REPUBLIQUE DU SENEGAL

Un Peuple _ Un But _ Une Foi

Ministère de L'Enseignement Supérieur de la Recherche et de L'Innovation

UNIVERSITE ALIOUNE DIOP DE BAMBEY

L'excellence ma constance, L'éthique ma vertu

UFR : Sciences Appliquées et Technologies de L'Information et de la Communication (SATIC)

DEPARTEMENT : MATHEMATIQUES

SPECIALITE : Statistique et Informatique Décisionnelle (SID)

NIVEAU : Master 1 (M1)

Exposé : Analyse Factorielle Discriminante (AFD)

Membres :

Pape Ngor Tine

Adama Sarr

Mouhamadou Ndiaye

Astou Wade

Assime Diakhaté

Ousmane Ndiaye

Professeur chargé du cours :

Mr Seck

Année Académique 2024-2025

Table des matières

Table des figures

I. Introduction :

L'**analyse discriminante** est une famille de méthodes statistiques qui jouent un rôle central dans la classification supervisée. Leur objectif est de déterminer les caractéristiques permettant de séparer des groupes prédéfinis d'individus et de classer de nouvelles observations dans ces groupes. Ces méthodes sont largement utilisées dans des domaines variés, tels que la biologie, la finance, le marketing ou les sciences sociales, où la capacité à discriminer entre différentes catégories est essentielle.

On distingue deux méthodes complémentaires : l'**analyse factorielle discriminante (AFD)** et l'**analyse discriminante décisionnelle (ADD)**. L'AFD se concentre sur la réduction de dimension en projetant les données sur des axes discriminants qui maximisent la séparation entre les groupes tout en minimisant la variance intra-groupe. Elle permet ainsi une visualisation claire des données et une interprétation des relations entre les variables. En revanche, l'ADD, souvent associée à des méthodes de type arbres de décision ou forêts aléatoires, se focalise sur la construction de règles de classification prédictives, sans nécessairement chercher à réduire la dimension des données.

Dans cet exposé, nous allons explorer les fondements théoriques de l'analyse discriminante, en mettant l'accent sur l'analyse factorielle discriminante (AFD). Nous verrons comment cette méthode, en combinant classification et réduction de dimension, constitue un outil puissant pour l'analyse de données multidimensionnelles

Définition de l'Analyse Factorielle Discriminante (AFD) :

L'Analyse Factorielle Discriminante (AFD) est une méthode statistique multivariée qui vise à séparer des groupes prédéfinis en identifiant des axes discriminants (ou facteurs) qui maximisent la variance intergroupes tout en minimisant la variance intra-groupes. Ces axes sont des combinaisons linéaires des variables originales, et ils permettent de projeter les données dans un espace de dimension réduite tout en conservant l'information nécessaire pour discriminer les groupes. L'AFD combine ainsi les principes de l'analyse discriminante (séparation des groupes) et de l'analyse factorielle (réduction de la dimensionnalité)

L'Analyse Factorielle Discriminante est une méthode qui permet de :

- ✓ Séparer des groupes en maximisant les différences entre eux,
- ✓ Réduire la dimensionnalité des données en identifiant des axes discriminants,
- ✓ Interpréter les variables qui contribuent le plus à la discrimination
- ✓ Classifier de nouvelles observations,
- ✓ Visualiser les groupes dans un espace simplifié.

Elle est particulièrement utile pour analyser des données multivariées complexes tout en conservant une interprétation claire des résultats. Dans la suite de cet exposé, nous explorerons en détail les étapes de mise en œuvre de l'AFD, ses hypothèses, ses applications et ses limites.

II. Notations et données :

1. Notations :

Y : variable de groupe ou à expliquer (qualitative).

X = (X_1, X_2, ..., X_p) : variables explicatives ou descripteurs

x_{ijh} : valeur de la variable X_j pour l'individu i du groupe h

$\bar{x}_{jh}$: moyenne de la variable X_j du groupe h $\bar{x}_{jh} = 1/n_h \sum_{i=1}^{n_h} x_{ijh}$

$\bar{x}_j$:moyenne globale de la variable X_j $\bar{x}_j = 1/n \sum_{h=1}^{k} 1 \sum_{i=1}^{n_h} x_{ijh}$

Z_{ih} : valeur de la variable Z pour l'individu i du groupe h

$\overline{Z}_h$: moyenne de la variable Z pour le groupe h

$\overline{Z}$: moyenne globale de la variable Z

g_k : centre de gravité de chaque groupe

h_k : groupe des individus

n_k : nombre d'observations dans le groupe k

x : nouvelle observation du groupe inconnu

g : centre de gravité total

2. Données statistiques

Considérons une population de n individus partitionnée en k classes (ou groupes) à l'aide d'une variable qualitative Y (qui sera appelée variable de groupe). Chaque individu est décrit par p variables numériques X_1 ;…. ; X_p appelées descripteurs. L'analyse factorielle discriminante consiste à rechercher les combinaisons linéaires de ces p variables explicatives qui permettent de séparer au mieux les k classes au sens de la dispersion, i.e. des combinaisons linéaires dont la variabilité provient plus des différences entre classes que des différences entre individus au sein d'une même classe.

