbf{X}\, \mathbf{Y} = \mathbf{Y}\, \mathbf{X} = \mathbf{I}$, la matrice $\mathbf{Y}$ est appelée *matrice inverse* de $\mathbf{X}$ et est notée $\mathbf{X}^{-1}$ (on a aussi $\mathbf{X} = \mathbf{Y}^{-1}$).

Une matrice diagonale $\mathbf{D}_I$ de termes diagonaux d_i non–nuls admet une matrice inverse diagonale $\mathbf{D}_I^{-1}$, dont les termes diagonaux sont égaux à $1/d_i$.

M-1.7 Rang d'une matrice

Une matrice non–nulle qui peut s'exprimer comme le produit d'une colonne par une ligne est dite *de rang 1*. Une matrice de rang 1 est constituée de lignes proportionnelles entre elles et de colonnes proportionnelles entre elles.

Si une matrice non–nulle n'est pas de rang 1, mais qu'elle peut s'écrire comme la somme de deux matrices de rang 1, on dit qu'elle est de rang 2, etc. D'où, par récurrence, la notion générale : si R est un entier positif (avec $R \geq 2$), une matrice est dite de rang R si, d'une part elle ne peut pas s'exprimer comme somme de $R-1$ matrices de rang 1, et si d'autre part elle peut s'exprimer comme somme de R matrices de rang 1, autrement dit comme produit d'une matrice à R colonnes par une matrice à R lignes. Par exemple, une matrice de rang 2 peut s'écrire comme la somme de deux matrices de rang 1 ou comme le produit d'une matrice à 2 colonnes par une matrice à 2 lignes.



Si une matrice est de rang $R > 1$, sa décomposition en R matrices de rang 1 n'est pas unique. Par ailleurs, on peut chercher une matrice de rang 1 qui soit, dans un sens à préciser, la meilleure approximation de la matrice de rang R, ce qui conduit à la *décomposition en valeurs singulières d'une matrice*, procédure centrale de l'analyse géométrique des données, cf. *Bases math.* §M-5 du chapitre 5 [*Nuage euclidien*], p. 136.

Chapitre 2
Mesures et variables

Il n'y a de science que du mesurable...
Lord Kelvin

Dans ce chapitre, nous introduisons les notions à la base des méthodes d'analyse des données multidimensionnelles, autour du concept central de dualité entre mesures et variables.

Nous présentons d'abord les notions de mesure et de variable (avec la notation de dualité) et celle de densité d'une mesure par rapport à une variable (§2.1), puis nous relions la notion de variable à celle d'échelle numérique (§2.2). Ensuite nous abordons la formalisation linéaire, dans le cadre des espaces vectoriels (§2.3), puis euclidiens, avec la méthodologie des moindres carrés et présentons les formulations matricielles (§2.4). Les *Bases mathématiques* de ce chapitre (§M-2) portent sur l'algèbre linéaire, principalement euclidienne ; on pourra les consulter avant de lire les paragraphes 2.3 et 2.4.

2.1 Premières notions

▷ **Exercice de mise en train : exemple Bénélux**

Dans le tableau ci–contre, on donne, pour chaque pays de l'Europe des Six, dans les années soixante, sa population (en milliers d'habitants), sa superficie (en milliers de kilomètres carrés) et sa densité de population au km^2.

	Pop.	Sup.	Densité
France	48 500	551	88
Italie	50 700	301	168
All.Fédérale	57 600	248	232
Pays–Bas	12 200	34	359
Belgique	9 300	30	310
Luxembourg	325	3	108

Exprimer la population, la superficie et la densité de population du *Bénélux* (regroupement des Pays–Bas, de la Belgique, et du Luxembourg) à partir des populations, superficies, densités des pays regroupés.

La population du Bénélux est la *somme* des populations des trois pays : $12\,200 + 9\,300 + 325 = 21\,825$ milliers d'habitants. La superficie est la *somme* des superficies : $34 + 30 + 3 = 67$ milliers de km^2. On en déduit la densité de population du Bénélux : $21\,825/67 = 326$ habitants au km^2. Pour exprimer la densité du Bénélux à partir des densités des pays regroupés, on écrit chaque population comme produit de la densité par la superficie, donc $12\,200 = 34 \times 359$ (aux arrondis près) pour les pays-Bas, etc. D'où $326 = \frac{(34\times359)+(30\times310)+(3\times108)}{34+30+3}$. La densité du regroupement est la *moyenne pondérée* par les superficies des densités des pays regroupés. On peut appliquer ces procédures à n'importe quel regroupement des six pays : les populations et les superficies s'ajoutent, les densités se moyennent.

L'exemple *Bénélux* illustre une situation courante en Statistique, dans laquelle les données se présentent sous la forme suivante, on a :

— un ensemble fini[1] (non–vide) I, qu'on appelle *support*, sur lequel on effectue des regroupements ; un regroupement de I étant défini par un sous–ensemble (non–vide) de I (éventuellement I lui–même) ;

— des *fonctions numériques* sur I (applications de I dans $\mathbb{R}$, ensemble des nombres réels), qui, dans le regroupement, se dérivent soit par *sommation*, soit par *moyennage*[2].

Nous étudierons maintenant ces deux types de fonctions numériques.

2.1.1 Mesures sur un support

Une fonction numérique sur I qui se *dérive par sommation* est appelée *mesure sur* I et notée, selon la notation indicielle des fonctions, avec *indices en bas* : $u_I = (u_i)_{i\in I}$.

Les nombres u_i sont appelés *masses ponctuelles* (ou masses, ou coefficients) de la mesure u_I. La somme[3] $\sum u_i$ est la *masse totale* de u_I ; si I' est une partie de I, la masse de u_I sur I' est $\sum\limits_{i\in I'} u_i$.

[1] *Notation* : le cardinal (nombre d'éléments) d'un ensemble fini I est souvent noté $/I/$ ou card I. Pour simplifier nous le noterons comme l'ensemble lui–même, c'est–à–dire I.

[2] Les dérivations dont il sera question dans ce livre sont des *dérivations statistiques* ; elles n'ont rien à voir avec la dérivation des fonctions au sens de l'analyse mathématique.

[3] Pour alléger les écritures, dans les cas où il n'y aura aucune ambiguïté, on ne précisera pas l'indice de sommation : ainsi $\sum\limits_{i\in I} u_i$ sera simplement écrit $\sum u_i$.

Dans la suite, nous désignerons souvent une partie de I par $I{<}c{>}$ (lire «I dans c») en utilisant la *notation de l'emboîtement* (cf. commentaire p. 36) ; on parlera alors de la classe c et on notera u_c la masse de la classe c, avec $u_c = \sum_{i \in I<c>} u_i$.

Mesure sur I :

$u_I : \begin{array}{l} I \to \mathbb{R} \\ i \mapsto u_i \end{array}$ est une mesure sur $I \iff I{<}c{>} \mapsto \sum_{i \in I<c>} u_i$

Exemple Bénélux. I est l'ensemble des six pays : $i1$ désigne la France, $i2$ l'Italie, …, $i6$ le Luxembourg. Notons u_i la population du pays i, ϖ_i sa superficie ; les fonctions $u_I : i \mapsto u_i$ et $\varpi_I : i \mapsto \varpi_i$ sont des mesures sur le support I.

◇ Si à toute partie non–vide de I on associe sa masse, et si à la partie vide on associe le nombre 0, on *prolonge* la fonction numérique sur I à l'ensemble des parties de I. Une mesure est une fonction numérique qui se *prolonge par sommation* ; cette propriété rattache la notion de mesure sur un ensemble fini à celle classique en mathématiques de mesure sur un espace mesurable (cf. exercice 2.4, p. 37).

Mesure de Dirac. On appelle *mesure ponctuelle* une mesure dont toutes les masses ponctuelles sauf une sont nulles. Une mesure ponctuelle de masse égale à 1 en i est appelée *mesure de Dirac* de i (appellation classique en théorie de la mesure) et est notée $\delta_I^i = (\delta_{i'}^i)_{i' \in I}$, avec $\delta_i^i = 1$ et $\delta_{i'}^i = 0$ pour $i' \neq i$.

