Chapitre 14

Statistique multivariée et géographie

L'analyse des données a plus d'un siècle d'existence. L'**analyse factorielle** consiste à transformer un tableau en une représentation graphique. L'analyse canonique fut développée par C. Spearman [Spearman, 1904] et K. Pearson au début du XX^{e} siècle. L'approche de Charles Spearman propose de comparer une matrice de données $\mathbf{T}$ avec le cas d'indépendance $\mathbf{T_0}$. Ainsi, il établit une égalité avec la meilleure approximation $\mathbf{T_1}$ et le reste des valeurs $\mathbf{T_2}$.

Charles Spearman (1863-1945)

$$\mathbf{T} - \mathbf{T_0} = \mathbf{T_1} + \mathbf{T_2} \tag{14.1}$$

Lindépendance est définie par une premier valeur propre triviale. La représentation graphique de Charles Spearman consiste à confronter l'approximation et le reste.

Vers 1930, H. Hotelling donna les bases de l'analyse en composantes principales [Hotelling, 1933]. Cela étant, ce ne fut que vers 1973 que J.-P. Benzécri proposa et étudia l'analyse des correspondances vue précédemment [Benzécri, 1973a] [Benzécri, 1973b].

Harold Hotelling (1895-1973)

Jean-Paul Benzécri (1932-2019)

Contrairement à la démarche utilisée en statistique inférentielle, on ne cherche pas à induire des lois valables pour la population entière à partir des résultats obtenus sur les individus observés. L'analyse des données s'apparente à de la statistique descriptive.

De fait, on retrouve ici les trois grandes méthodes de l'analyse des données :

1. les méthodes descriptives, dont l'objectif est de décrire et de résumer les informations obtenues ;
2. les méthodes explicatives (ou méthode d'analyse de dépendance), dont l'objectif est d'expliquer un phénomène, c'est-à-dire trouver un lieu, fonctionnel ou non, entre une ou plusieurs variables expliquées et une ou plusieurs variables explicatives ;

3. les méthodes de prévision, dont l'objectif est d'analyser et de résumer les informations obtenues.

Deux position scientifiques alimentent les analyses multivariées :

1. comprendre la production des résultats ;
2. comprendre la manière d'utiliser la méthode, c'est-à-dire admettre une « boîte noire ».

14.1 Méthodes descriptives

14.1.1 Méthodes de classification

Les méthodes de classification ont pour objectif de regrouper des individus, décrits par un certain nombre de variables, ou de caractères, en un nombre restreint de classes de sorte que :

1. les individus appartenant à une même classe soient le plus semblable possible ;
2. les classes soient bien séparées.

La plus utilisée est la **classification ascendante hiérarchique** (C.A.H.), qui est généralement couplée avec une analyse factorielle.

14.1.2 Analyse factorielle en composantes principales

L'analyse en composante principale (A.C.P.), due à K. Pearson et H. Hotelling, a pour but d'étudier les liens existants entre p variables mesurées sur n individus, d'éliminer les redondances (deux variables corrélées apportant à peu près la même information), ainsi que de remplacer les variables initiales par un petit nombre de variables (1, 2 ou 3), appelées **axes factoriels** ou **composantes principales** en fonction du nuage de points étudié. Ces variables sont des combinaisons linéaires des variables initiales non corrélées entre elles.

14.1.3 Analyse factorielle des correspondances

L'analyse factorielle des correspondances (A.C.P.) est une méthode proposée par J.-P. Benzécri vers 1973, pour l'étude des tableaux de contingence de deux variables qualitatives. Elle est devenue la méthode privilégiée pour la description des données qualitatives et un outil puissant pour le dépouillement des enquêtes.

Le tableau des données contient les fréquences observées des modalités de deux phénomènes. Le test du χ^2 permet de déterminer s'il existe une liaison entre

ces deux phénomènes, et l'analyse factorielle des correspondances décrit cette liaison par l'intermédiaire du tableau de contingence qui peut mettre en évidence des ressemblances entre les colonnes du tableau, entre les lignes ou des proximités entre les lignes et les colonnes par l'intermédiaire de son *mapping*.

La méthode s'étend sur plusieurs variables *via* l'analyse factorielle des correspondances multiples (A.C.M.).

14.1.4 Analyse factorielle discriminante

Sur l'ensemble des individus d'une population P, on étudie p caractères quantitatifs et un caractère qualitatif prenant un nombre fini k de modalités. La population est répartie en k classes.

