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].

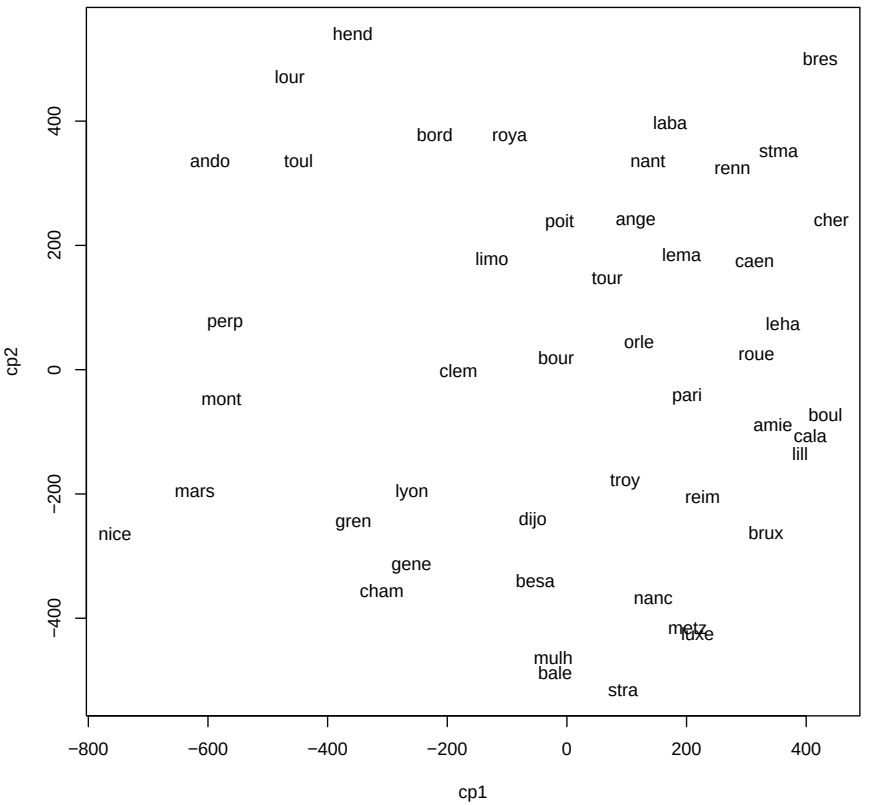

FIGURE 1 – *Villes : Positionnement de 47 villes à partir de la matrice de leurs distances kilométriques.*

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

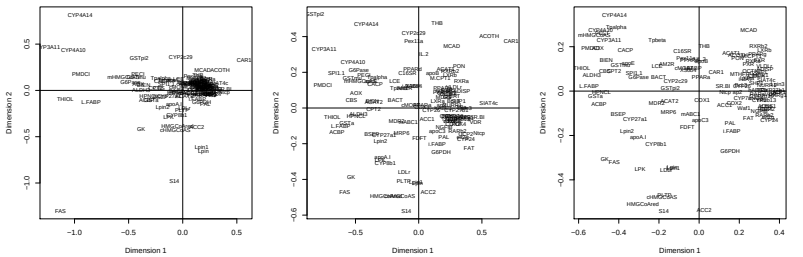

FIGURE 2 – *Souris : positionnement multidimensionnel des gènes sur les axes 1 et 2 selon 3 distances différentes : distance euclidienne (d_1 à gauche), corrélation (d_3 au centre), corrélation carrée (d_2 à droite).*

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.