Analyse des données multi-dimensionnelles

ECTS
5 crédits

Composante
Faculté des Sciences

Présentation

Description

Les données statistiques ne cessent de devenir plus massives. Préalablement à leur modélisation, il est indispensable de les explorer et d'en réduire la dimension en perdant le moins d'information possible. Tel est l'objectif de ce cours de statistique exploratoire multidimensionnelle. Sur le plan méthodologique, les outils qu'il utilise sont essentiellement ceux de la géométrie euclidienne. Les problèmes et notions statistiques y seront donc traduits dans le langage de la géométrie euclidienne avant d'être traités dans ce cadre. Les deux familles de méthodes exploratoires qui seront vues dans ce cours sont:

1) les méthodes de classification automatique, qui regroupent les observations en classes et réduisent leur disparité des observations à la disparité entre ces classes;

2) les méthodes d'analyse en composantes, qui recherchent les directions principales de disparité entre les observations et permettent de fournir de cette disparité des images interprétables en dimension réduite.

Objectifs

Faire le pont entre la géométrie euclidienne et la statistique exploratoire multidimensionnelle. Construire une compétence complète dans l'exploration des tableaux de données volumineuses et leur analyse préalable à la modélisation statistique.

Pré-requis nécessaires

Cours de géométrie euclidienne, d'espaces vectoriels normés et de réduction des endomorphismes.

Pré-requis recommandés : Cours de statistique descriptive univariée et bivariée. Bonne maîtrise du calcul matriciel.

Contrôle des connaissances

Contrôle continu (devoirs / miniprojets de maison) + contrôle terminal

Syllabus

I - Introduction :

a) Données multidimensionnelles, observations, variables, codages ; b) Traductions en nuages de points en espaces métriques euclidiens. c) Nécessité d'une réduction dimensionnelle : composantes / classes.

II - Écritures géométriques de quantités statistiques

Description univariée :

a) Moyenne, fréquence,

b) Variance et écart-type.

c) Centrage et réduction d'une variable.

Liaisons bivariées :

a) Liaison bivariée & conditionnement.

b) Covariance et corrélation de deux variables quantitatives.

c) R2 d'analyse de variance d'une variable quantitative sur une variable qualitative. d) Phi2 et T2 de deux variables qualitatives. e) Écriture unifiée des liaisons. f) Limites du bivarié & comment le dépasser.

III - Classification automatique

Dissemblance et ressemblance.

a) Mesures.

b) Ressemblance partielle vs globale.

Ressemblance partielle : classification logique/conceptuelle par treillis de Galois.

Ressemblance globale :

a) Partitionnement en espace métrique : méthode des K-means & raffinements.

b) Classification hiérarchique : indices, algorithme de CAH, critères de choix de partitions.

c) Classification mixte.

d) Interprétation des classes.

e) La classification sur variables.

IV - Analyses en composantes principales

ACP normée

a) Nuage des individus, inertie et ACP directes.

b) Nuage des variables, inertie et ACP duale.

c) Relations de dualité et interprétation jointe des graphiques.

d) Éléments supplémentaires & relation de dualité.

e) La première composante comme estimation d'une variable latente continue.

ACP générale (avec métriques quelconques)

a) Nuage des lignes et ACP directe.

b) Application au multidimensional scaling.

c) Quelle ACP des colonnes, pour quelles relations de dualité ?

d) Les aides à l'interprétation.

e) Éléments supplémentaires & relation de dualité.

f) Formule de reconstitution (décomposition en éléments singuliers).

Analyse des correspondances binaires

a) Le Phi2 comme inerties directe et duale des nuages de profils-lignes et de profils-colonnes.

b) Quelles métriques pour quelles relations de dualité : les positionnements barycentriques.

c) Interprétation jointe des graphiques.

d) Effet Guttman. E) Éléments supplémentaires.

Analyse des correspondances multiples.

a) Application de l'ACB à un tableau logique disjonctif complet.

b) Application de l'ACB à un tableau de Burt ; équivalence.

c) Relations barycentriques entre individus et modalités. Relations barycentriques entre modalités.

d) Effet Guttman. e) Éléments supplémentaires.

e) La première composante comme estimation d'une variable latente continue.

La pratique de l'ADM

a) Complémentarité de l'AF et de la CA.

b) Comment mener une bonne ADM.

Informations complémentaires

Volumes horaires :

CM : 21

TD : 21

TP : 0

Terrain : 0

Infos pratiques

Contacts

Responsable pédagogique

Xavier Bry

+33 4 67 14 35 78

xavier.bry@umontpellier.fr

Responsable pédagogique

Elodie Brunel-piccinini

+33 4 67 14 41 64

elodie.brunel-piccinini@umontpellier.fr