TD1 - Analyse en Composantes Principales sur les données de fertilité en Europe

T Nguyen

Présentation

On s'intéresse à la fertilité des femmes en Europe en 2012. Pour cela, on a construit un tableau de données avec 39 Pays d'Europe en lignes et en colonnes les tranches d'âge 15-19, 20-24, 25-29, 30-34, 35-39, 40 et plus. Dans une case du tableau on a la fertilité moyenne des femmes d'un pays pour une tranche d'âge. La fertilité correspond ici au nombre moyen d'enfants nés vivants pour 1000 femmes. Pour répondre à l'ensemble des questions, il vous faudra mettre en oeuvre une ACP à partir des données fournies dans le fichier "AnaDo_JeuDonnees_FertiliteEurope.csv"

Il vous est demandé de fournir un rapport répondant aux questions posées. Dans ce rapport, chaque étape de votre code devra être indiquée : chargement de la librairie de l'ACP, chargement des données, affichage des graphes et des données permettant de répondre aux questions, etc.

Pour fournir le rapport, vous pouvez au choix : compléter ce fichier et le compiler pour créer le PDF avec les réponses ; ou bien créer votre propre rapport répondant aux questions. Dans les deux cas, les étapes vous permettant de répondre aux questions devront apparaître (chargement des données, affichage des graphes, ...)

Questions

L'ensemble des questions vaut 19 points, plus un bonus de 0,5 points à la question 6. La qualité du rapport sera notée sur 2 points.

1) Importer les données et en faire l'ACP. Dans ce jeu de données, que représentent les individus et que représentent les variables ? (1.5p)

2) En analysant les deux graphes générés par l'ACP de la question 1), indiquer :

   a. Quelle est la variable la plus liée à la dimension 2 ? (0.5p)

   b. Quel pays a le taux de fertilité le plus élevé pour la classe 20-24 ans ? (1p)

   c. Que peut-on dire du taux de fertilité à tout âge des Danoises et des Hollandaises ? (1p)

   d. Que peut-on dire du taux de fertilité des Irlandaises comparativement aux autres pays d'Europe ? (1.5p)

   e. Citer quatre pays ayant un taux de fertilité très proche pour toutes les classes d'âge. (0.5p)

   f. Le taux de fertilité (pour toutes les classes d'âge) des Croates est-il proche de 0 ou proche du taux de fertilité moyen des pays d'Europe ? (1p)

   g. Que peut-on dire sur la fertilité des femmes de plus de 35 ans dans les pays où la fertilité des adolescentes est faible ? (2p)

3) En utilisant la fonction `summary`, indiquer le pourcentage d'inertie associé au plan principal. Ensuite, indiquer quel est le pays ayant le plus contribué à la création de l'axe 1, puis quels sont les trois pays ayant le plus contribué à la création de l'axe 2. Bien montrer la table des résultats ayant permis de répondre à cette question. (2.5p)

4) En utilisant la fonction `plot`, tracer le graphe des individus et des variables sur les dimensions 3 et 4. (1p)

5) En utilisant la fonction `dimdesc`, indiquer quelles sont les quatre variables ayant un taux de corrélation (positif ou négatif) de plus de 0.8 avec la première dimension. (1p)

6) Quels sont les 6 individus ayant le plus contribué à la construction du plan (sur les dimensions 1-2) ? (1.5p) Bonus : tracer un graphique pour répondre à cette question. (0.5p)

7) Tracer le cercle des corrélations sur les dimensions 1 et 2 en y affichant les deux variables ayant le plus contribué. (1p)

8) Faire l'ACP avec la variable supplémentaire "Région" puis tracer le graphe des individus sur les dimensions 1 et 2 avec le coloriage en fonction de la région. Sur ce graphe, ne sélectionner que les indivus qui ont une qualité de projection supérieure à 0.95. (2p) Comparer la région Europe de l'Est à la région Europe du Nord. (1p)

Template

Chargement de la librairie

Importation des données et affichage des graphes

On importe dans un premier temps les données de fertilité en Europe.

```
A COMPLETER
```

On fait l'ACP de ces données.

```
A COMPLETER
```

Interprétation de l'ACP

Pourcentage d'inertie du plan principal

```
APPEL A SUMMARY
```

Graphe sur les dimensions 3 et 4

```
A COMPLETER
```

Variables corrélées à la première dimension

```
APPEL A DIMDESC
```

Les quatre variables ayant un taux de corrélation de plus de 0.8 à la première dimension sont

Individus contribuant le plus à la création du plan

```
A COMPLETER
```

Les six individus ayant le plus contribué à la création du plan sont:

Cercle des corrélations

```
A COMPLETER
```

Graphique final

```
ACP avec la variable supplémentaire
```

```
AFFICHAGE DU GRAPHE
```