ANALYSE FACTORIELLE POUR LA DESCRIPTION DE DONNEES

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_proj1.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

Les données à analyser sont les résultats du premier tour des élections présidentielles de 2007 (publiées dans "Le Monde" au lendemain du 22 Avril 2007). Pour chacune des 23 régions françaises (22 régions métropolitaines + 1 "région" Outremer), on donne les effectifs de suffrages pour chacun des 12 candidats (en colonnes). L'objectif est d'analyser la structure des votes ainsi que les liaisons entre candidats et régions. Pour analyser ces résultats, vous mettrez en place une Analyse Factorielle des Correspondances. Le fichier de données est fourni. Il s'agit du fichier "elections.xls" mis à votre disposition sur moodle.

1.1 ETUDE DESCRIPTIVE DU TABLEAU DE CONTINGENCE – PROFILS-LIGNES ET PROFILS-COLONNES

Ouvrez le fichiers "elections.xls" et faites-en une copie que vous nommerez TD_elections.xls.

1.1 Renommez la première feuille de calcul : appelez-la "Tableau des effectifs" et, dans cette feuille, rajoutez la colonne $n_{i.}$ (effectifs totaux des lignes, à placer en dernière colonne) et la ligne $n_{.j}$ (effectifs totaux des colonnes, à placer en dernière ligne).

1.2 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez "Tableau des fréquences totales", calculez le tableau des fréquences (totales). Dans chaque cellule, les fréquences sont calculées par

$$f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}} = \frac{Effectif\,de\,la cellule(i,j)}{Effectif\,total}$$

Dans cette même feuille, calculez les profils-lignes moyens f.j, que vous placerez en dernière ligne de votre tableau :

$$f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

Dans cette même feuille, calculez les profils-colonnes moyens fi., que vous placerez en dernière colonne de votre tableau :

$$f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}}$$

1.2 ETUDE DESCRIPTIVE DU TABLEAU DE CONTINGENCE – TAUX DE LIAISON, CONTRIBUTION AU χ^2

Le taux de liaison entre la ligne i et la colonne j peut être calculée comme suit :

$$t_{ij} = \frac{f_{ij} - f_{i\bullet}.f_{\bullet j}}{f_{i\bullet}.f_{\bullet j}}$$

2.1.1 Quel est le domaine de valeurs que peut prendre un taux de liaison ?

2.1.2 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez "Tableau des liaisons", calculez le tableau des taux de liaison.

2.1.3 Comment interprétez-vous ces taux de liaison ?

La contribution χ^2(i,j) de la cellule (i,j) au coefficient du χ^2 peut être calculée comme suit :

$$\chi^2 = \sum_{ij} = \frac{(f_{ij} - f_{i\bullet}.f_{\bullet j})^2}{f_{i\bullet}.f_{\bullet j}}$$

2.2.1 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez " Tableau des contributions ", calculez le tableau des contributions.

2.2.2 Comment interprétez-vous ces contributions ?

2.2.3 Quelle est la valeur du coefficient du χ^2 sur nos données (somme de toutes les contributions) ?

2.2.4 Quelle est la contribution moyenne des cellules (somme des contributions / nombre de cellules) ?

2.2.5 Quel est la cellule ayant la plus forte contribution au χ^2 ?

2.2.6 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez " Tableau des contributions (%) ", calculez le tableau des contributions exprimés en pourcentage de la valeur du coefficient du χ^2.

2.2.7 Quels sont les avantages et les inconvénients de l'utilisation des contributions au χ^2 et des taux de liaison pour l'analyse des données ?

1.3 ETUDE DESCRIPTIVE DU TABLEAU DE CONTINGENCE – TABLEAUX DE DISTANCES DU χ^2

3.1 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez "Tableau des fréquences lignes", calculez le tableau des fréquences-lignes. Dans chaque cellule, les fréquences-lignes sont calculées par:

$$fl_{ij} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}} = \frac{Effectif\,de\,la cellule(i,j)}{Effectif\,total\,de\,la\,ligne\,i}$$

3.2 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez " Tableau des fréquences colonnes ", calculez le tableau des fréquences-colonnes. Dans chaque cellule, les fréquences-colonnes sont calculées par:

$$fc_{ij} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}} = \frac{Effectif\,de\,la cellule(i,j)}{Effectif\,total\,de\,la\,colonnes\,j}$$

