Analyse factorielle - Rappels et calculs matriciels

François Husson

Unité pédagogique de mathématiques appliquées - l'institut Agro

husson@agrocampus-ouest.fr

Analyse factorielle - Rappels et calculs matriciels

Rappels matriciels

Soit $\vec{u}$ et $\vec{v}$ deux vecteurs de taille p.
Le produit scalaire entre $\vec{u}$ et $\vec{v}$ s'écrit :

$$< \vec{u}, \vec{v} >= \|\vec{u}\| \ \|\vec{v}\| \cos(\vec{u}, \vec{v}) = u'v = u_1 v_1 + u_2 v_2 + ... + u_p v_p$$

En introduisant une métrique M, on écrit :

$$< \vec{u}, \vec{v} >_M = u' M v = m_1 \times u_1 v_1 + m_2 \times u_2 v_2 + ... + m_p \times u_p v_p$$

Rappels matriciels

Soit X une matrice centrée et n son nombre de lignes (centrer avec la fonction `scale`)
Vous utiliserez le jeu de données iris de R restreint aux 10 premières lignes et 4 premières colonnes

$X'X$ est la matrice ? ? ?

XX' est la matrice ? ? ?

Rappels matriciels

Soit X une matrice centrée et n son nombre de lignes (centrer avec la fonction `scale`)
Vous utiliserez le jeu de données iris de R restreint aux 10 premières lignes et 4 premières colonnes

$X'X$ est la matrice des covariances (au n près)

XX' est la matrice des produits scalaires

Analyse factorielle - Rappels et calculs matriciels

Décomposition en valeurs singulières

Soit $X_{(n,p)}$ une matrice, la SVD de X revient à trouver les matrices $U_{(n,r)}$, $\Lambda_{(r,r)}$ et $V_{(p,r)}$ telles que :

$$X = U\Lambda V' \quad \text{avec} \quad UU' = U'U = Id_n \quad \text{et} \quad VV' = V'V = Id_p$$

Exercice : Soit X une matrice centrée et n le nombre de lignes

- Diagonaliser (à l'aide de la fonction `eigen`) $COV = X'X$ et conserver valeurs propres et vecteurs propres
- Diagonaliser $PS = XX'$ et conserver valeurs propres et vecteurs propres
- Comparer les valeurs propres de COV et de PS
- Calculer la décomposition en valeurs singulières de X, i.e. les matrices U, Λ et V telles que $X = U\Lambda V'$ (fonction `svd`)
- Par rapport aux valeurs propres et vecteurs propres de COV et PS, que sont U, Λ et V ?

Formule de reconstitution

Calculer $\hat{X}^{(j)}$ la reconstitution des données en prenant uniquement les j premières colonnes de U et V et les j premières valeurs de Λ.

Calculer la somme des carrés des écarts entre X et $\hat{X}^{(2)}$.

Faire de même en faisant varier le nombre de colonnes conservées dans U et V. Que pouvez-vous dire ?

Analyse factorielle - Rappels et calculs matriciels

Lien avec l'ACP

Calculer XV et comparer avec les coordonnées des individus d'une ACP non normée (argument `scale.unit=FALSE` dans la fonction `PCA` de `FactoMineR`)

Calculer $X'U$ et comparer avec les coordonnées des variables d'une ACP non normée (en ACP le poids des individus est de $1/n$ avec n le nombre d'individus)

Fiche récapitulative de l'ACP

- Quels tableaux de données ? Exemples de jeu de données ?
- Quels objectifs ?
- Comment interpreter ?
- Comment considérer des individus supplémentaires ?
- Comment prendre en compte des variables qualitatives ?
- Comment prendre en compte des variables quantitatives supplémentaires ?
- Quelle différence entre ACP normée et non normée ?
- Dans un tableau avec 1 variable qualitative, les axes de l'ACP obtenus sur le tableau individus $\times$ variables quantitatives sont-ils identiques à ceux obtenus à partir des moyennes par modalité, moyennes pondérées par l'effectif de la modalité ? Donner un contre-exemple, expliciter les différences d'objectif OU démontrer l'égalité.

