TP2 : Analyse Discriminante PLS exploratoire (ADPLS)

Thomas Anne - laure

2025-11-05

Introduction

L'Analyse Discriminante PLS exploratoire (ADPLS) est une méthode multivariée qui combine les principes de la régression PLS (Partial Least Squares) et de l'analyse discriminante.
Son objectif est de construire des composantes linéaires des variables explicatives X maximisant la capacité de discrimination des groupes définis par la variable qualitative Y.

Autrement dit, on cherche à extraire des axes factoriels $f = XMu$ expliquant à la fois une grande part de la variance de X (inertie totale) et une forte part de la variance expliquée par les classes de Y.
Les notations utilisées sont celles du cours :

- $X \in \mathbb{R}^{n\times p}$: matrice des variables quantitatives centrées ;
- $Y \in \mathbb{R}^{n\times q}$: matrice d'indicatrices des q classes (non centrée) ;
- $W = \mathrm{diag}(w_i)$: matrice diagonale des poids individuels ;
- M : métrique définie positive sur l'espace des variables ;
- $f = XMu$: composante discriminante associée à la direction u ;
- $\Pi_Y = Y(Y'WY)^{-1}Y'W$: projecteur sur l'espace engendré par Y ;
- $\widehat{X} = \Pi_Y X$: projection de X sur cet espace.

Partie 1

1. Première identité : $\|f\|_W^2 R^2(f,Y) = \|\widehat{X}Mu\|_W^2$

Par définition, la proportion de variance de la composante f expliquée par Y est donnée par :

$$R^2(f,Y) = \frac{\|\Pi_Y f\|_W^2}{\|f\|_W^2}.$$

Or,

$$\Pi_Y f = \Pi_Y(XMu) = (\Pi_Y X)Mu = \widehat{X}Mu.$$

Ainsi, on obtient immédiatement :

$$\boxed{\|f\|_W^2 R^2(f,Y) = \|\widehat{X}Mu\|_W^2.}$$

Cette identité exprime que le produit de la norme pondérée de f par sa qualité de prédiction $R^2(f,Y)$ correspond à la norme pondérée de la projection de X sur l'espace des classes, transformée par Mu.
Elle constitue le point de départ du critère d'optimisation de l'ADPLS.

2. Programme de rang 1

On considère la matrice :

$$E = \widehat{X}'W\widehat{X}.$$

2-a) Interprétation de E

La matrice E représente la **matrice d'inertie inter-classes** ou la **covariance des centres de gravité des classes** projetés.
Si m_k désigne le centre de gravité de la classe k et m la moyenne globale, on peut écrire :

$$E = \sum_k n_k (m_k - m)(m_k - m)'.$$

Les valeurs propres de E mesurent la **dispersion inter-classes** dans les différentes directions, et permettent d'identifier les axes de plus grande discrimination.

2-b) Programme d'optimisation

On cherche le vecteur u maximisant la part de variance expliquée par les classes :

$$\max_{u'Mu=1} \|\widehat{X}Mu\|_W^2 = \max_{u'Mu=1} u'(MEM)u.$$

C'est un **problème de Rayleigh**, dont la condition d'optimalité conduit à :

$$EMu = \eta u,$$

où u est le **vecteur propre** M-unitaire associé à la plus grande valeur propre η.
La première composante discriminante s'écrit alors :

$$f_1 = XMu_1.$$

2-c) Symétrisation du problème

On introduit :

$$u^* = M^{1/2}u, \qquad X^* = XM^{1/2},$$

d'où $\widehat{X}^* = \Pi_Y X^*$.

Ainsi,

$$E^* = (\widehat{X}^*)'W\widehat{X}^* = M^{1/2}EM^{1/2}.$$

L'équation propre devient :

$$E^* u^* = \eta u^*.$$

La matrice E^* est **symétrique définie positive**, et u^* est un vecteur propre euclidien associé à la plus grande valeur propre η.
La première composante peut donc s'écrire :

$$\boxed{f_1 = X^* u_1^*.}$$

3. Programme de rang h

À partir de la deuxième composante, on impose la contrainte d'orthogonalité :

$$F'_{h-1} W f_h = 0,$$

où $F_{h-1} = [f_1, \dots, f_{h-1}]$.