Le but de l'analyse factorielle discriminante (A.F.D.) est de rechercher si ce caractère qualitatif a une influence sur les p variables mesurées et de déterminer, éventuellement, des **caractères discriminants**, définissant, sur l'ensemble des individus, une partition aussi proche que possible de la partition induite par la variable qualitative initiale. Dit autrement, il s'agit d'une technique statistique ayant pour objectif de décrire, expliquer et prédire l'appartenance à des groupes prédéfinis (classes, modalités de la variable à prédire, *etc*.) d'un ensemble d'observations (individus, exemples, *etc*.) à partir d'une série des variables prédictives (descripteurs, variables exogènes, *etc*.). Ainsi, l'analyse factorielle discriminante se ramène à une analyse en composantes principales, effectuée sur l'ensemble des centres de gravité des individus d'une même classe, chaque classe correspondant à une des k modalités de la variable qualitative initiale.

L'A.F.D. est utilisée dans de nombreux domaines statistiques : la statistique exploratoire [1], l'analyse de données [2], la reconnaissance des formes [3], particulièrement utilisé dans la **reconnaissance optique de caractères** [4] (R.O.C.), l'apprentissage automatique [5], la fouille des données [6], *etc*. L'A.F.D. se subdivise en deux approches : 1. l'analyse factorielle discriminante descriptive 2. l'analyse factorielle discriminante prédictive.

L'analyse factorielle discriminante prédictive permet de construire une **fonction de classement**, c'est-à-dire définissant des règles d'affectation, qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives. Elle est définie par un cadre probabiliste, qui permet de proposer une analyse discriminante linéaire : la **technique de la régression logis-**

1. *exploratory data analysis*
2. *data analysis*
3. *pattern recognition*
4. *optical character recognition* (O.C.R.) ou océrisation
5. *machine learning*
6. *data mining*

tique. Celle-ci est fondée sur un modèle de régression binomiale, lui-même issu d'un cas particulier du modèle linéaire généralisé. La technique fut inventée par J. Berkson en 1944 et 1951, et fut baptisé « modèle *logit* ». Un choix binaire (0 ou 1) entraîne :

Joseph Berkson (1889-1982)

$$\ln\left(\frac{\Pr(X=1)}{\Pr(X=0)}\right) = a_0 + a_1x_1 + \ldots + a_jx_j \tag{14.2}$$

ou

$$\ln\left(\frac{\Pr(X=1)}{1-\Pr(X=1)}\right) = b_0 + b_1x_1 + \ldots + b_jx_j \tag{14.3}$$

$$\Pr(X=1) = \frac{e^{b_0+b_1x_1+\ldots+b_jx_j}}{1+e^{b_0+b_1x_1+\ldots+b_jx_j}} \tag{14.4}$$

14.2 Méthodes explicatives

14.2.1 Méthodes de régression

Les méthodes de régression consistent à trouver une relation, linéaire ou non, entre une variable expliquée et une ou plusieurs variables explicatives, l'ajustement étant obtenu en général, par la **méthode des moindres carrés**. Elles sont utilisées dans le cas où toutes les variables explicatives sont quantitatives, comme cela a été expliqué lors de l'approche bivariée.

14.2.2 Analyse canonique

L'analyse canonique développée par H. Hotelling généralise la méthode de régression multiple, mais présente un intérêt théorique assez limité, car elle conduit à de grandes difficultés d'interprétation. Cette méthode cherche à synthétiser les relations pouvant exister entre deux groupes de variables, en déterminant les combinaisons linéaires des variables du premier groupe les plus corrélées à des combinaisons linéaires des variables du second groupe. Si le second groupe est constitué d'une seule variable, on retrouve la régression multiple.

14.2.3 Analyse de la variance

L'analyse de la variance est une méthode qui consiste à tester l'influence d'une ou plusieurs variables qualitatives sur une variable quantitative. Contrairement à son nom, il s'agit d'étudier des moyennes. On cherche à contrôler si une variation

des modalités prises par les variables explicatives, seules ou combinées, entraîne une variation de la variable expliquée Y.

Précédemment, l'étude portée sur deux variables : une quantitative et une qualitative. Dans cette partie, il s'agira d'étendre l'analyse à plusieurs variables qualitatives.

14.2.4 Analyse de la covariance

L'analyse de la covariance est une méthode qui généralise les méthodes de régression et de l'analyse de la variance.

14.3 Méthodes de prévision

Les méthodes de prévision concernent principalement l'analyse et la prévision des séries chronologiques. Celles-ci ont principalement pour but de mettre en évidence une tendance, une saisonnalité et un résidu à l'aide d'un modèle multiplicatif, le plus utilisé en gestion, ou d'un modèle additif.

Bibliographie

[Benzécri, 1973a] BENZÉCRI, J.-P. (1973a). L'analyse de données, t. 1, La taxinomie. Dunod, Paris.

[Benzécri, 1973b] BENZÉCRI, J.-P. (1973b). L'analyse de données, t. 2, L'analyse de correspondances. Dunod, Paris.

[Hotelling, 1933] HOTELLING, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6): 417–441.

[Spearman, 1904] SPEARMAN, C. (1904). "general intelligence", objectively determined and measured. American Journal of Psychology, 15(2):201–293.