Pour un ensemble I, on a I mesures de Dirac (cf. tableau avec $I = 6$), et toute mesure u_I s'écrit de manière unique comme somme pondérée des mesures de Dirac : $u_I = \sum u_i \, \delta_I^i$.

	δ_I^{i1}	δ_I^{i2}	δ_I^{i3}	δ_I^{i4}	δ_I^{i5}	δ_I^{i6}
$i1$	1	0	0	0	0	0
$i2$	0	1	0	0	0	0
$i3$	0	0	1	0	0	0
$i4$	0	0	0	1	0	0
$i5$	0	0	0	0	1	0
$i6$	0	0	0	0	0	1

La mesure sur I dont toutes les masses ponctuelles sont nulles est appelée *mesure nulle* et notée 0_I.

Contraste. On appelle *contraste* une mesure de masse totale nulle :
u_I est un contraste sur $I \iff \sum u_i = 0$.

Pondération. On appelle pondération une mesure dont les masses sont strictement positives, les masses sont alors appelées *poids*.

Une mesure dont toutes les masses ponctuelles sont égales est dite *uniforme* ; si toutes les masses sont égales à 1, on dit mesure *élémentaire* ou *de dénombrement* puisque la masse de la partie I' de I est égale à son cardinal, on la note $1_I = (1)_{i \in I}$.

Parmi les mesures sur un support I, on privilégie une *pondération*, que l'on note $\varpi_I = (\varpi_i)_{i\in I}$ (ϖ est une variante de la lettre π, qu'on pourra lire "omega") : I muni de cette pondération est appelé *support pondéré* et noté (I, ϖ_I). Par la suite, tout support est pondéré, par défaut, on le munit de la pondération élémentaire.

2.1.2 Variables sur un support pondéré

En statistique, une variable est une application à valeurs dans un ensemble d'observables. En *statistique linéaire*, deux types de variables interviendront principalement : les *variables numériques* (à valeurs dans un ensemble de nombres), et les *variables catégorisées* (à valeurs dans un ensemble fini). Les variables numériques sont privilégiées ; sauf mention contraire, *variable* voudra dire «variable numérique».

Une fonction numérique sur le support pondéré (I, ϖ_I) qui se *dérive par moyennage* est appelée *variable sur* I, et notée avec *indices en haut* : $x^I = (x^i)_{i\in I}$. Les nombres x^i sont appelés *valeurs* de la variable. La ϖ_I*–moyenne* de la variables x^I est égale $\sum \varpi_i x^i/\varpi$ (en posant $\varpi = \sum \varpi_i$) et notée Moy x^I ou $\overline{x}$. Au regroupement I<c>, on associe la moyenne sur I<c> selon la pondération ϖ_I, ou ϖ_I–moyenne sur I<c>, notée x^c, et égale à $(\sum_{i\in I<c>} \varpi_i x^i)/\varpi_c$ (avec $\varpi_c = \sum_{i\in I<c>} \varpi_i$).

Variable sur I :

$x^I : I \to \mathbb{R}$, $i \mapsto x^i$ est une *variable* sur $I \iff I<c> \mapsto x^c = \sum_{i\in I<c>} \frac{\varpi_i x^i}{\varpi_c}$

Contrairement à la dérivation par sommation, la dérivation par moyennage dépend de la pondération qui n'intervient qu'à un facteur près : si les poids sont remplacés par des poids proportionnels, la moyenne reste la même.

◇ Dans l'exemple *Bénélux*, la fonction $x^I : i \mapsto x^i$ qui, au pays i, associe sa densité de population, est une variable puisque la densité de tout regroupement de pays est la moyenne des densités de ces pays pondérées par les superficies.
Notation de dualité. La notation avec *indices bas* pour "ce qui s'ajoute" et *indices hauts* pour "ce qui se moyenne" reflète la dualité entre mesures et variables ; elle sera toujours utilisée dans ce livre[4].

[4]La notation fut introduite en analyse des correspondances par Benzécri (1966), voir aussi Benzécri (1973, p. 60-63). On trouve aussi une utilisation des indices hauts pour noter le conditionnement dans Neveu (1964).