ANALYSE EN COMPOSANTES PRINCIPALES

Romain Raveaux
[1]Laboratoire L3I – Université de La Rochelle

Octobre 24-11, 2008

N.B. Tous les travaux seront notés. Vous enverrez par mail à romain.raveaux01 at univ-lr.fr dans les délais qui vous seront précisés en cours :

- le classeur de votre tableur contenant vos calculs pour la partie TD
- un rapport que vous nommerez "nom_prenom_proj2.odt".
- Dans le rapport (rédigé en binôme), vous noterez vos réponses aux questions posées et illustrerez vos propos d'extraits de votre tableur. Bon travail !

1 Partie TD

Afin d'analyser les interactions de certains indicateurs sociaux-économiques à l'échelle mondiale. Nous nous appuierons sur les données brutes fournies dans le fichier "monde_indicateurs.xls" (Données publiées sur le site du "Monde" en Juin 2008). Pour chacun des 146 pays, on donne les taux effectifs pour chacun des 13 indicateurs sociaux-économiques (en colonnes). L'objectif est d'analyser la structure des indicateurs. (Exemple: y-a-t il un lien entre le taux de mortalité infantile et le produit intérieur brut ? ou Peut on en conclure qu'un fort PIB engendre une amélioration du système de santé?) Pour analyser ces résultats, vous mettrez en place une Analyse en Composantes Principales. Le fichier de données est fourni. Il s'agit du fichier "monde_indictaeurs.xls" mis à votre disposition sur moodle.

1 Idée de l'ACP

1.1 Sur quel type de données peut-on réaliser une ACP?

1.2 Soit un tableau de données comprenant en ligne différents pays et en colonne différents indicateurs économiques. Quels sont les objectifs d'une ACP sur un tel jeu de données?

1.3 Quelles sont les variables du problème ?

1.4 Quelles sont les observations ?

1.5 Dans quel espace vivent les observations/individus?

1.6 Dessiner un poisson et tracer la droite maximisant la variance du nuage de points (premier axe factoriel).

1.7 (VRAI OU FAUX) La variance des coordonnées des individus sur le premier axe factoriel est plus élevée que la variance des coordonnées sur le deuxième axe ?

1.8 Exprimer avec vos mots la notion de "centrer et réduire" une série de données.

1.9 Que vaut la moyenne d'une série numérique centrée réduite ?

1.10 Que vaut l'ecart-type d'une série numérique centrée réduite ?

1.11 Quelle est la relation entre la variance et la corrélation d'une série de données centrée-réduite ?

2 Partie TP

2 : Départ

2.1 Calculer la moyenne de chaque variable.

2.2 Calculer la variance de chaque variable.

2.2 Calculer l'écart-type de chaque variable.

2.3 Créer une nouvelle feuille, "centrée réduite", contenant les valeurs centrées-réduites de vos données de départ.

2.4 Vérifier la moyenne et l'écart-type de vos variables.

3: Nos variables

3.1 Afficher l'histogramme de la variable "DEN99", Afficher l'histogramme de la variable "DEN99-centrée réduite". Comparer les deux histogrammes?

3.2 à l'aide de l'outil "nuage de points", comparer deux à deux les variables ? Peut-on dire de chaque pair de variables qu'il existe une relation linéaire ?

$\longrightarrow$ Nuage de points : Sélectionner en X l'ensemble des données (matrice) (avec le libellé des variables) et Sélectionner en Y l'ensemble des données (avec le libellé des variables)

4 : Matrice de similarité

4.1 Créer une nouvelle feuille "matrice covariance"

4.2 Calculer la matrice de covariances à partir de vos données centrées réduites.

- XLStat $\longrightarrow$ Description des données $\longrightarrow$ Matrices de similarité $\longrightarrow$ Type de proximité : variance(n)

4.3 Calculer la matrice de corrélations à partir de vos données initiales.

- XLStat $\longrightarrow$ Description des données $\longrightarrow$ Matrices de similarité $\longrightarrow$ Type de proximité : pearson(n)

5 Analyse en Composantes Principales.

5.1 Appliquer l'analyse en composantes sur vos données initiales.

- XLStat $\longrightarrow$ Analyse de données $\longrightarrow$ ACP
- Tableau observations/variables
- intégrer les libellés des observations : colonne "CODE"
- intégrer les libellés des variable : première ligne de vos données.
- Type d'ACP : Pearson(n)

5.2 Sélectionner 3 couples d'axes (F1-F2);(F1-F3);(F2-F3)

5.3 Combien d'axes d'axe faudrait il conserver pour exprimer $\geqslant 90\%$ de l'information ?

5.4 Quel quantité d'information est perdue en projetant nos données sur les 2 axes (F1-F2) ?

5.5 Calculer les coordonnées des variables dans le cercle des corrélations (F1-F2).

Exemple : $X_{URB00} \longrightarrow$

$$x1_{URB00} = cor(URB00(initiales), URB00 - F1)$$

,

$$x2_{URB00} = cor(URB00(initiales), URB00 - F2)$$

.

5.6 Calculer les cosinus entre chaque paire de variables:

$$Exemple : \ cos(X_{URB00}, X_{ARG00})$$

- XLStat $\longrightarrow$ Description des données $\longrightarrow$ Matrices de similarité $\longrightarrow$ Type de proximité : cosinus

5.7 Trouver pour chaque variable ses deux variables les plus proches?

5.8 Interpréter le résultat du cercle des corrélations ?

6 : Visualisation (MDS)

6.1 Visualiser la matrice des corrélations avec l'outil MDS : XLStat $\longrightarrow$ Analyse de données $\longrightarrow$ MDS

6.2 Comparer ces résultats avec vos interprétations issues de l'ACP ?

7 Synthèse

7.1 Rédiger un petit paragraphe synthétisant et expliquant les interactions entre les différents indicateurs socio-économiques en étayant votre discours à l'aide de chiffres.

7.2 Illustrer votre rapport de graphiques (nuages de points, cercles des corrélations).