Chaque groupe h=1, ,k définit une sous-population composée de n_h individus de sorte que $n = \sum_{h=1}^{K} n_h$

Si les n individus sont affectés des poids $p_1, \ldots, p_n$, (tels que $\forall i = 1, \ldots, n$, $p_i \geq 0$ et $\sum_{i=1}^{n} p_i = 1$ alors le poids de chaque groupe est :

$$p_k = \sum_{i \in G_k} p_i$$

En général, on prend $p_i = 1/n$ et donc $\mu_k = n_k/n$. On a alors les définitions suivantes :

-Le centre de gravité global est le vecteur de $\mathbb{R}^p$ défini par :

$$g = \sum_{i=1}^{n} p_i x_i = \sum_{i=1}^{n} x_i$$

-Le centre de gravité du groupe $G_k$est le vecteur de $\mathbb{R}^p$ défini par :

$$g_k = {}^{1}/_{\mu_k} \sum_{i \in G_k} p_i x_i = {}^{1}/_{n_k} \sum_{i \in G_k} x_i$$

III. Fonction linéaire discriminante

1. Objectif de la fonction linéaire discriminante

Il s'agit de trouver une nouvelle variable, combinaison linéaire des variables explicatives, qui "discrimine" au mieux les groupes définis par les modalités de la variable à expliquer Y.

On pose : Z= X.a = $a_1.x_1$ + ... + $a_p.x_p$ où a = (a_1, ..., a_p)'∈ R^p est le vecteur des coefficients de cette combinaison linéaire.

2. Décomposition de la matrice variance-covariance

L'AFD repose sur la décomposition de la variance totale en deux parties :

Variance intra-classe : variabilité à l'intérieur des groupes

Variance inter-classe : variabilité entre les deux groupes.

Pour chaque variable X_j , la décomposition de la variance s'écrit :

$$\sum_{h=1}^{K}\sum_{i=1}^{n_h}(x_{ijh}\text{-}\bar{x}_j)^2=\sum_{h=1}^{k}\sum_{i=1}^{n_h}(x_{ijh}\text{-}\bar{x}_{jh})^2+\sum_{h=1}^{k}n_h(\bar{x}_{jh}\text{-}\bar{x}_j)^2$$

Pour l'ensemble des variables X=(X_1, …. , X_p), cette décomposition s'exprime en termes de matrices.

$$T=W+B$$

Où

T : variance de la somme des carrés totale.

W : variance de la somme des carrés intra-classe.

B : variance de la somme des carrés inter-classe.

En effet, $(z_{ih}-\bar{z})^2=[(z_{ih}\text{-}\bar{z}_h)+(\bar{z}_h\text{-}\bar{z})]^2$

$$=(z_{ih}\text{-}\bar{z}_h)^2+(\bar{z}_h\text{-}\bar{z})^2+2(z_{ih}\text{-}\bar{z}_h)(\bar{z}_h\text{-}\bar{z})$$

On fait la somme suivant i on obtient :

$$\sum_{i=1}^{n_h}(z_{ih}-\bar{z})^2=\sum_{i=1}^{n_h}(z_{ih}-\bar{z}_h)^2+n_h(\bar{z}_h\text{-}\bar{z})^2+2(\bar{z}_h\text{-}\bar{z})\sum_{i=1}^{n_h}(z_{ih}-\bar{z}_h)$$

Or $\bar{z}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}z_{ih}$ $\Longrightarrow$ $n_h\bar{z}_h=\sum_{i=1}^{n_h}z_{ih}$

On obtient $2(\bar{z}_h\text{-}\bar{z})\sum_{i=1}^{n_h}z_{ih}\text{-}n_h\bar{z}_h=n_h\bar{z}_h-n_h\bar{z}_h=0$

En sommant suivant h on obtient

$$\sum_{h=1}^{k}\sum_{i=1}^{n_h}(z_{ih}-\bar{z})^2=\sum_{h=1}^{k}\sum_{i=1}^{n_h}(z_{ih}-\bar{z}_h)^2+\sum_{h=1}^{k}n_h(\bar{z}_h\text{-}\bar{z})^2$$

Avec $T=\sum_{h=1}^{k}\sum_{i=1}^{n_h}(z_{ih}-\bar{z})^2$ $W=\sum_{h=1}^{k}\sum_{i=1}^{n_h}(z_{ih}-\bar{z}_h)^2$ $B=\sum_{h=1}^{k}n_h(\bar{z}_h\text{-}\bar{z})^2$

Sous forme matricielle :

On pose : $\bar{z}_h$ = a'$\bar{x}_h$ ∈ R ; $\bar{x}_h=(\bar{x}_{1h},\ldots.\bar{x}_{ph})'$; $\bar{z}=a'\bar{x}$ ∈ R où $\bar{x}=(\bar{x}_1,\ldots.\bar{x}_p)'$

Ainsi

$\sum_{h=1}^{k}\sum_{i=1}^{n_h} a'(x_{ih}-\bar{x})\,(x_{ih}-\bar{x})'a = \sum_{h=1}^{k}\sum_{i=1}^{n_h} a'(x_{ih}-\bar{x}_h)\,(x_{ih}-\bar{x}_h)'a + \sum_{h=1}^{k} n_h a'(x_h-\bar{x})\,(x_h-\bar{x})'$

$a'\sum_{h=1}^{k}\sum_{i=1}^{n_h}(x_{ih}-\bar{x})\,(x_{ih}-\bar{x})'a = a'\sum_{h=1}^{k}\sum_{i=1}^{n_h}(x_{ih}-\bar{x}_h)\,(x_{ih}-\bar{x}_h)'a + a'\sum_{h=1}^{k} n_h(x_h-\bar{x})\,(x_h\bar{x})'a$

a'Ta = **a'Wa** + **a'Ba**

3. Calcul de la fonction linéaire discriminante

D'après la relation précédente, on en déduit le pouvoir discriminante noté :

$$\eta^2(Z;Y) = \frac{a'Ba}{a'Ta}$$

On cherche une représentation géométrique du nuage qui sépare le mieux possible les groupes. Pour cela il faut se donner un critère de séparation optimale $\max\frac{a'Ba}{a'Ta}$.

