Première partie

Préparation, description et visualisation des données

Chapitre 2

Analyses Factorielles, MDS

L'objectif de ce chapitre est de faire un rappel (ou présenter une première approche) de méthodes utilisées en data mining pour visualiser et préparer des données multivariées : l'analyse en composantes principales, l'analyse des correspondances et le positionnement multidimensionnel. Ces techniques d'analyse descriptive seront utilisées, notamment, pour visualiser les données dans un sous espace représentatif, pour détecter des valeurs extrèmes ou abberrantes, pour aider au choix de variables.

2.1 Analyse en composantes principales

L'analyse en composantes principales est un outil de réduction de dimension qui permet de retirer la redondance ou la duplicité dans un ensemble de variables corrélées. L'ensemble initial est alors représenté par un ensemble réduit de variables, dérivées des variables observées. Les facteurs ainsi obtenus sont, en théorie, indépendants les uns des autres.

Soit $\{X_1, \cdots, X_p\}$ un ensemble de p variables observées sur n individus indépendants. On notera $\mathbf{x} = \{x_1, \cdots, x_p\}$ les observations. Pour tout $j \in \{1, \cdots, p\}$, $x_j \in \mathbb{R}^n$.

2.1.1 En préambule : problème à un facteur

Considérons que l'on veuille décrire les données à partir d'un seul facteur latent. Ce problème se pose par exemple lorsque l'on analyse des questionnaires de qualité de vie. Les patients interrogés répondent à un certain nombre d'items représentant des variables. Ils donnent par exemple une note de 1 à 10 à des questions du type : *Le soir vous vous endormez rapidement.* ou *Vous ressentez des douleurs.* Et on cherche à en déduire une information synthétique sur la qualité de vie du (ou des) patient(s) sous la forme d'un indice. Le modèle peut alors s'écrire :

$$X_j = \nu_j F + \epsilon_j \text{ pour tout } j \in \{1, \cdots, p\} \tag{2.1}$$

avec F un facteur commun aux variables X_j et ϵ_j une variable aléatoire centrée de variance σ_j^2 et indépendante de F. La constante ν_j est interprétée comme une portée et elle caractérise la contribution de la variable X_j au facteur F. Dans le modèle (2.1), le terme $\nu_j F$ joue le rôle d'un effet moyen.

On pourrait aussi interpréter ce modèle comme un modèle de régression entre X_j et F. Cependant, il est important de noter ici que la variable F est latente (non observée).

Remarques : Si le facteur F est normé, c'est à dire si $var(F) = 1$ alors on remarque que

$$Corr(X_j, F) = \nu_j$$

et

$$Corr(X_j, X_k) = \nu_j \nu_k \text{ si } j \neq k$$

$$Corr(X_j, X_j) = \nu_j^2 + \sigma_j^2$$

La covariance est complètement déterminée par la décomposition en facteurs.

Si on suppose de plus que les erreurs ϵ_j sont toutes de même variance σ, l'équation (2.1) conduit au modèle de l'analyse en composantes principales.

$$\mathbf{X} = \nu^T F + \epsilon$$

avec ϵ un vecteur aléatoire centré de matrice de covariance σ^2Id. Et on peut alors considérer l'analyse principale comme une méthode d'inférence qui va permettre d'estimer ν et F.

2.1.2 Analyse en composantes principales

ACP par minimisation de l'erreur

Considérons maintenant le cas où on a autant de facteurs que de variables. On peut alors écrire

$$\mathbf{X} = \sum_{j=1}^{p} z_j F_j$$

Si on veut extraire q facteurs latents, on peut décomposer l'équation précédente sous la forme

$$\mathbf{X} = \sum_{j=1}^{q} z_j F_j + \sum_{j=q+1}^{p} b_j F_j$$

Et on cherche, logiquement, la représentation qui conduit à l'erreur $\epsilon = \sum_{j=q+1}^{p} z_j F_j$ de plus faible variance ou de manière équivalente

On cherche donc les matrices $\mathbf{z}^*$ et $\mathbf{F}^*$ telles que

$$(\mathbf{z}^*, \mathbf{F}^*) = arg \min_{\{(z,F) \in \mathbb{R}^q \times \mathbb{R}^q, \mathbf{F}\mathbf{F}^T = Id\}} Var\left(\mathbf{X} - \sum_{j=1}^{M} z_j F_j\right)$$

En pratique, on ne sait pas calculer cette variance. On l'estime à partir des observations. Et on montre que la solution unique est donnée par les composantes principales $\hat{F}$ et las axes pricipaux $\hat{z}$, vecteurs propres de la matrice de variance-covariance.