On pose également :

$$D' = F'_{h-1} W X.$$

Le problème d'optimisation devient alors :

$$\max_{u'Mu=1,\, D'Mu=0} u'MEMu.$$

3-a) Équation propre restreinte

L'écriture du lagrangien et les conditions d'optimalité conduisent à l'équation :

$$\Pi_{D^\perp} EMu = \lambda u,$$

avec

$$\Pi_{D^\perp} = I - D(D'MD)^{-1} D'M,$$

qui est le **projecteur sur le sous-espace orthogonal à** D pour la métrique M.

3-b) Sous-espace orthogonal

Puisque $u \in \langle D^\perp \rangle$, on peut écrire l'équation sous la forme :

$$\Pi_{D^\perp} EM \Pi_{D^\perp} u = \lambda u.$$

3-c) Forme symétrique

En posant $u^* = M^{1/2} u$, on obtient l'équation propre symétrique :

$$\boxed{M^{1/2} \Pi_{D^\perp} \widehat{X}' W \widehat{X} \Pi'_{D^\perp} M^{1/2} u^* = \lambda u^*.}$$

Cette formulation assure que la matrice à diagonaliser est **symétrique**, ce qui garantit l'orthogonalité des composantes extraites.

4. Indicateurs de qualité des composantes

Pour chaque composante $f = XMu$, on définit :

$$S(f) = \frac{\|f\|_W^2}{\mathrm{tr}(X'WX)}, \qquad R^2(f, Y) = \frac{\|\widehat{X}Mu\|_W^2}{\|f\|_W^2}.$$

- $S(f)$ mesure la **part de l'inertie totale** de X expliquée par la composante f.
- $R^2(f, Y)$ mesure la **part de la variance de** f expliquée par les classes Y.
- Le produit $S(f)\, R^2(f, Y)$ représente le **pouvoir discriminant** de la composante.

Ces indicateurs permettent d'évaluer la qualité globale du modèle ADPLS et l'importance relative de chaque axe discriminant.

5. Représentations graphiques

Les résultats de l'ADPLS peuvent être visualisés sous deux formes complémentaires :
le **plan des individus et des centres de gravité des classes**, et le **plan des variables**.

5-a) Centres de gravité des classes

Les coordonnées des q centres de gravité des classes sur les H premiers axes discriminants sont données par :

$$\boxed{(Y'WY)^{-1}Y'W\widetilde{F}_H,}$$

où $\widetilde{F}_H = [\tilde{f}_1, \ldots, \tilde{f}_H]$ contient les composantes réduites.
Ces points résument la position moyenne de chaque classe dans l'espace discriminant.

5-b) Représentation des variables

Chaque variable x_j peut être représentée dans le plan (h, m) par ses corrélations avec les composantes :

$$\left(\frac{\langle x_j, f_h \rangle_W}{\|x_j\|_W \, \|f_h\|_W}, \; \frac{\langle x_j, f_m \rangle_W}{\|x_j\|_W \, \|f_m\|_W} \right).$$

Ces coordonnées permettent d'interpréter la signification de chaque axe discriminant en fonction des variables les plus corrélées.
Le **cercle unité** sert de repère visuel pour évaluer la qualité de représentation des variables dans le plan factoriel.

