Analyse de données
M1 Statistique et économétrie - 2012
V. Monbet
Analyse canonique des correlations

A travers ce TD, nous allons apprendre à mettre en oeuvre l'analyse canonique des corrélations (en utilisant SAS et R), mais nous reverrons aussi les méthodes d'analyse multivariée étudiées précédemment.
Il est important de s'assurer à chaque étape qu'on sait interpréter les différents résultats obtenus et qu'on en sera capable de façon indépendante sur d'autres données.

1 Test psychologique

Un chercheur a recueilli des données sur trois variables psychologiques, quatre variables scolaires (notes obtenues à des tests normalisés) et le sexe pour 600 étudiants de collège. Il s'intéresse à la façon dont l'ensemble des variables psychologiques (série 1) est lié aux variables scolaires et au sexe. En particulier, le chercheur s'intéresse au nombre de dimensions nécessaires pour comprendre l'association entre les deux ensembles de variables.

Le fichier de données, `mmreg.csv`, contient 600 observations sur huit variables. Les variables psychologiques sont le locus de contrôle, la perception de soi et la motivation. Les variables scolaires sont des tests standardisés en lecture, écriture, mathématiques et sciences. En outre, la variable sexe est une variable définie sur $\{0, 1\}$, le 1 indiquant une étudiante.

1.1 Analyse descriptive

Lire les données et en faire une analyse descriptive. On pourra, par exemple, mener une première analyse descriptive multidimensionnelle en utilisant l'ACP. On cherchera en particulier à mettre en évidence des liens entre variables ou/et des individus aberrants.

```
mm <- read.table("mmreg.csv", sep = ",", header = TRUE)
head(mm)
xtabs(~female,data=mm)
```

1.2 Analyse canonique des correlations

Pour réaliser l'analyse canonique des corrélations, nous allons utiliser le package R `CCA`. De façon à simplifier son accès aux étudiants utilisant les postes de la salle, un fichier se trouvant sur la page web du cours regroupe l'ensemble des fonctions du package. Pour les utiliser, il suffit alors de taper sous R :

```
rep = "monrepertoire " ; # indiquer ici le chemin complet du répertoire
                         # où vous stockez le fichier CCA.R
source(paste(rep,"CCA.R",sep="")) ;
```

Pour les autres (ceux qui utilisent leur portable sous windows), installer le package puis taper `library(CCA)`.

1. Estimer et interpréter la matrice de corrélation des données complètes puis des deux groupes de données. D'abord indépendamment du sexe.

   ```
   # definition les deux ensembles de variables
   psych <- mm[,1:3]
   acad <- mm[,4:7]
   sex <- as.factor(mm[,8])
   pairs(psych)
   pairs(acad)
   matcor(psych,acad) # correlations
   ```

   Puis par sexe. Discuter les résultats obtenus.

   ```
   matcor(psych[sex==0,],acad[sex==0,]) # correlations pour les hommes
   matcor(psych[sex==1,],acad[sex==1,]) # correlations pour les femmes
   ```

2. Réaliser une analyse canonique à l'aide de la fonction `cc`. Interpréter les coefficients canoniques comme on interprète les coefficients d'un modèle de régression.

   ```
   cc1 <- cc(psych,acad)
   # Affichage des corrélations canoniques
   cc1$cor
   cor(cc1$scores$xscores,cc1$scores$yscores)
   # Affichage des coefficients canoniques
   cc1$xcoef
   lm(cc1$scores$xscores[,1]~locus_of_control+self_concept+motivation,data=mm)
   cc1$ycoef
   lm(cc1$scores$yscores[,1]~read+write+math+science,data=mm)
   ```

   The raw canonical coefficients are interpreted in a manner analogous to interpreting regression coefficients i.e., for the variable `read`*, a one unit increase in reading leads to a .0446 decrease in the first canonical variate of set 2 when all of the other variables are held constant. Here is another example : being female leads to a .6321 decrease in the dimension 1 for the academic set with the other predictors held constant.*

   Quand les variables ont des variances très différentes, les coefficients normalisés peuvent être plus faciles à interpréter.

   ```
   # coefficients normalisés pour psych
   sd<-sd(psych)
   s1<-diag(sd)
   (cc1$xcoef.norm <- s1 %*% cc1$xcoef)
   ```

```
# coefficients normalisés pour acad
sd<-sd(acad)
s2<-diag(sd)
(cc1$ycoef.norm <- s2 %*% cc1$ycoef)
```

3. Interpréter ensuite les corrélations entre les variables et les axes canoniques. On peut s'aider des valeurs et des graphiques. En vous aidant des instructions ci-dessous tracer la projection des variables sur le 1er et 2nd plan factoriel, puis celle des individus.

```
 dim1 = 1 ; dim2 = 2 ;
plot(NULL,xlim=c(-1.5,1.5),ylim=c(-3,3),
   xlabel=paste("Dim",1,sep=" "),ylabel=paste("Dim",1,sep=" "))
text(cc1$xcoef[,dim1],cc1$xcoef[,dim2],
   labels=c("locus","self","motiv"))
text(cc1$ycoef[,dim1],cc1$ycoef[,dim2],
   labels=c("read","write","math","science"),col="red")
dev.new()
plot(NULL,xlim=c(-1.5,1.5),ylim=c(-3,3))
text(cc1$xcoef.norm[,dim1],cc1$xcoef.norm[,dim2],
   labels=c("locus","self","motiv"))
text(cc1$ycoef.norm[,dim1],cc1$ycoef.norm[,dim2],
   labels=c("read","write","math","science"),col="red")
```

4. Recommencer les analyses précédentes en séparant des hommes des femmes. Les résultats sont-ils différents ? Commenter les résultats obtenus.