Plus généralement, si les observations sont définies dans un espace muni d'une métrique M et sont pondérées par la matrice de poids diagonale D, on formule le problème ainsi :

$$(\mathbf{z}^*, \mathbf{F}^*) = arg \min_{\{(z,F)\in\mathbb{R}^q\times\mathbb{R}^q, \mathbf{F}\mathbf{F}^T=Id\}} ||\mathbf{x} - \sum_{j=1}^{q} z_j F_j||^2_{(M,D)}$$

Si l'espace est euclidien, par définition

$$||\mathbf{x} - \sum_{j=1}^{q} z_j F_j||^2_{(M,D)} = D(\mathbf{x} - \sum_{j=1}^{q} z_j F_j)^T M(\mathbf{x} - \sum_{j=1}^{q} z_j F_j)$$

et sa solution donnée par

$$\sum_{j=1}^{q} z_j F_j = \sum_{j=1}^{q} \lambda_j u_k v_k^T$$

avec U et V des matrices unitaires.
Remarque : Le choix de la métrique M et/ou de la matrice de pondération D a un impact sur les résultats et notamment sur les projections des individus sur les plans factoriels. Certaines métriques permettent par exemple de mettre en évidence les individus atypiques (voir TD).

ACP par projection

Une approche équivalente consiste à considérer qu'on cherche des facteurs F_k orthonormés qui sont une combinaison linéaire des variables d'origine $X_1, \cdots, X_p$ et tels que le sous espace engendré par les F_k, $k = 1, \cdots q$, soit le sous espace de dimension q dans lequel le nuage de points $\mathbf{x}$ est d'inertie maximale. L'*inertie* est définie comme la somme des distances au carré des points à leur centre de gravité. Dans le cas où les variables dont quantitatives, c'est aussi aussi n fois la somme des variances de chacune des variables, soit la trace de la matrice de variance-covariance.

Finalement, nous cherchons le vecteur u tel que la projection du nuage sur u ait une inertie (ou une variance) maximale. Notons $\tilde{\mathbf{x}}$ le nuage de points centré (et éventuellement réduit). La projection de l'échantillon des $\tilde{\mathbf{x}}$ sur u s'écrit :

$$P_u(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}} \cdot u$$

la variance empirique de $P_u(\tilde{\mathbf{x}})$ vaut donc :

$$\frac{1}{n} P_u(\tilde{\mathbf{x}})^T P_u(\tilde{\mathbf{x}}) = u^T \cdot \underbrace{\frac{1}{n}\tilde{\mathbf{x}}^T\tilde{\mathbf{x}})}_{\hat{\Sigma}} \cdot u$$

où $\hat{\Sigma}$ est la matrice de covariance empirique de $\tilde{\mathbf{x}}$. Ainsi, pour le premier vecteur propre, on cherche un vecteur unitaire u^* tel que

$$u^* = arg \max_{\{u\in\mathbb{R}^n, u^T u=1\}} u^T \Sigma u$$

En introduisant les multiplicateurs de Lagrange, ce problème est équivalent à

$$(u^*, \lambda_1) = arg \max_{\{u\in\mathbb{R}^n, \lambda\in\mathbb{R}\}} u^T \Sigma u - \lambda(u^T u - 1)$$

En annulant la dérivée de l'expression ci-dessous, on vérifie que u^* est le vecteur propre associé à la valeur propre λ_1 de Σ. La valeur propre λ_1 est la variance empirique sur le premier axe de l'ACP.