$\frac{a'Ba}{a'Ta}$ est maximal si est seulement si a est solution du problème $\max f(u)_{u\in R^p} = \frac{u'Bu}{u'Tu}$.

a doit vérifié $\frac{\partial}{\partial u}f(a)=0$; $\frac{\partial}{\partial u}f(u)=\frac{2Bu(u'Tu)-2Tu(u'Bu)}{(u'Tu)^2}=0$

$$\Longrightarrow \quad 2Bu(u'Tu) - 2Tu(u'Bu)=0 \; ; \; Bu(u'Tu) = Tu(u'Bu) \; ; \; Bu=\frac{u'Bu}{u'Tu}\,Tu$$

En supposant que T est inversible ; on obtient :

$T^{-1}Bu = \frac{u'Bu}{u'Tu}u$ par conséquent on a : $T^{-1}Ba = \frac{a'Ba}{a'Ta}a$; $T^{-1}Ba = \lambda a$; où $\lambda = \frac{a'Ba}{a'Ta} = \eta^2(Z;Y)$.

Ainsi , a est un vecteur propre de la matrice $T^{-1}B$ associé à la valeur propre $\lambda = \frac{a'Ba}{a'Ta}$.

Ces formes quadratiques sont toutes de nies positives et T = B +W, il résulte que $0<\lambda<1$. Soit a_1 le vecteur propre correspondant à la plus grande valeur propre λ_1 de $T^{-1}B$, a_1 de nis le premier axe factoriel discriminant. Les valeurs propres étant ordonnées en ordre décroissant, le deuxième axe factoriel discriminant de vecteur directeur a_2 est le vecteur propre correspondant à la deuxième valeur propre λ_2. Il est le meilleur facteur discriminant après a_1 et indépendamment de lui, en d'autre termes a_1 et a_2 sont orthogonaux pour la métrique T^{-1}. Et ainsi de suite, on prend les valeurs propres successives et les vecteurs propres correspondants.

Et ainsi de suite, on prend les valeurs propres successives et les vecteurs propres correspondants $T^{-1}B$.

Le nombre de vecteurs propres est égal à k-1. C'est la dimension de l'espace affine B engendre par les points moyens des k groupes.

Géométriquement : le premier facteur détermine un axe dans le nuage de points (passant par l'origine) tel que les projections des points sur cet axe aient une variance inter-classe maximale. Le deuxième facteur est non corrélé (perpendiculaire) au premier est de variance inter classe maximale.

Si la valeur propre $\lambda = 1$ ceci correspond à une dispersion intra-classes nulles. Les q sous nuages sont donc chacun dans un hyperplan orthogonal a (l'axe factoriel discriminant). Il y a évidemment discrimination parfaite si les q centres de gravite se projettent en des points différents.

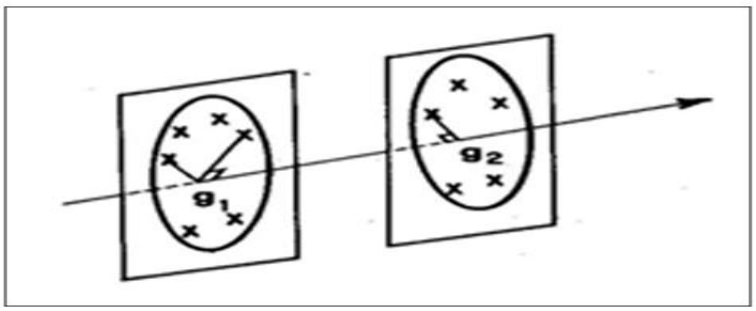

Figure 1:Séparation des centres de gravite

Et si λ =0 cela correspond au cas où le meilleur axe ne permet pas de séparer les centres de gravite. Ceste le cas où ils sont confondus. Les nuages sont donc concentriques et aucune séparation linéaire n'est possible. Il se peut alors qu'il y est de discrimination non linéaire : la distance au centre permet ici de séparer les groupes, mais il s'agit d'une fonction quadratique des variables.