L'ACP normée

Avec $M = diag(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \dots \frac{1}{\sigma_p^2})$ et N la matrice diagonale des poids des lignes $(1/n)$, les valeurs propres et vecteurs propres de $ZMZ'N$ et $Z'NZM$ donnent les résultats de l'ACP normée :

```
data(iris)
X <- as.matrix(iris[1:10,1:4])
X <- scale(X,scale=FALSE) ## centrer
n <- nrow(X)   ;   N <- diag(rep(1/n,n))
M <- diag(1/(apply(X,2,var)*(n-1)/n)) ## métrique pour normer

diagPS <- eigen(X %*% M %*% t(X) %*% N)
SVD <- svd(N^0.5 %*% X  %*%  M^(0.5))
SVD$u/diagPS$vectors[,1:4]

pca <- PCA(X,gr=F)
pca$eig[,1] ; diagPS$values[1:4]; SVD$d^2

coordInd <- SVD$u[,1:4]  %*%  diag(SVD$d)
coordInd*sqrt(n)/ pca$ind$coord
coordInd2 <- N^0.5 %*% X %*% M^0.5 %*% SVD$v
coordInd2 / pca$ind$coord

coordVar <- SVD$v  %*%  diag(SVD$d)
coordVar[,1:4]/ pca$var$coord

coordVar2 <- (M^0.5%*%t(X)%*%N^0.5%*%SVD$u)
coordVar2[,1:4]/pca$var$coord
```

Analyse factorielle - Rappels et calculs matriciels

Fiche récapitulative sur l'AFC

- Quels types de tableaux de données ? Exemple de jeu de données ?
- Quels objectifs ?
- Comment interpreter ?
- Considérer le jeu de données Nobel avec le code suivant :

```
fichier <- "https://husson.github.io/MOOC_AnaDo/AnaDo_JeuDonnees_Nobel_avecMaths.csv"
Nobel <- read.table(fichier, header=TRUE, sep=";", row.names=1, check.names=FALSE)
Nobel <- Nobel[1:8,]
```

Comparer les objectifs et les résultats de l'ACP et ceux de l'AFC sur ce jeu de données. Bien expliciter la différence.

L'AFC comme une SVD

1. $P = N/n$, D_r et D_c les matrices diagonales des marges lignes et colonnes de P
2. Calcul la matrice des résidus standardisés :
   $S = D_r^{-1/2}(P - rc')D_c^{-1/2}$
3. Calcul la SVD : $S = U\Lambda V'$
4. Coordonnées des lignes : $F = D_r{-}1/2U\Lambda$
5. Coordonnées des lignes : $G = D_c{-}1/2V\Lambda$

```
library(FactoMineR) ; data(children) ; X <- as.matrix(children[1:8,1:5])
ca <- CA(X, graph=FALSE)

P = X/sum(X)
r = apply(P,1,sum)
c = apply(P,2,sum)
invDr=diag(1/r)
invDc=diag(1/c)

S = invDr^0.5 %*% (P-r %*% t(c)) %*% invDc^0.5
res <- svd(S)
F <- invDr^0.5 %*% res$u[,1:4] %*% diag(res$d[1:4])
F/ca$row$coord

G <- invDc^0.5 %*% res$v[,1:4] %*% diag(res$d[1:4])
G/ca$col$coord
```

Analyse factorielle - Rappels et calculs matriciels

1. Rappels matriciels
2. SVD
3. ACP
4. AFC
5. ACM

Fiche récapitulative sur l'ACM

- Quels types de tableaux de données ? Exemple de jeu de données ?
- Quels objectifs ?
- Comment interpreter ?
- Considérer le jeu de données `tea` du package `FactoMineR` et faire l'ACM avec sur le tableau avec uniquement les variables 14 et 18. Puis faire l'AFC sur le tableau de contingence croisant ces 2 variables.

```
library(FactoMineR)
data(tea)
don <- tea[, c(14,18)]
TabCont <- table(don)
```

Comparer les objectifs et les résultats de l'ACM et ceux de l'AFC sur ce jeu de données. Bien expliciter la différence.

L'ACM comme une SVD

Soit Z le tableau disjonctif de X, centré par colonne.
En posant $M = diag(\frac{n}{n_1}, \frac{n}{n_2}, \ldots \frac{n}{n_l})$ avec n_k le nombre d'occurrences de la modalité k, et N la matrice diagonale du poids de chaque ligne $(1/n)$, les valeurs propres et vecteurs propres de $ZMZ'N$ et $Z'NZM$ donne les valeurs propres et vecteurs propres de l'ACM qui permettent de calculer les coordonnées des individus et des modalités (à des constantes près) :

```
library(FactoMineR)
data(tea)
X <- tea[,10:18]
n <- nrow(X)
Z <- tab.disjonctif(X)
M <- diag(n/apply(Z,2,sum))
Z <- scale(Z, scale=FALSE)
N <- diag(rep(1/n,n))
U <- eigen(Z %*% M %*% t(Z) %*% N)
V <- eigen(t(Z) %*% N %*% Z %*% M)
mca <- MCA(X,graph=FALSE)

mca$eig[1:6,1] * ncol(X) / U$val[1:6]
U$vect[,1:4]%*%diag(sqrt(U$val[1:4]))/mca$ind$coord[,1:4]*sqrt(n/ncol(X))  # F = ZMV = U Lambda
M%*%V$vect[,1:4]%*%diag(sqrt(U$val[1:4]))/mca$var$coord[,1:4]            #G =  MV Lambda
```