Exemple : Si $X \in \mathbb{R}^2$ et centré, alors rechercher le premier axe principal revient à rechercher l'équation de la droite $Y = a_1X_1 + a_2X_2$ qui est telle que la variance de Y soit maximale. On cherche donc a_1 et a_2 tels que $Var(a_1X_1+a_2X_2)$ soit maximale. Autrement dit $a_1 = Cov(X_1, X_2)/Var(X_1)$ et $a_2 = Cov(X_1, X_2)/Var(X_2)$.

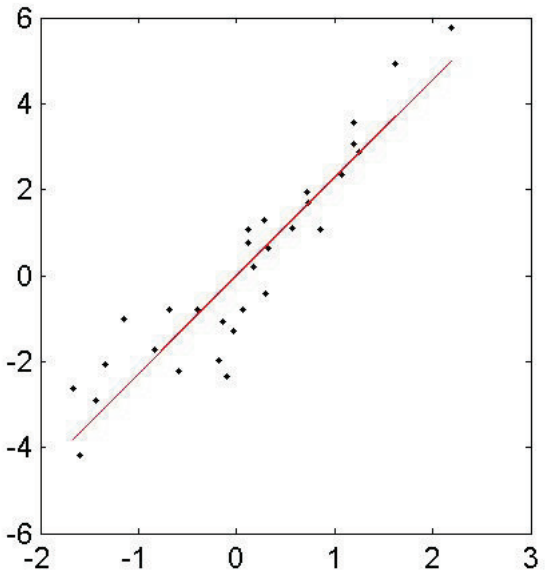

FIGURE 2.1 – ACP en 2 dimensions. L'axe en trait continu représente le premier axe principal

A voir : ecrire les estimateurs des axes principaux et des composantes principales comme des estimateurs de max de vraisemblance quand le modèle de l'ACP est vu comme un modèle linéaire à effet fixe signal plus bruit, avec une bruit gaussien.

2.1.3 Représentations graphiques

L'analyse en composantes principales est souvent utilisée pour donner une représentation graphique des individus et des variables.

Les individus

Les graphiques obtenus permettent de représenter au mieux les distances euclidiennes inter-individus mesurées par la métrique M. En pratique, on projette orthogonalement les obsservations $\mathbf{x}$ sur les plans factoriels. Et les coordonnées de $\mathbf{x}_i - \bar{\mathbf{x}}$ sur le sous espace de dimension q sont les q premiers éléments de la matrice $C = U\Lambda^{1/2}$.

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

$$r_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

Tandis que la qualité de la représentation de chaque point est donnée par

$$cs_i^2 = \frac{\sum_{j=1}^{q} c_{ij}^2}{\sum_{j=1}^{p} c_{ij}^2}$$

La contribution de chaque individu à l'inertie du nuage permet de détecter les observations les plus influentes et éventuellement aberrantes.

$$\gamma_i = \omega_i \frac{\sum_{j=1}^{p} c_{ij}^2}{\sum_{j=1}^{p} \lambda_j}$$

On peut projeter des individus supplémentaires $\mathbf{s}$ sur un sous espace factoriel en calculant ses coordonnées :

$$V_q^T M(\mathbf{s} - \bar{\mathbf{x}})$$

Ici V_q joue le rôle d'une matrice de changement de base/

Les variables

Les graphiques obtenus permettent de représenter "au mieux" les corrélation entre les variables et , si celles ci ne sont pas réduites, leurs variances. On obtient le cercle des corrélations par projection D-orthogonale sur le sous espace factoriel. La coordonnées de x_j sur u_k est données par

$$\sqrt{\lambda} v_{jk}$$

La qualité de la représentation de chaque x_j est mesurée par

$$\frac{\sum_{j=1}^{q} \lambda_j v_{jk}^2}{\sum_{j=1}^{p} \lambda_j v_{jk}^2}$$

Représentation simultanée

2.1.4 Exemple

A titre d'exemple, on considère un jeu de données établissant la composition du lait de 25 espèces de mammifères. On mesure 5 variables : la teneur en protéines, en lactose, en graisse, en eau et en minéraux.