3.3 La distance du χ^2 sert à mesurer la distance entre lignes ou entre colonnes. Pour les lignes, elle est définie de la manière suivante :

$$d_{\chi^2}(L_1, L_2) = \sum_j \frac{(fl_{1j} - fl_{2j})^2}{f_{\bullet j}}$$

La distance du χ^2 entre profils-colonnes est similaire (en remplaçant les fl par des fc). Il est intéressant d'utiliser la distance du χ^2 plutôt que la distance Euclidienne pour deux raisons :

- La distance du χ^2 entre 2 lignes ne dépend pas du poids respectif des colonnes
- La distance du χ^2 vérifie la propriété d'équivalence distributionnelle : si on regroupe deux modalités lignes, les distances entre les profils-colonne, ou entre les autres profils-lignes restent inchangées.

3.3.1 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez "Tableau des distances entre lignes ", donnez le tableau des distances entre lignes. Attention, pour plus de simplicité vous devrez utiliser une formule matricielle dans votre tableur. Une fois le contenu de la cellule tapé, n'oubliez pas d'appuyer sur Maj+Ctrl+Entrée pour valider cette formule matricielle !

3.3.1.1 Quels sont les deux profils-lignes les plus proches au sens de la distance du χ^2 ?

3.3.1.2 Quel est le profil-ligne le plus éloigné du profil-ligne moyen selon la distance du χ^2 ?

3.3.2 Dans une nouvelle feuille de calcul du même classeur, que vous nommerez "Tableau des distances entre colonnes", donnez le tableau des distances entre colonnes. Attention, pour plus de simplicité vous devrez utiliser une formule matricielle dans votre tableur. Une fois le contenu de la cellule tapé, n'oubliez pas d'appuyer sur Maj+Ctrl+Entrée pour valider cette formule matricielle !

3.3.2.1 Quels sont les deux profils-colonnes les plus proches au sens de la distance du χ^2 ?

3.3.2.2 Quel est le profil-colonne le plus éloigné du profil-colonne moyen au sens de la distance du χ^2 ?

2 PARTIE TP

2.1 ELECTIONS PRESIDENTIELLES – ANALYSE FACTORIELLE DES CORRESPONDANCES

La méthode d'Analyse Factorielle des Correspondances peut être vue comme une sorte de double ACP (menée à la fois sur les lignes et les colonnes du tableau de contingence) et fournissant une décomposition pertinente de la matrice de variance-covariance des taux de liaison pondérés par les coefficients $f_{i\bullet}.f_{\bullet j}$

$$f_{i\bullet}.f_{\bullet j}t_{ij}^2$$

L'AFC permet d'obtenir de nouvelles variables, les facteurs, qui permettent une représentation aussi fidèle que possible de la répartition des modalités, et ce avec moins de variables que le nombre de variables initiales.

L'AFC peut être utilisée dans des situations variées, y compris sur des données qui ne constituent pas stricto sensu un tableau de contingence.

Cependant, il faut noter que l'AFC mettra toujours en évidence des attractions et des répulsions entre modalités lignes et modalités colonne, même si les variables étudiées sont indépendantes. En effet, lorsqu'on travaille sur un échantillon et que le χ^2 du tableau de contingence n'est pas significativement élevé ($p - value$ élevé), l'effet mis en évidence est simplement aléatoire et ne reflète aucune réalité. Il ne faut donc appliquer l'AFC que si le test d'indépendance du χ^2 est significatif ($p - value < 0,05$). Ici cet aspect ne sera pas traité. Nous considérons toujours que l'hypothèse nulle H0 d'indépendance peut être rejetée.

En vous aidant de l'aide en ligne d'XLstat, lancez l'AFC sur le tableau de contingence "election.xls".

5.1 Vérifiez que la valeur du $\{\chi^2 \times n_{\bullet\bullet}\}$ que vous avez obtenue dans la partie TD (question 2.2.3) est bien la même que celle qui vous est donnée par XLstat.

5.2 Vérifiez que la valeur du chi^2 que vous avez obtenue dans la partie TD (question 2.2.3) correspond bien à l'inertie totale su système donnée par XLstat.