5. On peut aussi utiliser la fonction `MFA` du packge `FactoMineR`. On appelle, par exemple,

```
mfa <- MFA(mm[,1:7],group=c(3,4),type = c("c","c"),
   name.group=c("psy","acad"))
   barplot(mfa$eig[,1],main="Eigenvalues",names.arg=1:nrow(mfa$eig))
mm[,8]<-factor(mm[,8])
mfa.sex <- MFA(mm[,1:8],group=c(3,4,1),type = c("c","c","n"),
   name.group=c("psy","acad","sex"))
```

2 Valeur de différents modèles de voiture

En vous aidant de l'exercice précédent, faire l'analyse canonique des corrélation de la table de données `cars.tex` (source : Härdle, Simar, *Applied Multivariate Statistical Analysis*). Ces données sont issues d'une enquête au près de 40 personnes qui ont noté les caractéristiques suivante pour différents modèles de voitures : Economique, Equipement, Design, Sportif, Sécurité , Facileité à conduire. On considérera pour premier groupe de variable le prix et la valeur de la voiture et pour second groupe les autres variables numériques. En déduire, en justifiant vos réponses, que

– le prix et la valeur sont inversement corrélés ;
– les deux groupes de variables ne sont pas indépendants ;
– le premier axe canonique peut être interpréter comme un indice de prix et valeur de la voiture ;
– le second axe canonique est principalement form par les caractéristiques du modèle (ces variables peuvent donc être interprétées comme une appréciation de la valeur de la voiture) ;
– certaines variable ont un effet négatif sur le prix (les quelles ?).

3 Enquête de satisfaction auprès d'employés de commerces

On propose dans cette seconde partie d'analyser les relations entre des caracteristiques liées au poste de travail d'employés de commerces de détail et des variables liées à la satisfaction.

On ne dispose pas ici de données brutes mais uniquement de la matrice de corrélation de l'ensemble des variables estimée à partir de 784 individus. Un exemple de commandes SAS est proposé ci-dessous.

1. Quel(s) type(s) de graphiques ne pourra-t-on pas réaliser ici ?
2. Quel est le nombre maximum d'axes canoniques que l'on peut considérer ?
3. Réaliser une analyse canonique des corrélations. Expliquer comment on sélectionne le modèle.
4. Une question plus difficile (à faire à la fin du TD ?) : en s'inspirant des macro `gacpvx` et `gacpix` (ou de l'aide en ligne de l'instruction `ANNOTATE`) projeter les variables sur le premier plan factoriel.
5. Interpréter les résultats. Répondre, par exemple, aux questions suivantes en justifiant les réponses :
   (a) Quelles sont les deux variables de travail qui sont le plus fortement liées au 1er (resp 2nd) axe canonique de satisfaction.
   (b) Quelles sont les deux variables de satisfaction qui sont le plus fortement liées au 1er (resp 2nd) axe canonique du travail.
   (c) Quelles variables contribuent au troisième axe factoriel. Est-ce significatif ?
   (d) Certaines variables n'apportent-elles aucune information utile ?

   Aurait-on pu utiliser une autre méthode pour mettre en évidence des liens entre les variables ? Laquelle ? Mettre cette méthode en oeuvre et interpréter les résultats. Quels sont les avantages et inconvénient des deux méthodes utilisées ici ?

```
title 'Canonical correlation (CORR matrix input)';
data jobsat(type=corr);
   input _name_ $1-6 _type_ $8-11 x1-x5 y1-y7;
   label x1='Feedback' x2='Task Significance'
 x3='Task Variety' x4='Task Identity' x5='Autonomy'
         y1='Supervisor Satisfaction'  y2='Career-Future Sat'
```

```
        y3='Financial Satisfaction' y4='Workload Satisfaction'
        y5='Company Satisfaction' y6='Kind-of-work Sat'
        y7='General Satisfaction';
/*  Note:
  The _NAME_ values *must* match the variable names for valid
  correlation matrix input.  Only the lower half need be entered. */
list;datalines;
X1     CORR 1   .   .   .   .   .   .   .   .   .   .   .
X2     CORR .49 1   .   .   .   .   .   .   .   .   .   .
X3     CORR .53 .57 1   .   .   .   .   .   .   .   .   .
X4     CORR .49 .46 .48 1   .   .   .   .   .   .   .   .
X5     CORR .51 .53 .57 .57 1   .   .   .   .   .   .   .
Y1     CORR .33 .30 .31 .24 .38 1   .   .   .   .   .   .
Y2     CORR .32 .21 .23 .22 .32 .43 1   .   .   .   .   .
Y3     CORR .20 .16 .14 .12 .17 .27 .33 1   .   .   .   .
Y4     CORR .19 .08 .07 .19 .23 .24 .26 .25 1   .   .   .
Y5     CORR .30 .27 .24 .21 .32 .34 .54 .46 .28 1   .   .
Y6     CORR .37 .35 .37 .29 .36 .37 .32 .29 .30 .35 1   .
Y7     CORR .21 .20 .18 .16 .27 .40 .58 .45 .27 .59 .31 1
       N    784 784 784 784 784 784 784 784 784 784 784 784
;
proc cancorr data=jobsat
     redundancy smc    /* Squared Multiple Rs */
     ncan=2            /* Print only 2 canonical variates */
     vprefix = job   vname='Job characteristics'
     wprefix = sat   wname='Satisfaction measures';
   var  x1-x5;
   with y1-y7;
   run;
```