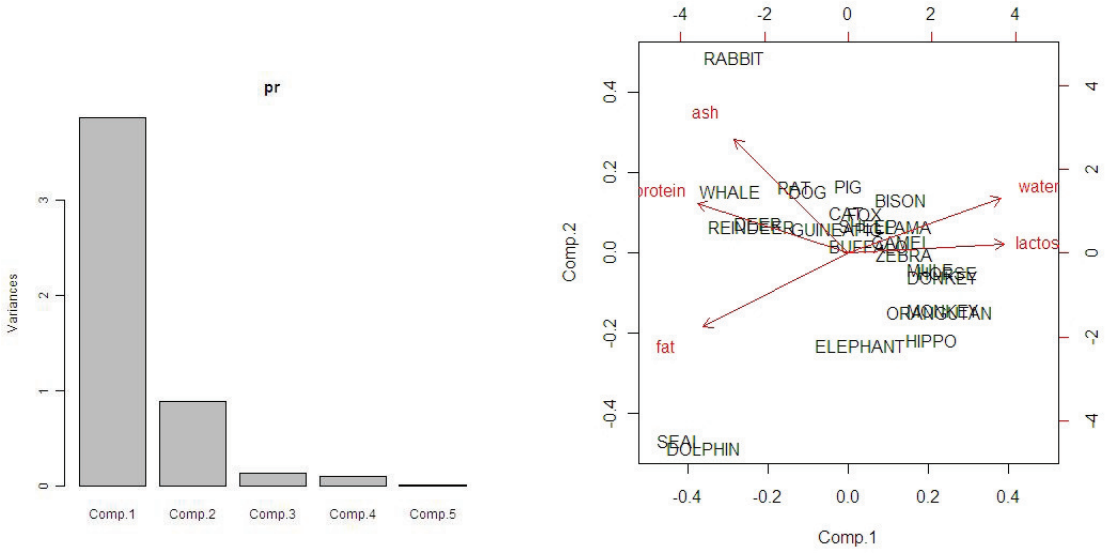

FIGURE 2.2 – Composition du lait - Ebouli des valeurs propres (à gauche) et représentation simultannée sur le premier plan principal de l'ACP (à droite)

2.2 Analyse Factorielle des Correspondances (Multiples)

On considère dans cette section p variables qualitatives observées simultanément sur n individus de poids identiques $1/n$.

Exemple - Considérons le jeu données suivant dans lequel on caractérise différentes races de chien en fonction de 7 variables portant sur des caractéristiques de physique, de caractère et d'utilité.

	Taille	Poids	Velocité	Intelligence	Affectivité	Agressivité	Utilité
beauceron	3	2	3	2	2	2	utilite
basset	1	1	1	1	1	2	chasse
ber_allem	3	2	3	3	2	2	utilite
boxer	2	2	2	2	2	2	compagnie
bull-dog	1	1	1	2	2	1	compagnie
bull-mass	3	3	1	3	1	2	utilite

Tableau disjonctif complet

Il est difficile de travailler directement avec un tel tableau de données. En effet, on ne peut pas considérer ces données comme des données quantitatives. Par exemple, ça n'a pas de sens de considérer qu'il y a une distance équivalente entre les classes 1 et 2 de la variable Poids et de la variable Intelligence. En conséquence, il est d'usage de recoder les données et de construire le *tableau disjonctif complet*.

Le *tableau disjonctif complet* est tel que chaque ligne correspond à un individu et chaque colonne à une modalité. On note K le nombre total de modalités. Et les observations x_{ij} sont codées 1 si l'individu i a la modalité j et 0 sinon. Notons X le tableau disjonctif complet.

Dans l'exemple, on obtient :

	T1	T2	T3	P1	P2	P3	V1	V2	V3	I1	I2	I3
beauceron	0	0	1	0	1	0	1	0	0	0	1	0
basset	1	0	0	1	0	0	1	0	0	1	0	0
ber_allem	0	0	1	0	1	0	1	0	0	1	0	0
boxer	0	1	0	0	1	0	0	1	0	0	1	0
bull-dog	1	0	0	1	0	0	1	0	0	0	1	0
bull-mass	0	0	1	0	0	1	1	0	0	1	0	0

A1	A2	A3	Ag1	Ag2	Ag3	Ut	Ch	Co
0	1	0	0	1	0	1	0	0
1	0	0	0	1	0	0	1	0
0	1	0	0	1	0	1	0	0
0	1	0	0	1	0	0	0	1
0	1	0	1	0	0	1		
1	0	0	0	1	0	1	0	0

