Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Séance 9. Les méthodes statistiques multivariées descriptives

Maxime Forriez[1,a]

[1] Institut de géographie, 191, rue Saint-Jacques, Bureau 105, 75 005 Paris,
[a]maxime.forriez@sorbonne-universite.fr

24 octobre 2025

1 Questions de cours

Les réponses comptent pour 25 % de la note finale du parcours « confirmés ».

1. Pourquoi la géographie a-t-elle si souvent recours à la statistique multivariée ?
2. Expliquez ce que vous avez compris des enjeux scientifiques des méthodes descriptives (un paragraphe par méthode).
3. Pour chaque méthode, résumez les éléments que vous devez calculer et mesurer.

   Attention ! Je vous demande une explication textuelle, et non mathématique, mais vous devez respecter les étapes de calcul.
4. Existe-t-il des points communs entre les différentes méthodes factorielles ? ou sont-elles mathématiquement indépendantes ?
5. Pour l'ensemble des méthodes, quels sont les tableaux de données existant ? Définissez-les et expliquez quelles méthodes les utilisant ?
6. Quel est le lien entre le choix de la méthode factorielle et le type de variables ? Selon vous, quelle est l'analyse factorielle la plus générale ? Justifier votre réponse.
7. Quelles différences existe-t-il entre une régression par les moindres carrés et une régression orthogonale ?
8. En analyse factorielle, que signifient les valeurs propres d'un axe d'un point de vue statistique ?
9. Qu'est-ce et à quoi sert la distance du χ^2 ?

10. Quelles différences entre analyse factorielle des correspondances et analyse des correspondances multiples, et entre analyse en composantes principales et analyse des correspondances multiples ?
11. À quoi servent les profils lignes et les profils colonnes
12. Quels sont les types de *mapping* ? Comment les interpréter ?
13. Comment expliqueriez-vous la généralisation de l'A.F.C. en A.C.M. ?
14. Comment distinguer classification et classement ? Expliquez les statistiques de regroupements ? Trouvez les situations d'utilisation de l'une ou de l'autre.
15. Construisez une explication de l'intérêt de faire une A.F.M. ? Pourquoi existe-t-il autant de modèle en A.F.M. ? Quelle est sa particularité par rapport à l'ensemble des méthodes factorielles précédentes ? Existe-t-il des objets `Python` permettant de les calculer ?
16. À quelle branche mathématique appartiennent les méthodes factorielles ?

2 Mise en œuvre avec Python

La sous-partie « Bonus » vous permet d'obtenir des points supplémentaires.

2.1 Objectifs

— Apprendre à utiliser la bibliothèque `Scikit-learn`

— Apprendre à utiliser la bibliothèque `Prince`

2.2 Manipulations

Le fichier obtenu compte pour 25 % de la note finale du parcours « confirmés ».

1. Faire une A.C.P. avec le fichier `france-temperatures.csv`.
   a. Récupérer les données.
   b. Isoler la colonne des individus `"Villes"` et isoler les colonnes numériques en utilisant la méthode `drop(columns = ["Villes"])`
   c. Centrer-réduire les données numériques en utilisant `StandardScaler()` et `fit_transform()`.
   d. Faire une A.C.P. avec 12 facteurs sur les données numériques centrées-réduites avec l'objet `PCA` et sa méthode `fit()`. L'afficher sur la console. Que constatez-vous ?
   e. Afficher la variance expliquée, la variance expliquée en pourcentage et calculer les valeurs propres.
   f. En utilisant `Pandas` et sa méthode `DataFrame`, afficher le résultat de l'A.C.P.
   g. Calculer les coordonnées des individus et faire un graphique permettant de créer l'image du *mapping* des individus (les deux premiers facteurs) dans un dossier `img`.

h. Calculer et afficher la contribution des individus aux facteurs. Calculer et afficher la qualité de la projection des individus (les cosinus carrés). Qu'en concluez-vous ?

i. Calculer les coordonnées des variables et créer l'image du cercle de corrélation associé dans un dossier `img`.

N.B. Vous pouvez vous aider du site https://fxjollois.github.io/cours-2019-2020/lp-iot–python-ds/seance2-ACP-classif.html.

2. Faire une A.C.M. avec le fichier `chiens.csv`.
   a. Récupérer les données.
   b. Isoler la colonne `Race` (les individus) et isoler dans un tableau `Pandas` les colonnes `Taille`, `Poids`, `Vitesse`, `Intelligence`, `Affection`, `Agressivité`, `Fonction`, `Origine` (les variables).
   c. Transformer le tableau des variables en tableau disjonctif complet (T.D.C.) en utilisant la méthode `Pandas` `get_dummies()`
   d. L'A.C.M. se calcule avec la bibliothèque `Prince` à partir du T.D.C. calculé par `Pandas` et dans ce format avec les méthodes `MCA` et `fit()`. Calculer une A.C.M. à huit facteurs.
   e. Calculer les valeurs propres.
   f. Calculer les coordonnées des lignes et des colonnes des variables.
   g. Créer une image du *mapping* des deux premiers facteurs dans un dossier `img`.
   h. Calculer les qualités des données des lignes et des colonnes (les cosinus carrés).

2.3 Bonus

Faire une C.A.H. à partir des résultats de l'A.C.P. étudiée.

Attention ! Il faudra vous documenter au-delà de la documentation fournie par le `GitHub`.