Partie 2

```r
# ======================== ADPLS (version finale corrigée) ======================== #
# Métrique M = I (conforme au sujet pédagogique)
# X : variables quantitatives (centrées-réduites)
# Y : indicatrices de la variable qualitative (non centrées)
# H : nombre d'axes
# w : poids (par défaut : 1/n)
# ================================================================================ #

adpls <- function(X, Y, H = 2, w = NULL, center_scale = TRUE, verbose = FALSE) {

  # ---- Préparation ------------------------------------------------------------
  X <- as.matrix(X)
  Y <- as.matrix(Y)
  n <- nrow(X); p <- ncol(X)

  if (nrow(Y) != n) stop("X et Y doivent avoir le même nombre de lignes.")

  if (is.null(w)) w <- rep(1 / n, n)
  w <- as.numeric(w)
  if (length(w) != n) stop("w doit avoir longueur = nrow(X).")

  W <- diag(w)

  # ---- Centrage-réduction de X ------------------------------------------------
  if (center_scale) {
    Xsc <- scale(X, center = TRUE, scale = TRUE)
  } else {
    Xsc <- X
  }

  # ---- Métrique M = I ---------------------------------------------------------
  M <- diag(1, p)

  # ---- Projecteur sur espace de Y ---------------------------------------------
  YtWY <- t(Y) %*% (W %*% Y)
  if (qr(YtWY)$rank < ncol(YtWY)) stop("Y' W Y est singulière.")

  PiY <- Y %*% solve(YtWY) %*% t(Y) %*% W

  # ---- X^ = Projection de X dans espace de Y ----------------------------------
  Xhat <- PiY %*% Xsc

  # ---- Matrice E (p×p) --------------------------------------------------------
  E <- t(Xhat) %*% (W %*% Xhat)
  E <- (E + t(E)) / 2  # symétrisation

  # ---- Préparation stockage ---------------------------------------------------
  U_list <- vector("list", H)
  F <- matrix(0, n, H)
  Ftilde <- matrix(0, n, H)

  S_vec <- numeric(H)
  R2_vec <- numeric(H)
```

```r
var_coords <- matrix(0, p, H)

total_inertia <- sum(diag(t(Xsc) %*% (W %*% Xsc)))

# ========================== BOUCLE PRINCIPALE ============================== #
for (h in 1:H) {

  if (verbose) message("---- Axe ", h, " ----")

  if (h == 1) {
    # ---- Axe 1 : diagonalisation de E -----------------------------------
    ev <- eigen(E, symmetric = TRUE)
    u <- ev$vectors[, 1]
    u <- u / sqrt(sum(u^2))

  } else {
    # ---- Axes suivants : contrainte d'orthogonalité ----------------------
    Fprev <- F[, 1:(h-1), drop = FALSE]   # n × (h-1)

    # D = X' W Fprev   (p × (h-1))
    D <- t(Xsc) %*% (W %*% Fprev)

    # PiDperp = I - D(D'D)^-1 D'
    inner <- t(D) %*% D

    if (qr(inner)$rank < ncol(inner)) {
      inner <- inner + diag(1e-10, ncol(inner))
    }

    PiDperp <- diag(1, p) - D %*% solve(inner) %*% t(D)

    # Projeter E dans D^
    Eproj <- PiDperp %*% E %*% PiDperp
    Eproj <- (Eproj + t(Eproj)) / 2

    ev <- eigen(Eproj, symmetric = TRUE)
    u <- ev$vectors[, 1]

    # projet final + normalisation
    u <- PiDperp %*% u
    u <- u / sqrt(sum(u^2))
  }

  # ---- Calcul score ------------------------------------------------------
  f <- as.numeric(Xsc %*% u)
  norm_f2 <- as.numeric(t(f) %*% (W %*% f))

  if (norm_f2 <= 0) stop("Norme nulle : erreur.")

  ftilde <- f / sqrt(norm_f2)

  # ---- Indicateurs -------------------------------------------------------
  S_val <- norm_f2 / total_inertia
```

```r
  fhat <- PiY %*% f
  top_val <- as.numeric(t(fhat) %*% (W %*% fhat))

  R2_val <- top_val / norm_f2

  # ---- Corrélations variables ----------------------------------------------
  var_corrs <- numeric(p)

  for (j in 1:p) {
    xj <- Xsc[, j]
    denom_xj <- sqrt(as.numeric(t(xj) %*% (W %*% xj)))
    if (denom_xj == 0) next

    var_corrs[j] <- as.numeric(t(xj) %*% (W %*% f)) / (denom_xj * sqrt(norm_f2))
  }

  # ---- Stockage ------------------------------------------------------------
  U_list[[h]] <- u
  F[, h] <- f
  Ftilde[, h] <- ftilde
  S_vec[h] <- S_val
  R2_vec[h] <- R2_val
  var_coords[, h] <- var_corrs
}

# ==================== Centres de gravité ================================= #
centers <- solve(YtWY, t(Y) %*% (W %*% Ftilde))   # q × H

rownames(centers) <- if (!is.null(colnames(Y))) colnames(Y) else paste0("class", 1:ncol(Y))
colnames(centers) <- paste0("axe", 1:ncol(Ftilde))

# ================ Packaging final ======================================== #
U_mat <- do.call(cbind, U_list)
rownames(var_coords) <- colnames(X)
colnames(var_coords) <- paste0("axe", 1:ncol(var_coords))

colnames(F) <- paste0("f", 1:ncol(F))
colnames(Ftilde) <- paste0("f_tilde", 1:ncol(Ftilde))
colnames(U_mat) <- paste0("u", 1:ncol(U_mat))

return(list(
  U = U_mat,
  F = F,
  Ftilde = Ftilde,
  S = S_vec,
  R2 = R2_vec,
  centers = centers,
  var_coords = var_coords,
  Xhat = Xhat,
  E = E,
  M = M,
  w = w
))
```

```
}
```

Exemple d'application avec les données `Datagenus` du premier TP :

```
data <- read.table("Datagenus.csv", header = TRUE)

X <- scale(as.matrix(data[, paste0("gen", 1:27)]), center = TRUE, scale = TRUE)
grp <- as.factor(data$forest)
Y <- model.matrix(~ grp - 1)

out <- adpls(X, Y, H = 3, w = NULL, center_scale = TRUE)

# Tableau indicateurs
data.frame(
  Axe = 1:3,
  S = round(out$S, 3),
  R2 = round(out$R2, 3),
  Discriminant = round(out$S * out$R2, 3)
)
```

```
##   Axe     S    R2 Discriminant
## 1   1 0.184 0.200        0.037
## 2   2 0.116 0.178        0.021
## 3   3 0.051 0.090        0.005
```

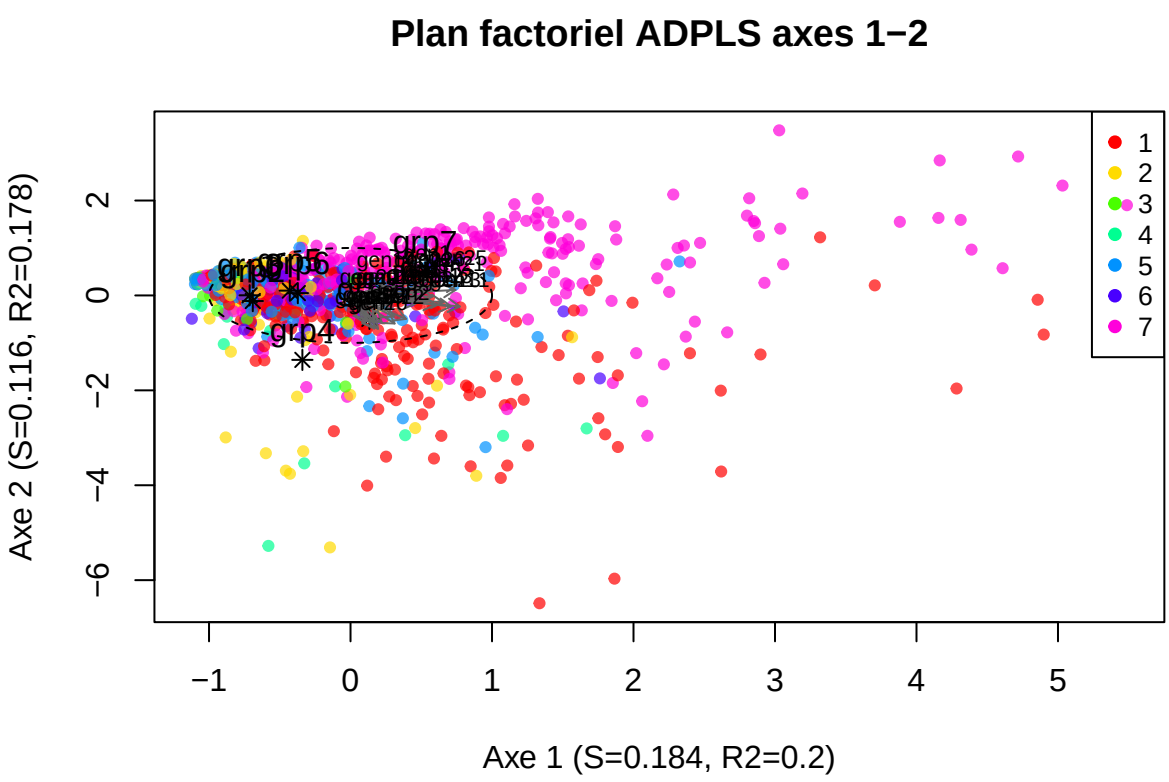
Le tableau obtenu présente les valeurs des indicateurs $S(f_h)$, $R^2(f_h, Y)$ et du produit $S(f_h) \times R^2(f_h, Y)$ pour les trois premières composantes extraites :

