SY09 Printemps 2015

TP 1
Statistique descriptive, Analyse en composantes principales

1 Statistique descriptive

1.1 Le racket du tennis

Données

Début 2016, une série d'articles paraît dans la presse à propos de matchs de tennis dont le déroulement ou l'issue auraient été arrangés. Les journalistes se basent sur une série de données agrégées dont l'étude met en évidence un certain nombre de matchs suspects. Nous vous proposons d'étudier ces données dans le cadre de ce TP.

Le jeu de données considéré anonymous-betting-data.csv caractérise 129271 prises de position décrits par 16 variables. On commencera par télécharger le fichier de données et le script de pré-traitements à appliquer aux données brutes. On exécutera ensuite le code suivant :

```
books <- read.csv("anonymous-betting-data.csv")
source("pretraitements.R")
```

Les pré-traitements consistent, dans un premier temps, à déclarer les variables booléennes comme « logicals », et à renommer les variables qualitatives nominales. Dans un second temps, les paris correspondant à des prises de position atypiques sont supprimés : en effet, de telles prises de position ne sont pas nécessairement malveillantes, mais le retour à une cote « normale » en cours de match pourrait être considéré comme suspect. De même, les paris correspondant à des matches annulés sont supprimés. Deux variables sans intérêt pour notre étude sont laissées de côté. Les données restantes, stockées dans le tableau de données books.sel, feront l'objet de vos analyses.

Les variables disponibles dans ce tableau de données sont :

1. match_book_uid, match_uid, winner et loser : les identifiants du pari, du match, du gagnant, et du perdant ;
2. book : l'identifiant du bookmaker ;
3. year : l'année du match ;
4. odds_winner_open, odds_winner_close, odds_loser_open et odds_loser_close : les cotes[1] du joueur qui finalement gagnera le match, à l'ouverture et à la clôture des paris ; et les mêmes informations pour le joueur finalement perdant ;
5. implied_probs_winner_open, implied_probs_winner_close, implied_probs_loser_open et enfin implied_probs_loser_close : les probabilités de gain du match à l'ouverture et à la fermeture des paris, déduites des cotes précédentes ;
6. moved_towards_winner : une variable indiquant si la cote a évolué en faveur du joueur qui a finalement gagné le match.

1. Il est à noter que ce sont des « cotes *contre* » : ce nombre reflète la croyance que le joueur *perdra* le match.

Questions

On pourra adopter la stratégie d'étude suivante.

1. Faire une analyse descriptive générale des données, en répondant par exemple aux questions suivantes : combien de matches concernent-elles, combien de joueurs, sur quelle période de temps ? (Ces suggestions ne constituent évidemment pas une liste exhaustive.)
2. On s'intéressera ensuite aux joueurs. En particulier, on commencera par calculer le nombre de matches gagnés, perdus, et donc joués par chaque joueur. On pourra catégoriser les joueurs en fonction de leur propension à gagner les matches, et représenter l'information du nombre de matches joués (gagnés, perdus) en fonction du niveau du joueur.
3. Enfin, on cherchera à étudier plus particulièrement les matches suspects. Pour cela, on calculera les évolutions de probabilité de gain du match pour le gagnant et le perdant, ainsi que l'évolution de la probabilité en valeur absolue (probabilité du gagnant ou du perdant : cela revient au même, la somme des deux étant égale à 1).

   Un match sera considéré comme suspect si l'un des paris au moins présente une évolution de probabilité supérieure à 0.1 en valeur absolue (seuil de significativité fixé par les experts du domaine).

   On pourra se laisser guider par les questions suivantes :

   (a) combien y a-t-il de matches suspects ? Peut-on les caractériser ?

   (b) Quels sont les bookmakers impliqués dans ces matches ?

   (c) Quels sont les joueurs suspectés d'être associés à des malversations [2] ? Combien sont-ils ? Combien de joueurs sont impliqués dans un grand nombre de défaites [3] suspectes (plus de 10) ?

2. On pourra s'intéresser aux gagnants de matches pour lesquels la probabilité du perdant a évolué significativement, et de même aux perdants de matches caractérisés par une évolution significative de la probabilité du gagnant.

3. Il est plus facile d'influencer un résultat de match en le perdant contre toute attente, qu'en le gagnant contre toute attente...

1.2 Données crabs

Données

Le jeu de données considéré, disponible dans la bibliothèque de fonctions MASS, est constitué de 200 crabes décrits par huit variables (trois variables qualitatives, et cinq quantitatives). Charger le jeu de données et sélectionner les variables quantitatives en utilisant le code R suivant :

```
> library(MASS)
> data(crabs)
> crabsquant <- crabs[,4:8]
```

Questions

1. Effectuer dans un premier temps une analyse descriptive des données. Existe-t-il des différences de caractéristiques morphologiques selon l'espèce ou le sexe ? Semble-t-il possible d'identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs mesures de ces caractéristiques ?
2. Dans un second temps, on étudiera la corrélation entre les différentes variables. Quelle en est vraisemblablement la cause ? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène ?

2 Analyse en composantes principales

2.1 Exercice théorique

Trois variables mesurées sur quatre individus fournissent le tableau suivant :

$$\begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 4 & 1 & 2 \end{pmatrix}.$$

On associe les mêmes pondérations à tous les individus, et on munit $\mathbb{R}^p$ de la métrique euclidienne.

1. Calculer les axes factoriels de l'ACP du nuage ainsi défini. Quels sont les pourcentages d'inertie expliquée par chacun de ces axes ?
2. Calculer les composantes principales ; en déduire la représentation des quatre individus dans le premier plan factoriel.
3. Tracer la représentation des trois variables dans le premier plan factoriel.
4. Calculer l'expression $\sum_{\alpha=1}^{k} \mathbf{c}_\alpha \mathbf{u}'_\alpha$ pour les valeurs $k = 1, 2$ et 3. À quoi correspond cette somme lorsque $k = 3$?

2.2 Utilisation des outils R

L'objectif de cet exercice est de se familiariser avec les fonctions R permettant d'effectuer une ACP, en particulier les fonctions `princomp`, `summary`, `loadings`, `plot` et `biplot`. Remarquons qu'il existe une autre fonction `prcomp` qui effectue les calculs de manière différente ; on ne l'utilisera pas ici.

– En utilisant ces fonctions, effectuer l'ACP du jeu de données notes étudiées en cours. Montrer comment on peut retrouver tous les résultats alors obtenus (valeurs propres, axes principaux, composantes principales, représentations graphiques, ...).
– On s'intéresse à l'affichage des résultats de la fonction `princomp`. Qu'affichent les fonctions `plot` et `biplot` ? Détailler plus particulièrement le fonctionnement de la fonction `biplot` redéfinie pour la classe `princomp` (accessible par `biplot.princomp`) et de ses différentes options.

2.3 Traitement des données Crabs

Données

Comme dans l'exercice 1, on s'intéressera aux données crabs, et plus particulièrement aux descripteurs quantitatifs. On commencera donc par charger les données et sélectionner les variables quantitatives en utilisant le code R suivant :

```
> library(MASS)
> data(crabs)
> crabsquant<-crabs[,4:8]
```

Questions

Cette étude vise à utiliser l'ACP pour trouver une représentation des crabes qui permettent de distinguer visuellement différents groupes, liés à l'espèce et au sexe.

1. Tester tout d'abord l'ACP sur `crabsquant` sans traitement préalable. Que constatez-vous ? Comment pouvez-vous expliquer ce phénomène à la lumière des analyses menées au paragraphe 1.2 ?
2. Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes.