Analyse Multivariée

TP n°2: Analyse Discriminante PLS exploratoire

I - Théorie

On considère deux matrices de données décrivant n individus (en ligne) : $X=[x^1,\dots,x^p]$ codant p variables numériques centrées ; $Y=[y^1,\dots,y^q]$ codant une variable qualitative à q modalités par ses indicatrices non centrées. On note $W=diag(w_i;i=1,\dots,n)$ la matrice des poids des individus. L'espace $\mathbb{R}^p$ est muni de la métrique d'ACP de X, notée M.

On rappelle que l'ADPLS est fondée sur la maximisation par chaque composante $f=XMu$ du critère produit : $\|f\|_W^2 R^2(f,Y)$ sous la contrainte $u'Mu=1$.

1. Montrez que : $\|f\|_W^2 R^2(f,Y)=\|\hat{X}Mu\|_W^2$, où $\hat{X}=\Pi_Y X$.

2. Programme de rang 1 :

a) Soit la matrice $E=\hat{X}'W\hat{X}$. Interprétez-la.

b) Soit le programme : $\max\limits_{u'Mu=1}\|\hat{X}Mu\|_W^2$, dont la solution fournit la première composante f^1.
Montrez que $f^1=XMu_1$, où u_1 est le vecteur propre M-unitaire de EM associé à sa plus grande valeur propre.

b) Montrez que si l'on pose $u^*=M^{1/2}u$, $X^*=XM^{1/2}$ et $\hat{X}^*=\Pi_Y X^*$, u_1^* est le vecteur propre I-unitaire de la matrice *symétrique* $E^*=\hat{X}^{*\prime}W\hat{X}^*$ associé à sa plus grande valeur propre. Montrez qu'alors, $f^1=X^*u_1^*$.

3. Programme de rang h :

On désire obtenir des composantes deux à deux orthogonales. On note $F^{h-1}=[f^1,\dots,f^{h-1}]$. La h ième composante f^h doit donc vérifier la contrainte d'orthogonalité : $F^{h-1\prime}Wf^h=0$.

a) Montrez que la résolution du programme $\underset{\substack{u'Mu=1\\ D'Mu=0}}{Max}\; u'MEMu$, où $D'=F^{h-1\prime}WX$, conduit à rechercher u solution de :

$$\Pi_{D^\perp}EMu=\lambda u \quad (5), \text{ où } \Pi_{D^\perp}=I-D(D'MD)^{-1}D'M \text{ et } \lambda \text{ maximale.}$$

b) Montrez qu'alors, $u\in\langle D^\perp\rangle$. Déduisez-en que u est de façon équivalente solution de :

$$\Pi_{D^\perp}EM\Pi_{D^\perp}u=\lambda u \text{ associé à } \lambda \text{ maximale.}$$

c) Montrez que cette dernière équation équivaut à :

$$M^{1/2}\Pi_{D^\perp}\hat{X}'W\hat{X}\Pi_{D^\perp}'M^{1/2}u^*=\lambda u^*,$$

qui caractérise la diagonalisation d'une matrice *symétrique*.

4. Pour toute composante discriminante f, interprétez les deux indicateurs suivants :

$$S(f)=\frac{\|f\|_W^2}{tr(X'WX)} \quad ; \quad R^2(f,Y)=\frac{\|\hat{X}Mu\|_W^2}{\|f\|_W^2}$$

5. Représentations graphiques :
Dans le plan (h, m) direct, l'individu i sera représenté par ses coordonnées selon les composantes *réduites* : $(\tilde{f}_i^h, \tilde{f}_i^m)$.
On veut également représenter dans les plans discriminants les centres de gravité des classes de Y correspondant à ses q modalités. Montrez que ces centres de gravité ont pour coordonnées sur les H premiers axes discriminants :

$$(Y'WY)^{-1} Y'W \tilde{F}^H \quad \text{, où} \quad \tilde{F}^H = [\tilde{f}^1, \dots, \tilde{f}^H]$$

Dans le plan dual, la variable x^j sera représentée par ses corrélations avec les composantes discriminantes : $\left(\frac{\langle x^j | f_i^h \rangle_W}{\|x^j\|_W \|f^h\|_W}, \frac{\langle x^j | f_i^m \rangle_W}{\|x^j\|_W \|f^m\|_W} \right)$.

II - Programmation

1. Programmez le calcul des H premières composantes de l'ADPLS.

2. Programmez le calcul des indicateurs $S(f)$ et $R^2(f, Y)$ pour ces H premières composantes.

3. Programmez le calcul des coordonnées des centres de gravité des q classes sur les H axes discriminants.

4. Programmez le calcul des coordonnées des variables de X .

5. Programmez l'affichage des individus et des centres de gravité de classes dans un plan (h, m) choisi par l'utilisateur. Programmez l'affichage des variables dans ce plan (h, m) et faire figurer le cercle unité sur ce graphique.

III - Application: types forestiers du bassin du Congo

1. Chargez le fichier *genus*. Procédez, avec votre programme, à l'ADPLS exploratoire de la variable Y = *forest* sur les variables de composition arborée X = [*gen1*, ... , *gen27*], en centrant-réduisant ces dernières variables au préalable et en choisissant la métrique M adaptée .

2. Interprétez soigneusement les résultats obtenus à partir des indicateurs et des graphiques. Prenez soin de dépasser l'interprétation isolée de chaque composante en interprétant les plans entiers.