- La première composante (axe 1) concentre la plus grande part de l'inertie de X ($S_1 = 18{,}4\%$) et présente la liaison la plus forte avec Y ($R_1^2 = 0{,}20$). Son pouvoir discriminant $S_1 R_1^2 = 0{,}037$ est le plus élevé : elle constitue l'axe discriminant principal entre les forêts.
- La deuxième composante (axe 2) apporte une contribution plus faible ($S_2 R_2^2 = 0{,}021$), traduisant une discrimination complémentaire, souvent entre classes proches ou sous-groupes de forêts.
- La troisième composante (axe 3) ne joue qu'un rôle mineur ($S_3 R_3^2 = 0{,}005$), indiquant que la structure discriminante est essentiellement portée par les deux premiers axes.

Les deux premiers axes résument donc l'essentiel de la séparation entre les types de forêts. On peut ainsi se limiter au plan 1-2 pour l'interprétation graphique.

```
# ------------------- Graphique plan 1-2 ----------------------------------------
plot_adpls_corrected <- function(out, Y = NULL, Y_labels = NULL, h = 1, m = 2,
                                 show_variables = TRUE, circle = TRUE,
                                 main = NULL, cex_ind = 0.7, cex_cent = 1.0) {
  `%||%` <- function(a, b) if (!is.null(a)) a else b

  Ftilde <- out$Ftilde
  centers <- out$centers
  var_coords <- out$var_coords

  if (is.null(Y_labels)) {
```

```r
    if (is.null(Y)) stop("Donner Y (indicatrices) ou Y_labels (facteur).")
    Ymat <- as.matrix(Y)
    lab <- apply(Ymat, 1, function(r) which(r == 1))
    Y_labels <- factor(lab)
  }

  xind <- Ftilde[, h]; yind <- Ftilde[, m]
  classes <- as.factor(Y_labels)
  n_classes <- length(levels(classes))

  # palette simple
  cols <- rainbow(n_classes)

  plot(xind, yind,
       col = adjustcolor(cols[as.numeric(classes)], alpha.f = 0.7),
       pch = 19, cex = cex_ind,
       xlab = paste0("Axe ", h, " (S=", round(out$S[h],3), ", R2=", round(out$R2[h],3), ")"),
       ylab = paste0("Axe ", m, " (S=", round(out$S[m],3), ", R2=", round(out$R2[m],3), ")"),
       main = main %||% paste0("Projection ADPLS : axes ", h, "-", m))

  legend("topright", legend = levels(classes), col = cols, pch = 19, cex = 0.8)

  # Centres de gravité
  points(centers[, h], centers[, m], col = "black", pch = 8, cex = cex_cent)
  text(centers[, h], centers[, m], labels = rownames(centers), pos = 3)

  # Variables (flèches)
  if (show_variables) {
    arrows(rep(0, nrow(var_coords)), rep(0, nrow(var_coords)),
           var_coords[, h], var_coords[, m], length = 0.07, col = "grey40")
    text(var_coords[, h], var_coords[, m], labels = rownames(var_coords), cex = 0.7, pos = 3)
    if (circle) {
      angles <- seq(0, 2*pi, length.out = 300)
      lines(cos(angles), sin(angles), lty = 2)
    }
  }
}

# ------------------- Affichage graphique ---------------------------------------
plot_adpls_corrected(out, Y = Y, Y_labels = grp, h = 1, m = 2,
                     show_variables = TRUE, circle = TRUE,
                     main = "Plan factoriel ADPLS axes 1-2")
```