5.3 Vérifiez que les valeurs des poids qu'XL stat associe à chacune des lignes i (colonne "poids ") correspondent bien à la ième coordonnée du profilcolonne moyen $f_{i.}$

5.4 Vérifiez que les valeurs des poids qu'XL stat associe à chacune des colonnes j correspondent bien à la jème coordonnée du profilligne moyen $f_{.j}$

5.5 Le nombre de valeurs propres produites par la recherche des facteurs principaux est égal au minimum du nombre de lignes et du nombre de colonnes du tableau de contingence. Cependant, la première valeur propre est systématiquement égale à 1, et n'est pas mentionnée dans les résultats. Les autres valeurs propres sont des nombres positifs inférieurs à 1 et leur somme est égale à la valeur du coefficient du χ^2. Pour le tableau de données considéré, combien de facteurs proposez-vous de conserver ?

5.6 Combien de facteurs faut il garder pour exprimer plus de 90% de l'information ?

5.7 Quels sont les régions qui ont le plus contribué à la formation des axes factoriels que vous avez retenu ?

5.8 Quels sont les candidats qui ont le plus contribué à la formation des axes factoriels que vous avez retenus ?

5.9 Représentez les profils-lignes et les profils-colonnes dans le plan constitué des deux premiers axes factoriels, puis des axes factoriels 2 et 3. Notez que, dans l'hyperplan factoriel constitué de tous les facteurs, la distance Euclidienne entre 2 profils-lignes (resp. entre 2 profilscolonnes) est égale à la distance du χ^2 entre les profils-lignes (resp. profils-colonne) initiaux. L'AFC fournit donc une représentation graphique permettant de comparer rapidement deux profils-lignes (resp. deux profils-colonnes).

5.9.1 Interprétez les résultats obtenus du point de vue des profils-lignes :

- à partir des Coordonnées principales (lignes) établir une matrice de dissimilarité.
- XLstat $\longrightarrow$ Description de données $\longrightarrow$ Matrice de dissimilarité $\longrightarrow$ Type de proximité : Euclidienne.
- Pour chaque région déterminer les deux régions les plus proches.

5.9.2 Interprétez les résultats obtenus du point de vue des profils-colonnes

- à partir des Coordonnées principales (colonnes) établir une matrice de dissimilarité.
- XLstat $\longrightarrow$ Description de données $\longrightarrow$ Matrice de dissimilarité $\longrightarrow$ Type de proximité : Euclidienne.
- Pour chaque candidats déterminer les deux candidats les plus proches.

5.9.3 La proximité entre un point-ligne L et un point-colonne C ne possède pas d'interprétation géométrique immédiate. En revanche, l'angle de sommet O (centre du plan factoriel) et dont les côtés passent par L et C a la propriété suivante :

- si l'angle (OL, OC) est aigu, la modalité-ligne L et la modalité colonne C s' "attirent ". Le cosinus de l'angle tend vers 1. (taux de liaison positif)
- si l'angle (OL, OC) est obtus, la modalité-ligne L et la modalité colonne C se "repoussent" (taux de liaison négatif). Le cosinus de l'angle tend vers -1.
- si l'angle (OL, OC) est droit, la modalité-ligne L et la modalité colonne C n'interagissent pas (taux de liaison voisin de 0).Le cosinus de l'angle tend vers 0.

Interprétez les résultats obtenus en croisant profils-lignes et profils-colonnes. Pour ce faire suivez la procédure suivante :

- Créer une nouvelle feuille : "région_candidat"
- Copier les coordonnées principales des régions suivant les axes F1-F2.
- à la suite copier les coordonnées principales des candidats suivant les axes F1-F2.
- XLstat $\longrightarrow$ Description de données $\longrightarrow$ Matrice de similarité $\longrightarrow$ Type de proximité : Cosinus.
- Pour chaque candidats déterminer la région la proches (tend vers 1).
- Pour chaque candidats déterminer la région la plus éloignée (tend vers -1).

5.9.4 Concluez en statuant sur les régions qui se démarquent des autres (elles sont isolées sur le grahpique)?

5.9.5 Concluez en statuant sur les candidants qui se démarquent des autres ?

5.9.6 Concluez en statuant sur les régions qui présentent des accointances (elles sont regroupées sur le grahpique)?

5.9.7 Concluez en statuant sur les candidats qui présentent des accointances ?

5.9.8 Que pouvez vous conclure des relations candidats-régions.