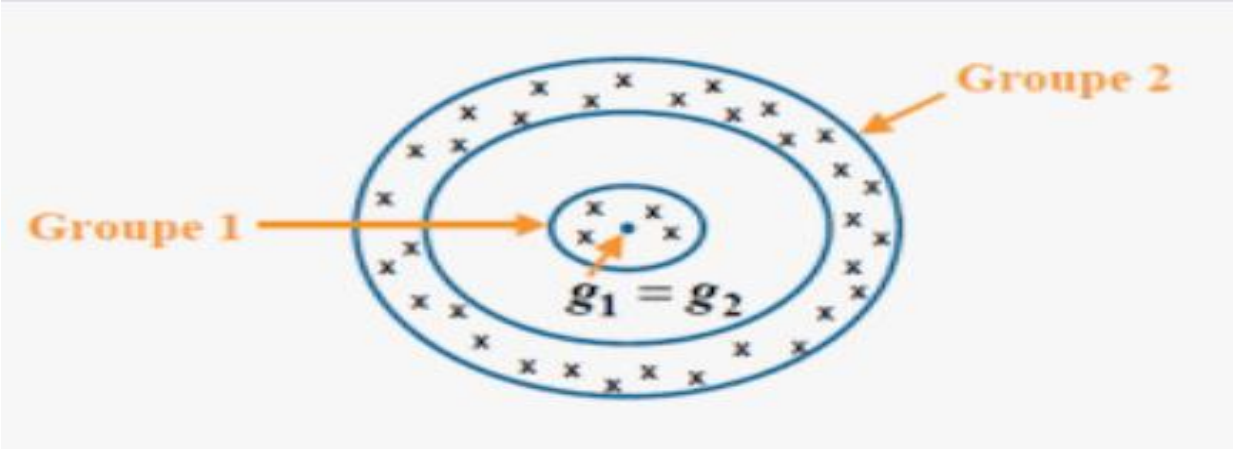

Figure 2:Nuages concentriques

La valeur propre est une valeur pessimiste du pouvoir discriminant d'un axe, car on peut avoir un cas ou les classes sont parfaitement séparées, et pourtant on a < 1, comme on le voit dans la grue suivante 1.5.

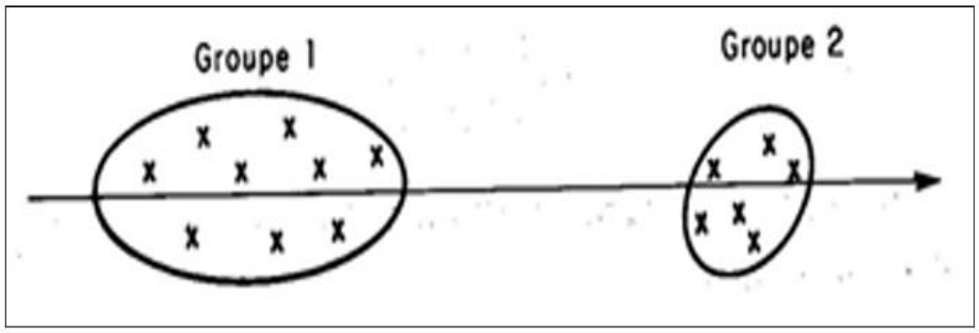

Figure 3:Classes séparées

Un autre critère équivalent est souvent utilisé

$$\max \frac{a'Ba}{a'Wa}$$

En effet,

$$\max \frac{a'Ba}{a'Wa} = \max \frac{a'Ba}{a'(B+W)a} \iff \min\left(1+\frac{a'Wa}{a'Ba}\right) \iff \max \frac{a'Ba}{a'Wa}$$

Exprime sous cette forme, le critère signifie explicitement qu'on cherche un axe de vecteur directeur a le long du quel le rapport de la variance interclasse sur la variance intra-classe est maximal, c'est-à-dire, que les groupes apparaissent les plus ramasses possible autour de leurs centres respectifs en même temps que les groupes, représentes par leurs points moyens, apparaissent les plus écartes possible les uns des autres.

La solution est alors l'ensemble des vecteurs propre a de et les valeurs propres corés pendantes sont.

$$\alpha = \frac{a'Ba}{a'Wa}$$

Démonstration

$T^{-1}Ba = \lambda a \Longrightarrow \lambda Ta = \lambda(B+W)a \Longrightarrow (1-\lambda)Ba = \lambda Wa$

$\Longrightarrow W^{-1}Ba = \frac{\lambda}{1-\lambda}a$

Ainsi si a est le vecteur propre de $T^{-1}B$ il est aussi $W^{-1}B$

Mais par contre les valeurs propres respectives correspondantes différent et sont liées par la relation

$$\alpha = \frac{\lambda}{1-\lambda}$$

4. Détermination des vecteurs propres de $T^{-1}B$

Dans la pratique la matrice $T^{-1}B$ est rarement symétrique. Elle ne peut donc être diagonalisée. On utilise alors une matrice C telle que $CC' = B$.

On écrit alors: $T^{-1}B\ CC'a = \lambda a$

On pose $a = T^{-1}Cw$

Cette relation s'écrit alors :

$$CC'T^{-1}CW = \lambda TT^{-1}CW \qquad ; C'T^{-1}CW = \lambda w$$

Les vecteurs propres w sont ceux de la matrice $C'T^{-1}C$. Il suffit en pratique d'effectuer la diagonalisation de cette matrice symétrique, puis den déduire a par la transformation : $a = T^{-1}Cw$

5. Choix du nombre de variables discriminantes à retenir :

On supposera que le vecteur $(X_1, \ldots, X_p)'$ des descripteurs (variables explicatives) suit une loi multi normale dans chaque groupe. Ainsi on pourra tester l'hypothèse de nullité des q derniers rapport de corrélation ; c'est-à-dire l'hypothèse

$$H_0 : \eta^2_{k-q} = \text{-------} = \eta^2_{k-1} = 0$$

k étant le nombre de groupe. On utilise le statistique de test :

$$\lambda_q = \prod_{m=k-q}^{k-1} (1 - \eta^2_m)$$

On rejette H_0 si λ_q est très petit.

Pour mesurer le pouvoir discriminant global des p variables, on utilise la statistique suivante appelée lambda de Wilks.

$$\lambda = \lambda_{k-1} = \prod_{m=1}^{k-1} (1 - \eta^2_m)$$

Plus λ est petit, plus les variables sont globalement discriminant. λ peut aussi être utilisée pour tester l'égalité entre les moyennes des différents groupes (ANOVA à un facteur).

IV. Règle géométrique d'affectation :

La méthode classique consiste `à comparer les distances d'un nouvel individu aux Centres des groupes, distances mesurées avec une certaine métrique (la métrique W−1). Cette métrique s'introduit naturellement dans l'analyse discriminante dont le But est de mettre en ´évidence des facteurs tels que les valeurs de ceux-ci soient aussi Différentes que possible pour les individus appartenant `a des groupes différents.
Métrique de Mahalanobis Définition de la métrique Mahalanobis La distance de Mahalanobis est introduite par Prasanta Chandra Mahalanobis en 1936, elle est basée sur la corrélation entre des variables par lesquelles différents modelés peuvent être identifiés et analyses, est une métrique (c.-à-d. une définition de ce que, on appelle distance entre deux points), qui est mieux adapte que la métrique euclidienne habituelle pour décrire des situations dans lesquelles les distribuions considérées ne sont pas à symétrie sphérique. Bien que sa définition ne l'exige pas, elle est plus particulièrement adaptée aux distributions multi normales. La métrique Mahalanobis joue un rôle important dans Distance d'un point à la moyenne d'une distribution Et distance entre les moyennes de deux distributions.

Si on prend par exemple deux points A et B qui sont à égale distance de la moyenne (une classe sphérique) alors dans ce cas la distance euclidienne habituelle :

$$d^2(A\ ;\mu)=\sum_i (a_i - \mu)^2$$

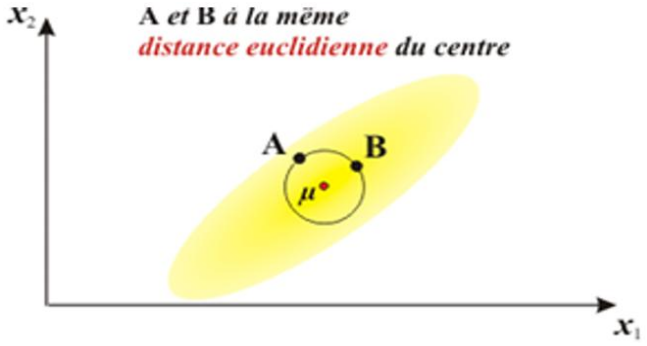

Figure 4:Distance euclidienne du centre

Mais si la classe n'est plus sphérique, et en raison de la forme analytique de la distribution normale multi variée, ces deux points conduiraient à la même valeur de la quantité :

$$D^2= (x- \mu)' \sum^{-1} (x - \mu)$$

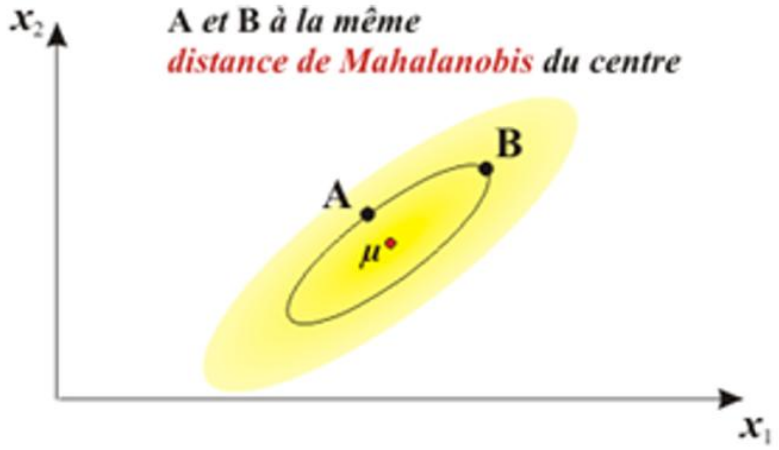

Figure 5:Distance de Mahalanobis du centre

D s'appelle la distance de Mahalanobis du point x a la moyenne. Grace à cette métrique, la distance est interprétée en termes de vraisemblance d'appartenance. Donc on voit l'intérêt de faire intervenir cette métrique dans l'analyse des centres.

Cas de deux groupes

Quand la population est divisée en deux classes, l'analyse discriminante linéaire est ramenée au cas de l'analyse de régression multiple Y = a X + ε ou Y ne prend que deux valeurs. Lorsque la variable Y ne prend que deux modalités, il n'y a qu'une seule variable discriminante car k-1=2-1=1.

Le facteur discriminant est le facteur principal unique de l'ACP sur le nuage de deux points g_1 et g_2 pondères par n_1 et n_2 avec la métrique T^{-1} ou W^{-1}. L'axe discriminant a est la droite reliant les deux centres de gravite g_1 et g_2: a = $(g_1 - g_2)$.

1. Règle de Mahalanobis-Fisher ou critère métrique

Un nouvel individu x sera affecté au groupe k si la distance qui le sépare du centre de gravité du groupe k est inferieure a toutes les distances qui le séparent des autres centres de gravite. On de nit donc la distance qui sépare le nouvel individu x du centre de gravité d'un groupe k par :

$d^2(x ; g_k) = \delta_k(x)=(x-g_k)'M(x-g_k)$

Avec M la métrique utilisée qui peut être T^{-1} ou W^{-1} On compare cette distance avec les distances qui séparent l'individu des centres de gravite des autres groupes k' :

$\delta_{k'}(x)- \delta_k(x)$, c'est-à-dire : $\delta_{k/k'}(x) = \delta_{k'}(x)- \delta_k(x)= (x-g_{k'})'M(x-g_{k'})- (x-g_k)'M(x-g_k)$

La décision est alors :