Tableau de Burt

On appelle tableau de Burt le tableau $\mathcal{B} = X^T X$. On peut écrire $\mathcal{B} = (B_{k,k'})_{k,k'=1,\cdots,K}$ où

- si $k \neq k'$, $B_{k,k'}$ est la table de contingence des variables X_k et $X_{k'}$,
- si $k = k'$, B_{kk} est une matrice diagonale contenant les effectifs marginaux de X_k dans la diagonale, notés $n^k_{c_1}, \cdots, n^k_{c_k}$.

Propriétés :

- $\mathcal{B}$ est symétrique.
- La somme des lignes (resp. des colonnes) de $\mathcal{B}$ est $Kn^k_l, l = 1, \cdots, c_k$.
- La somme des éléments de $\mathcal{B}$ est $K^2 n$.

Remarque : si on considère les données du tableau disjonctif X comme des observations de variables qualitatives, alors le tableau de Burt représente la variance de X à un facteur multiplicatif près.

AFCM

L'AFCM est une double ACP du tableau de Burt (ou du tableau disjonctif complet). En pratique, on cherche les q premiers vecteurs propres $u_1, \cdots, u_q$ de la matrice

$$\left(\frac{1}{K}\mathcal{B}D\right)^2$$

avec D la matrice diagonale définie par $D = \text{diag}(D_1, ...D_p)^{-1}$ où $D_k = \text{diag}(n^k_1, , ...n^k_{c_k})/n$.

Les composantes principales sont alors

$$C_j = nDBDU_j$$

Elles permettent la représentation simultanée de toutes les modalités.

La contribution de la modalité c de la variable X_k à l'Šinertie de l'Šaxe j est donnée par

$$\frac{\frac{n_c^k}{mn}(c_i^j)^2}{\lambda_j}$$

Attention ! En AFCM, seules certaines valeurs propres du tableau de Burt sont considérées. Aussi, on ne peut donner de mesure de la qualité globale de la représentation à partir des valeurs propres. Seules les contributions des modalités à l'inertie selon les axes sont interprétées, selon le même principe qu'en AFC.

En pratique,

- On interprète les proximités et les oppositions entre les modalités des différentes variables
- On privilégie les interprétations sur les modalités suffisamment éloignées du centre du graphique
- Les rapports de valeurs propres ne sont pas interprétables mais on regarde la décroissance des valeurs propres pour choisir la dimension (exemple figure 2.3).
- Seules les contributions des modalités à l'inertie selon les axes sont interprétables

Exemple des races de chiens
Revenons à l'exemple des races de chiens. Nous obtenons les graphiques de la figure 2.3.

