

Échelonnement multidimensionnel

(Multidimensional scaling MDS)

Objectif de la méthode

L'échelonnement multidimensionnel ou positionnement multidimensionnel a deux objectifs:

- Déterminer si certains facteurs sont négligeables dans un tableau de donnée.
- Créer une représentation visuelle des données

Données manipulées

Matrice de similarité aussi appelée matrice de proximité

Exemple

	Atlanta	Boston	Chicago	Washington	Denver	Los Angeles	Miami	NYC	Seattle	San Francisco	New Orleans
Atlanta	0	934	585	542	1209	1942	605	751	2181	2139	424
Boston	934	0	853	392	1769	2601	1252	183	2492	2700	1356
Chicago	585	853	0	598	918	1748	1187	720	1736	1857	830
Washington	542	392	598	0	1493	2305	922	209	2328	2442	964
Denver	1209	1769	918	1493	0	836	1723	1636	1023	951	1079
Los Angeles	1942	2601	1748	2305	836	0	2345	2461	957	341	1679
Miami	605	1252	1187	922	1723	2345	0	1092	2733	2594	669
NYC	751	183	720	209	1636	2461	1092	0	2412	2577	1173
Seattle	2181	2492	1736	2328	1023	957	2733	2412	0	681	2101
San Francisco	2139	2700	1857	2442	951	341	2594	2577	681	0	1925
New Orleans	424	1356	830	964	1079	1679	669	1173	2101	1925	0

Description

Le Multidimensional Scaling (MDS) est une méthode d'analyse de données largement utilisée dans les domaines du marketing et de la psychométrie, particulièrement dans les pays anglo-saxons. Le principe de la méthode consiste à reconstituer une carte d'individus à partir d'une matrice de proximités (similarités ou dissimilarités) entre les individus.

Dans le cas idéal où l'on dispose d'une matrice donnant les distances entre des points dans le plan (par exemple, les distances entre les villes d'une région), le Multidimensional Scaling reconstitue la carte des points, à une rotation/symétrie près.

Construction

Étant donné N points x_1, x_2, \dots, x_N dans un espace de dimension p , le positionnement multidimensionnel consiste à représenter ces points dans un espace de dimension $m < p$ par N nouveaux points y_1, y_2, \dots, y_N en conservant les proximités. On se donne pour cela une matrice de distance D qui peut être définie par la distance euclidienne $d_{ij} = \|x_i - x_j\|_2$. Si on part de valeurs de similarité, il faut les convertir en valeurs de vraie distance mathématique, car il faut conserver à l'esprit que distance et similarité sont des notions opposées : plus faible est la distance, plus grande est la similarité, et réciproquement.

Présenté sous cet angle, le positionnement multidimensionnel est une technique de réduction de dimension, au même titre que l'analyse en composantes principales. En pratique, le positionnement multidimensionnel consiste à trouver N vecteurs y_1, y_2, \dots, y_N de taille m qui minimisent une fonction de coût $S(y_1, y_2, \dots, y_N)$ appelée *stress*. Plus le stress est proche de 0, meilleure est la représentation.

Les différents types de l'échelonnement multidimensionnel (classique, métrique ou non métrique) correspondent en fait aux différentes formes possibles de la fonction de coût.

MDS classique

$$S(y_1, y_2, \dots, y_N) = \sum_{i \neq j} (b_{ij} - \langle y_i, y_j \rangle)^2$$

MDS métrique

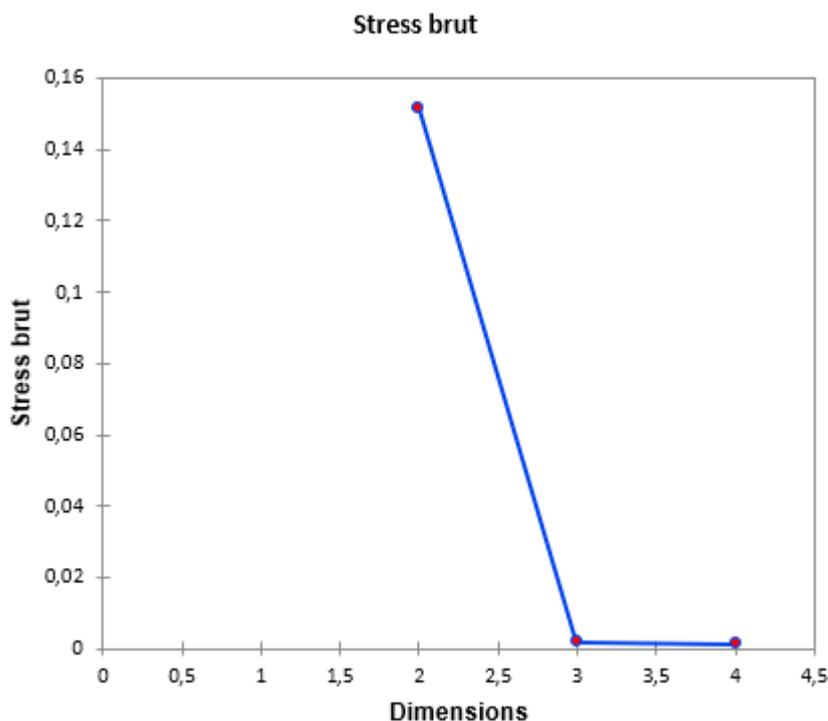
$$S(y_1, y_2, \dots, y_N) = \sum_{i \neq j} (d_{ij} - \|y_i - y_j\|)^2$$