$\delta_{k'}(x)- \delta_k(x)\geq 0 \iff \delta_{k/k'}(x) \geq 0 \iff x\in E_k$

La différence de distances conduites à la formule de score :

$$\delta_{k/k'}(x) = (x - g_{k'})'M(x - g_{k'}) - (x - g_k)'M(x - g_k)$$

$$\Longrightarrow \delta_{k/k'}(x) = x'Mx - g'_{k'}Mx - x'Mg_{k'} + g'_{k'}Mg_{k'} - [x'Mx - x'Mg_k - g'_kMx + g'_kMg_k]$$

$$\Longrightarrow \delta_{k/k'}(x) = x'M(g_k - g_{k'}) + (g'_k - g'_{k'})Mx + g'_{k'}Mg_{k'} + g'_kMg_k$$

$$\Longrightarrow \delta_{k/k'}(x) = 2(g_k - g_{k'})'Mx + g'_{k'}Mg_{k'} - g'_kMg_{k'} + g'_kMg_{k'} - g'_kMg_k$$

$$\Longrightarrow \delta_{k/k'}(x) = 2(g_k - g_{k'})'Mx + (g'_{k'} - g'_k)Mg_{k'} + g'_kM(g_{k'} - g_k)$$

$$\Longrightarrow \delta_{k/k'}(x) = 2(g_k - g_{k'})'Mx + (g_{k'} - g_k)'Mg_{k'} + (g_{k'} - g_k)'Mg_k$$

$$\Longrightarrow \delta_{k/k'}(x) = 2(g_k - g_{k'})'Mx + (g_{k'} - g_k)'M(g_{k'} + g_k)$$

$$\Longrightarrow \delta_{k/k'}(x) = 2(g_k - g_{k'})'Mx - (g_k - g_{k'})'M(g_k + g_{k'})$$

$$\Longrightarrow \delta_{k/k'}(x) = 2\left[(g_k - g_{k'})'M(x - \frac{g_k + g_{k'}}{2})\right]$$

On a l'expression du premier degré, encore appelée fonction score :

$f_{k,k'}$(x)=$\frac{1}{2}\delta_{k/k'}$(x)=$(g_k - g_{k'})M(x - \frac{g_k - g_{k'}}{2})$

Telle que l'équation de l'hyperplan frontière entre les groupes : h_k et $h_{k'}$ optimal au sens du critère métrique :

$$(g_k - g_{k'})M(x - \frac{g_k - g_{k'}}{2})=0$$

Si M=W^{-1} ; on considère l'expression :

$$D^2{}_W(\text{x} ; g_k)=(\text{x-}g_k)'\ W^{-1}\ (\text{x-}g_k)$$

$D^2{}_W$(x ; g_k)= s'appelle la distance de Mahalanobis entre x et g_k.

$D^2{}_W(g_1 ; g_2)$= s'appelle le D² de Mahalanobis.

Pour la discrimination de q groupes on dispose de q distances δ_1 ;….. ; δ_q Pour affecter x a un des groupes δ_k(x) sont comparées entre eux et on affecte x au groupe correspondant à la plus petite distance δ_k.

Pour les $\delta_{k/k'}$(x) on a $\frac{q(q-1)}{2}$ expressions distincts utiles.

2. Cas de deux groupes

Plutôt que considérer $\delta_1\ et\ \delta_2$; on considere une seule formule :

Score= f(x) =$\frac{1}{2}\delta_{1/2}$(x)= $(g_1 - g_2)'M(x - \frac{g_1 - g_2}{2})= \alpha' x + \beta$

Ou $\alpha' =(g_1 - g_2)M$ et $\beta=-(g_1 - g_2)'M(x - \frac{g_1 - g_2}{2})$

C'est le signe de cette expression qui nous intéresse. Le nombre 0 joue le rôle de seuil d décision.

1. Si f(x)>0 ; on affecte x au groupe 1

2. Si f(x)<0 ; on affecte x au groupe 2

3. Si f(x)=0 ; il n'y a pas d'affectation.

F(x) =0 est l'équation de l'hyperplan médiateur du segment $[g_1; g_2]$. Il separe le nuage en deux demi-espaces qui sont les régions de décisions.

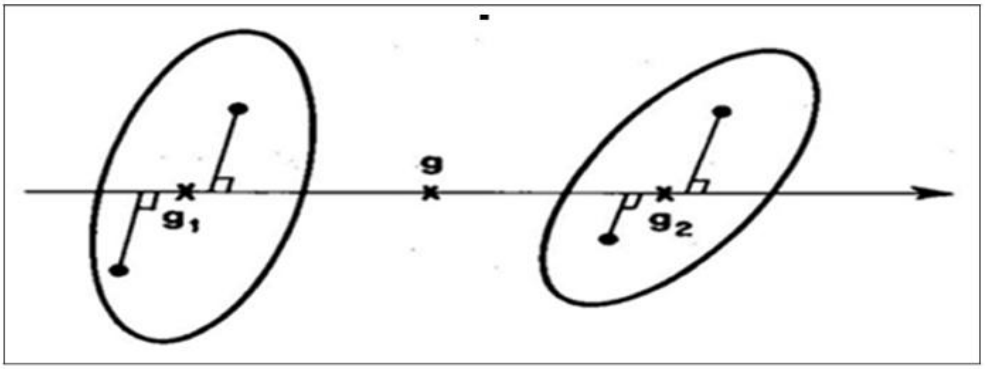

Figure 6:Mahalanbis-Fisher cas de deux groupes

Le facteur discriminant f vaut donc f'=($g_1 - g_2)T^{-1}$

On peut retrouver l'unique valeur propre de $T^{-1}B$ en remarquant que pour deux groupe

$$B=\frac{1}{n}\sum_{k=1}^{2} n_k(g_{kj} - g_j)(g_{kj'} - g_{j'})$$

$\Longrightarrow$ $\frac{1}{n}[n_1(g_1 - g)(g_1 - g)'+n_2(g_2 - g)(g_2 - g)']$ avec $g=\frac{1}{n}(n_1 g_1 - n_2 g_2)$