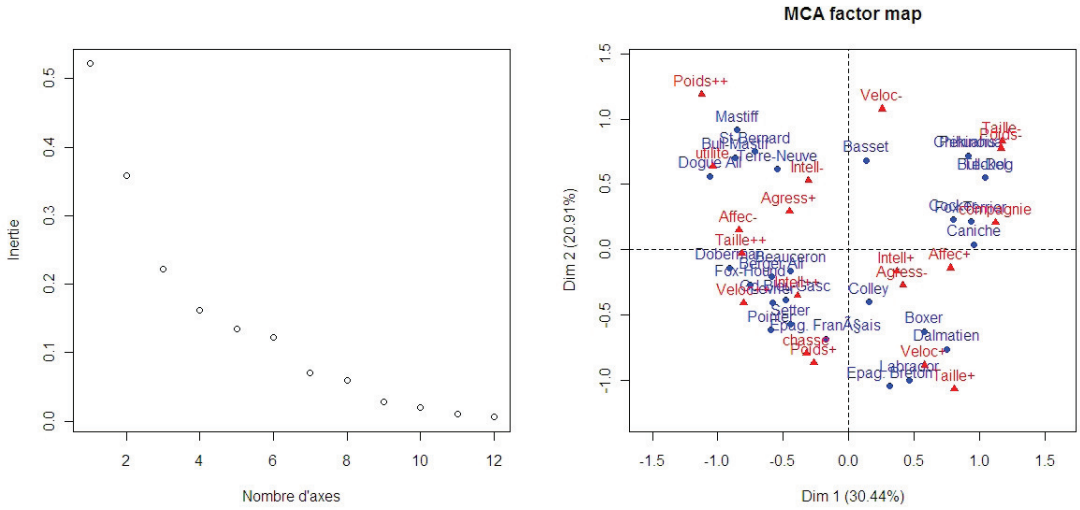

FIGURE 2.3 – AFCM des données sur les races de chiens. Ebouli des valeurs propres (à gauche) et projection simultanée sur le premier plan factoriel (à droite).

2.3 Positionnement multidimensionnel

Considérons n individus. Contrairement aux sections précédentes on suppose ici que les données sont les $1/2n(n-1)$ valeurs d'un indice (de distance, similarité ou dissimilarité) observées ou construites pour chacun des couples d'individus. Ces valeurs sont stockées dans une matrice. L'objectif du positionnement multidimensionnel (multidimensional scaling ou MDS ou ACP d'un tableau de distances) est de construire, à partir de la matrice, une représentation euclidienne des individus dans un espace de dimension réduite qui rapproche "au mieux" les indices observés. Si la dimension de l'espace est assez faible, ceci permet alors d'obtenir une représentation graphique.

2.3.1 Distance entre variables

L'un des intérêts pratiques du positionnement multidimensionnel est d'aider à comprendre les structures de liaison dans un grand ensemble de variables. On obtient ainsi des indications pour choisir un sous-ensemble de variables. Cette approche nécessite de définir des distances entre variables.

Variables quantitatives

Soient X et Y deux variables dont les observations sur les mêmes n individus sont notés $\mathbf{x}$ et $\mathbf{y}$. On suppose sans restriction que $\mathbf{x}$ et $\mathbf{y}$ sont centrés.

On définit la distance suivante où $\text{cor}^2(X,Y)$ est la corrélation entre X et Y.

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

On remarque ici que la distance entre deux variables est associée à une notion de dépendance entre les variables, représentée par la corrélation.

Variables quantitatives

Considérons maintenant deux variables quantitatives, X à r modalités et Y à s modalités.

$$d^2(X,Y) = 2(1 - T^2(X,Y)) \tag{2.2}$$

avec le T de Tschuprow défini par

$$T = \sqrt{\frac{\chi^2_{\text{observé}}}{n\sqrt{\nu}}}$$

où $\nu = (r-1)(s-1)$ et

$$\chi^2_{\text{observé}} = n\left[\sum_{l=1}^{r}\sum_{h=1}^{s}\frac{n_{lh}^2}{n_{l+}n_{+h}} - 1\right]$$

Ici n_{lh} représente le nombre d'individus tels que $x = l$ et $y = h$, n_{l+} est le nombre d'individus tels que $x = l$ et n_{+h} le nombre d'individus tels que $y = h$.

On observe, selon l'équation (2.2), que la distance entre deux variables X et Y est d'autant plus grande que le $T^2(X,Y)$ est proche de 0. Or le T est une fonction croissante du $\chi_{\text{observé}}$. Et le

premier terme du $\chi_{\text{observé}}$ donne une approximation empirique du rapport

$$\frac{P(X=x, Y=y))}{P(X=x)P(Y=y)}$$

On rappelle que si X et Y sont indépendantes (et donc distantes), ce rapport est égal à 1 et, en conséquence, le T de Tschuprow proche de 0. Et on retrouve bien, comme dans le cas de deux variables quantitatives, que la notion de distance est associée à une mesure de dépendance entre les variables.

Variables qualitative et quantitative

Lorsque l'on considère un ensemble de variables constitué de variables quantitatives et qualitatives, on doit considérer une distance adaptée. On propose ici la distance suivante

$$d^2(X,Y) = 2(1 - R_c^2(X,Y))$$

où R_c désigne le rapport de corrélation.