MDS non métrique

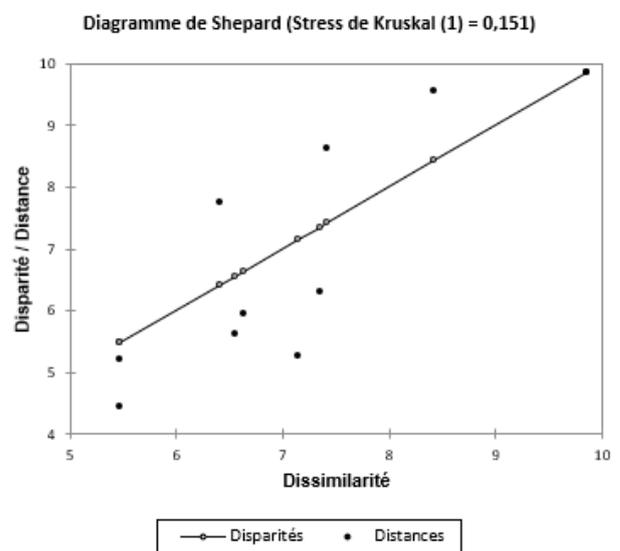
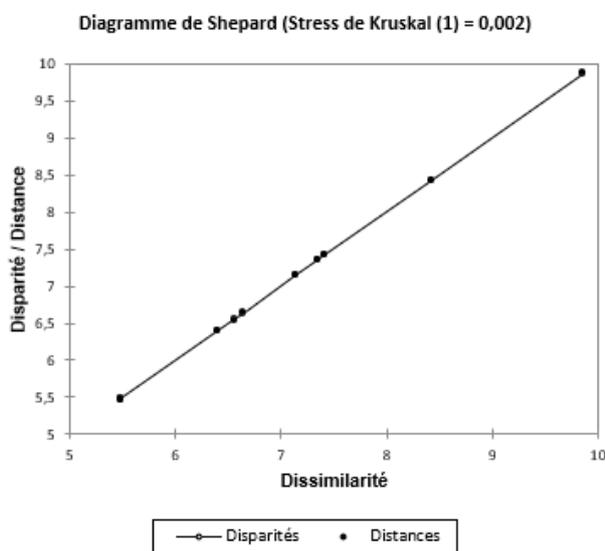
$$S(y_1, y_2, \dots, y_N) = \sum_{i \neq j} (d_{ij} - f(\|y_i - y_j\|))^2$$

Interprétation des résultats d'un MDS

Le premier tableau montre l'évolution du stress en fonction du nombre de dimensions de l'espace de représentation. On note une rupture très nette entre les dimensions 2 et 3, et une stabilité entre les dimensions 3 et 4 (il est normal que la représentation de 5 objets soit parfaite dans un espace à 4 dimensions).



Remarque : il n'existe pas de méthode statistique rigoureuse pour évaluer la qualité et la fiabilité d'une représentation issue d'un MDS. Néanmoins l'observation du diagramme de Shepard permet d'avoir une idée générale de la qualité de la représentation. Le diagramme de Shepard correspond à un nuage de points, dont les abscisses sont les dissimilarités observées, et les ordonnées, les distances dans la configuration issue du MDS. Plus les points sont dispersés, moins le graphique est fiable.



Annexe :

Wikipédia : https://fr.wikipedia.org/wiki/Analyse_des_donn%C3%A9es#Positionnement_multidimensionnel

https://fr.wikipedia.org/wiki/Positionnement_multidimensionnel

XLStat : <https://help.xlstat.com/s/article/quelle-methode-danalyse-multivariee-choisir?language=fr>

<https://www.youtube.com/watch?v=ACYE8oCgzJU>