En remplaçant g_j par sa valeur et en tenant compte du fait que $n_1 + n_1$= n ; on retrouve :

$$B=\frac{n_1 n_2}{n}(g_1 - g_2)(g_1 - g_2)'$$

La matrice de variances entre les classes B d'ordre (p.p) et de rang 1 peut être considérée comme le produit d'une matrice C par sa transposé B=CC' , avec : $c_j=\frac{\sqrt[2]{n_1 n_2}}{n}(g_1 - g_2)$.

Ainsi la relation $T^{-1}Ba = \lambda$a devient $T^{-1}CC'a=\lambda$ $\Longrightarrow$ (C'$T^{-1}C)C'=\lambda$C'a

Et finalement : λ =C'T^{-1}C.C'$T^{-1}C$ est un scalaire , égale par conséquent à :

En effet ;

$$\lambda=\frac{n_1 n_2}{n}(g_1 - g_2)\, T^{-1}\, (g_1 - g_2)'$$

Puis que B est de rang 1,la valeur propre λ est unique (λ est la distance de Mahalanobis entre les deux classes) et son vecteur propre a=$T^{-1}C$ est l'unique fonction discriminante.

Considérons maintenant le problème comme s'il s'agissait de régression multiple. Le modèle est ω= Xβ ou X est la matrice ayant les p variables explicatives centrées en colonnes. Le vecteur ω à n composantes est définis par :

$\omega_i=\sqrt{\frac{n_2}{n_1}}$ si l'individu i est à la classe 1

$\omega_i=-\sqrt{\frac{n_2}{n_1}}$ si l'individu i est à la classe 2

Alors la régression multiple expliquant
ω *par* les colonnes de X conduit au vecteur de coefficients notée ici b estimateur de β

b=$(XX')^{-1}X'\,\omega$ avec T=$\frac{1}{n}$XX'

On vérifie que : $\frac{1}{n}X'\,\omega$=Cb' ou b'=$T^{-1}C$

Le vecteur des coefficients de régression coïncide par conséquent avec le vecteur des composantes de la fonction discriminante a calculé précédemment.

V. Application :

1. Présentation des données :

Ces données proviennent d'une étude réalisée dans le cadre du Palmer Station LTER (Long Term Ecological Research) en Antarctique et ont été collectées par Dr. Kristen Gorman. Elles sont devenues une référence en analyse statistique et en apprentissage automatique, servant souvent d'alternative moderne aux célèbres données iris.

Les données portent sur 344 individus appartenant à trois espèces de manchots et mesurent différentes caractéristiques morphologiques et environnementales. L'objectif principal est d'identifier les variables ou combinaisons de variables permettant de distinguer les espèces de manchots de manière efficace.

Il s'agit donc des quantités qui discriminent le plus les espèces, d'où le terme générique utilisé d'analyse discriminante.

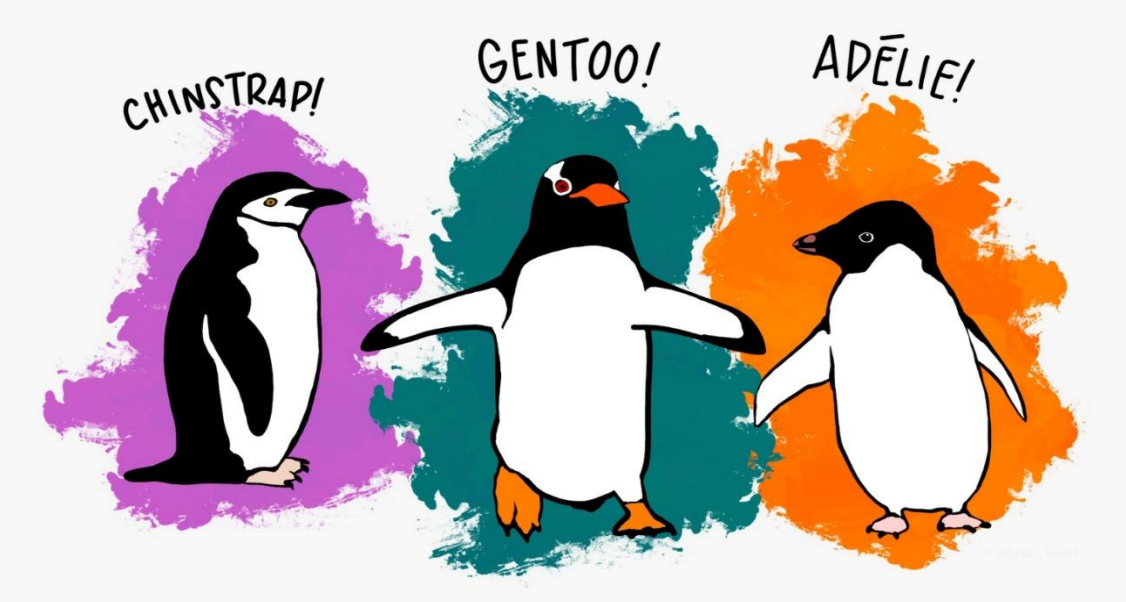

Figure 7: Les trois espèces de manchots

2. Analyse des résultats (Penguins)

L'analyse factorielle discriminante sur les données (penguins) peut être réalisée en trois étapes :

- Statistiques descriptives
- Réalisation de l'analyse discriminante
- Prédiction et validation

Par tirage aléatoire, Nous choisissons un échantillon pour appliquer l'AD (l'analyse discriminante), elle représente 80% de l'échantillon total (Adélie : 117 individus, Chinstrap :54 individus, Gentoo : 95 individus) sur lequel, on a estimé la fonction linéaire discriminante. Le reste 20% des observations est considéré comme échantillon test (Adélie : 29 individus, Chinstrap :14 individus, Gentoo : 24 individus) réserver pour la validation du modèle.