Le rapport de corrélation évalue le rapport entre une variable indépendante qualitative nominale ou ordinale et une variable dépendante quantitative. Il compare la variation totale de la variable dépendante aux variations non expliquées par la variable qualitative.

$$R_c^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}}$$

en notant X la variable qualitative, x_i ses observations et μ sa moyenne.

$$\text{Variation totale} = \sum_{i=1}^{n}(x_i - \mu)^2$$

et si on note μ_k la moyenne de X dans la catégorie k de la variable qualitative,

$$\text{Variation expliquée} = \sum_{k=1}^{K}(\mu_k - \mu)^2$$

2.3.2 Recherche d'une configuration de points

Soit D une matrice de distance de terme général δ_{ij} et $B = HAH$ la matrice centrée en lignes et en colonnes associée.

On cherche une configuration $\vec{X}^* = (x_{ia}^*)$, $a = 1, \ldots, q$ de n points dans un espace de dimension q, en général euclidien, tel que le point $\vec{x}_i^* = (x_{i1}^*, \ldots, x_{ip}^*)^T$ représente de façon unique l'objet i et la distance euclidienne entre les points $\vec{x}_i^*$ et $\vec{x}_j^*$

$$d_{ij}(\vec{X}^*) = \|\vec{x}_i^* - \vec{x}_j^*\| = \sqrt{\sum_{a=1}^{p}\left(x_{ia}^* - x_{ja}^*\right)^2}$$

approche δ_{ij}, pour toute paire d'objets (i,j).

Dans le cas d'une matrice D euclidienne supposée de rang r, la solution est obtenue en éxécutant les étapes suivantes :

1. Contruction de la matrice A de terme général $-\frac{1}{2}d_{jk}^2$
2. Calcul de la matrice des produits scalaires par double centrage $B = HAH$
3. Diagonalisation de $B = U\Delta U^T$
4. Les coordonnées d'une configuration, appelées *coordonnées principales*, sont les lignes de la matrice $X = U\Delta^{1/2}$.

On remarque que dans le cas euclidien le MDS et l'ACP sont connectés et peuvent conduire à des résultats identiques.

L'intérêt du MDS apparaît evidemment lorsque l'on cherche la meilleure représentation euclidienne de distances non euclidiennes entre les individus. En ce sens, le MDS "généralise" l'ACP.

Théorème 1 *Si $\mathcal{D}$ est une matrice de distance, pas nécessairement euclidienne, alors pour une dimension q fixée, la configuration issue du MDS de distance $\hat{\mathcal{D}}$ qui rend $\sum_{j,k=1}^{n}(d_{jk}^2 - \hat{d}_{jk}^2)$ minimum.*

On peut aussi formuler ce théorème en terme de produit scalaires associés à $\mathcal{D}$.

2.3.3 Exemple

Dans cet exemple, on donne une représentation graphique des proximités géographiques entre quelques grandes villes européennes obtenue à partir de leurs distances kilométriques par la route. La structure du réseau routier fait que cette distance n'est pas euclidienne. On observe cependant sur le graphique (fig. 2.4) qui donne une approximation euclidienne que la distance euclidienne est très proche de la distance "routière".

```
library(MASS)
require(graphics)

eurodist
loc <- cmdscale(eurodist)
x <- loc[,1]
y <- -loc[,2]
plot(x, y, type="n", xlab="", ylab="", main="cmdscale(eurodist)")
text(x, y, rownames(loc), cex=0.8)
```

2.3.4 Application à la sélection de variables

Voir td no 5.

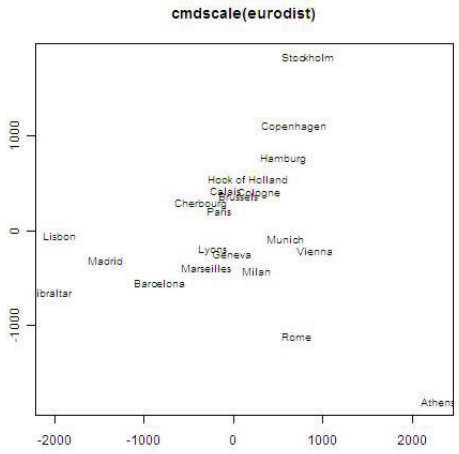

FIGURE 2.4 – Représentation, dans un plan, des distances entre quelaues grandes villes européennes