Le graphique ci-dessus représente les forêts projetées sur le plan factoriel (1-2). Chaque point correspond à une forêt, colorée selon sa classe (*forest type*), et les croix noires indiquent les centres de gravité des classes. Les flèches grises correspondent aux genres d'arbres (x_j), projetés en fonction de leur corrélation avec les axes.

L'axe 1 constitue l'axe discriminant principal, comme l'indiquent les valeurs S_1 et R_1^2. Les classes 1 (rouge) et 7 (magenta) se situent très à droite du plan, traduisant des valeurs fortement positives de f_1. À l'inverse, les classes 2 (jaune), 3 (vert), 4 (cyan) et 5 (bleu) sont regroupées autour de 0, indiquant une composition floristique beaucoup plus proche entre elles. Les forêts des classes 1 et 7 se distinguent donc nettement des autres selon les variables les plus corrélées à l'axe 1.

L'axe 2 apporte une discrimination complémentaire, mais plus faible. Certaines classes notamment la classe 1 présentent une dispersion verticale plus importante, suggérant une variabilité interne ou la présence de sous-groupes. À l'inverse, la classe 7 reste compacte sur cet axe.

Les classes 2, 3, 4, 5 et 6 se chevauchent fortement sur le plan. Du point de vue des abondances des genres d'arbres, ces types de forêts sont proches : la séparation entre elles n'est pas nette et elles ne présentent pas de signature discriminante forte sur les deux premiers axes. Cette observation est cohérente avec les faibles valeurs de $S_2R_2^2$ et $S_3R_3^2$.

Les centres de gravité sont bien alignés sur l'axe 1 : les classes les plus éloignées (1 et 7) sont nettement différenciées, tandis que les autres demeurent proches les unes des autres, confirmant la faible discrimination qui les sépare. L'axe 1 concentre donc l'essentiel du pouvoir discriminant, l'axe 2 n'apportant qu'un complément limité.

En conclusion, le plan (1-2) met en évidence une séparation nette des classes 1 et 7 du reste des données, tandis que les autres classes se superposent largement. Cette structure est cohérente avec les indicateurs, qui montrent que le premier axe porte la majorité de l'information discriminante.

Conclusion

L'Analyse Discriminante PLS exploratoire appliquée aux données *Datagenus* met en évidence une structure discriminante claire entre certains types de forêts. La première composante explique une part importante de l'inertie de X ($S_1 = 0.184$) et entretient une relation notable avec la variable de groupe Y ($R_1^2 = 0.200$). La deuxième composante apporte une information complémentaire mais plus faible ($S_2 R_2^2 = 0.021$), tandis que les composantes suivantes présentent un pouvoir discriminant limité.

L'analyse graphique confirme ces résultats :

- les classes 1 et 7 sont nettement séparées sur l'axe 1 ;
- les classes 2, 3, 4, 5 et 6 se chevauchent fortement ;
- les centres de gravité sont alignés sur l'axe 1, soulignant son rôle majeur dans la séparation.

L'ADPLS permet ainsi d'identifier les principales directions expliquant la variabilité inter-forêts et de distinguer clairement les classes les plus atypiques, tout en révélant la proximité structurelle entre les autres groupes. Dans l'ensemble, elle fournit une représentation cohérente et conforme aux objectifs de la méthode : extraire des composantes optimisant la relation entre les variables X et la variable de groupe Y.