

Analyse des données avancée

Chapitre 12 : Distances et dissimilarités



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Que veut dire « se ressembler »

Ce qu'une distance doit vérifier

Classer sur les axes factoriels

Ce qu'il faut retenir

Que veut dire « se ressembler »

Une question qui paraît simple

Le directeur commercial demande : « regroupez-moi les clients qui se ressemblent ».

Se ressemblent selon quoi ? Selon leurs notes, selon leur panier, selon leur fréquence de visite ? Et si l'on prend les notes, faut-il compter la propreté autant que le prix ?

Ce que la classification exige, et l'ACP n'exigeait pas

L'ACP se contentait des données : la géométrie découlait des corrélations.

La classification exige une décision préalable : celle d'une distance. Aucun algorithme ne la choisira à votre place, et le résultat en dépend entièrement.

Distance	Formule	Quand l'employer
Euclidienne	$\sqrt{\sum_j (x_{ij} - x_{i'j})^2}$	variables quantitatives réduites
Manhattan	$\sum_j x_{ij} - x_{i'j} $	si l'on craint les valeurs extrêmes
du χ^2	$\sum_j \frac{1}{c_j} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2$	profils, tableaux de contingence
de Jaccard	$1 - \frac{ A \cap B }{ A \cup B }$	présence / absence, paniers d'achat

Le choix de la distance précède le choix de la méthode — et pèse davantage.

Deux clients « proches » ne le sont que pour une distance donnée. Sur des variables non réduites, l'euclidienne classe selon la variable la plus dispersée : **centrer et réduire est une décision, pas une formalité.**

La distance du χ^2 a une propriété que les autres n'ont pas — *l'équivalence distributionnelle* : regrouper deux colonnes de profils identiques ne change rien au résultat. C'est ce qui fait d'elle la distance des tableaux de contingence.

La distance de Jaccard ignore les doubles absences : deux clients qui n'ont acheté *ni* l'un *ni* l'autre produit ne se ressemblent pas pour autant.

Ce qu'une distance doit vérifier

Distance et dissimilarité

Une **distance** vérifie : $d(i, i) = 0$, $d(i, j) = d(j, i)$, et l'inégalité triangulaire $d(i, k) \leq d(i, j) + d(j, k)$.

Une **dissimilarité** ne vérifie que les deux premières.

Pourquoi la nuance compte au chapitre 18

Quand on demande à un client « ces deux enseignes se ressemblent-elles ? », rien ne garantit l'inégalité triangulaire : il peut juger A proche de B , B proche de C , et A très loin de C .

Les jugements humains ne forment pas une distance. C'est précisément pour cela que le positionnement multidimensionnel existe : il cherche la carte qui approche au mieux des dissimilarités qui n'en sont pas.

Notre fichier mélange les deux natures

Huit notes quantitatives, cinq variables qualitatives. Aucune distance usuelle ne les traite ensemble.

Trois solutions, par ordre de préférence croissante : **découper les notes en tranches** et tout traiter en qualitatif; utiliser un **indice mixte** comme celui de Gower; ou **faire une analyse factorielle d'abord** et classer sur les coordonnées obtenues.

La troisième est la bonne, et c'est celle du cours

On classe sur les **premiers axes factoriels**, jamais sur les variables brutes.

Ces axes sont quantitatifs, non corrélés, et de même nature quelle que soit l'origine des données. La distance euclidienne y retrouve tout son sens.

Classer sur les axes factoriels

Pourquoi passer par l'ACP avant de classer

1. **Débruiter** : les derniers axes ne portent que de l'erreur de mesure.
2. **Décorrélérer** : sur des variables corrélées, la distance euclidienne compte plusieurs fois la même information.
3. **Homogénéiser** : une seule échelle, quelle que soit la nature des données de départ.

Le deuxième point, sur nos données

Prix, choix et qualité sont corrélés autour de 0,70. Deux clients qui diffèrent sur ces trois notes diffèrent **trois fois pour une seule raison**.

La distance euclidienne sur les huit notes brutes surpondère donc le bloc « offre ». **Sur les composantes, le problème disparaît par construction** : elles sont non corrélées.

Une question distincte de celle du chapitre 7

Pour **représenter**, deux axes suffisaient — on ne dessine pas en dimension trois.

Pour **classer**, on peut en garder davantage : rien n'oblige à visualiser. L'usage est de conserver les axes jusqu'à 80 ou 90 % de l'inertie.

Notre choix : les deux premiers axes

Ils portent 77,3 %, et le troisième ne pèse que 4,8 % — du bruit.

On classera donc sur F_1 et F_2 . Avantage supplémentaire, et il n'est pas mince : la partition obtenue sera directement lisible sur le plan que nous savons déjà interpréter.

Ce qu'il faut retenir

Six points

1. Classer exige de **choisir une distance** — décision que l'ACP n'imposait pas.
2. Euclidienne pour les variables réduites, **khi-deux** pour les profils, **Jaccard** pour les présences-absences.
3. Une **dissimilarité** ne vérifie pas l'inégalité triangulaire; les jugements humains non plus.
4. Sur données **mixtes**, la meilleure voie est de classer sur les **axes factoriels**.
5. Trois raisons : **débruiter, décorréler, homogénéiser**.
6. Nous classerons sur F_1 et F_2 , soit **77,3 %** de l'information.

La suite

La distance est choisie. Reste à construire les groupes — et deux familles d'algorithmes s'y emploient, avec des logiques opposées.

On commence par celle qui ne demande pas de fixer le nombre de classes à l'avance.