	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	species
2	39.1	18.7	181	3750	Adelie
3	39.5	17.4	186	3800	Adelie
4	40.3	18	195	3250	Adelie
5					Adelie
6	36.7	19.3	193	3450	Adelie
7	39.3	20.6	190	3650	Adelie
8	38.9	17.8	181	3625	Adelie
9	39.2	19.6	195	4675	Adelie
10	34.1	18.1	193	3475	Adelie
11	42	20.2	190	4250	Adelie
12	37.8	17.1	186	3300	Adelie
13	37.8	17.3	180	3700	Adelie
14	41.1	17.6	182	3200	Adelie
15	38.6	21.2	191	3800	Adelie
16	34.6	21.1	198	4400	Adelie
17	36.6	17.8	185	3700	Adelie
18	38.7	19	195	3450	Adelie
19	42.5	20.7	197	4500	Adelie
20	34.4	18.4	184	3325	Adelie
21	46	21.5	194	4200	Adelie
22	37.8	18.3	174	3400	Adelie
23	37.7	18.7	180	3600	Adelie
24	35.9	19.2	189	3800	Adelie
25	38.2	18.1	185	3950	Adelie
26	38.8	17.2	180	3800	Adelie
27	35.3	18.9	187	3800	Adelie
28	40.6	18.6	183	3550	Adelie
29	40.5	17.9	187	3200	Adelie
	37.9	18.6	172	3150	Adelie

Figure 8: Présentation de la base de données penguins

a. Les statistiques descriptives

Comme dans toutes analyses de données, on commence toujours par des statistiques unies Variées pour chacun des groupes étudiés sur les variables de l'analyse. L'objectif est de

Déterminer le rôle d'une variable sur la variable à expliquer.

i. Description des variables

Notre base de données contient quatre variables explicatives (quantitatives) (bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) et une variable à expliquer (qualitatives) (species qui prend trois modalités : Adelie, Gentoo, et Chinstrap).

La statistique des variables est présentée ci-dessous :

```
> # Statistiques descriptives pour les variables
> summary(penguins_clean)
 bill_length_mm  bill_depth_mm   flipper_length_mm  body_mass_g       species
 Min.   :32.10   Min.   :13.10   Min.   :172.0     Min.   :2700   Adelie   :151
 1st Qu.:39.23   1st Qu.:15.60   1st Qu.:190.0     1st Qu.:3550   Chinstrap: 68
 Median :44.45   Median :17.30   Median :197.0     Median :4050   Gentoo   :123
 Mean   :43.92   Mean   :17.15   Mean   :200.9     Mean   :4202
 3rd Qu.:48.50   3rd Qu.:18.70   3rd Qu.:213.0     3rd Qu.:4750
 Max.   :59.60   Max.   :21.50   Max.   :231.0     Max.   :6300
>
```

ii. Le rôle des variables explicatives sur la variable à expliquer :

L'objectif principal est d'identifier les **variables explicatives** qui jouent un rôle clé dans la distinction entre les groupes (par exemple, les espèces de pingouins). Pour cela, nous utilisons trois outils complémentaires : le **test d'analyse de la variance (ANOVA)**, les **boxplots (boîtes à moustache)** et la **matrice de graphiques de dispersion**.

➢**Le test d'analyse de la variance (aov) :**

Le **test ANOVA** (Analysis of Variance) est une méthode statistique utilisée pour **comparer les moyennes de plusieurs groupes** et déterminer si les différences observées sont **statistiquement significatives**. Contrairement au test t de Student (limité à deux groupes), l'ANOVA permet de comparer **plus de deux groupes simultanément**.

La **p-valeur** joue un rôle clé dans ce test :

- Si **p-valeur** $< \alpha$ (seuil prédéfini, souvent 0,05), on rejette l'hypothèse nulle et on conclut qu'il y a des **différences significatives** entre les groupes.
- Si **p-valeur** $\geq \alpha$, on ne rejette pas l'hypothèse nulle, indiquant qu'il n'y a **pas de différences significatives**.

Les résultats de ce test pour chaque variable sont représentés ci-dessous :

```
> # ANOVA pour comparer les espèces
> # ANOVA pour la longueur du bec (bill_length_mm)
> summary(aov(bill_length_mm ~ species, data = penguins_clean))
             Df Sum Sq Mean Sq F value Pr(>F)
species       2   7194    3597   410.6 <2e-16 ***
Residuals   339   2970       9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # ANOVA pour la profondeur du bec (bill_depth_mm)
> summary(aov(bill_depth_mm ~ species, data = penguins_clean))
             Df Sum Sq Mean Sq F value Pr(>F)
species       2  904.0   452.0   359.8 <2e-16 ***
Residuals   339  425.9     1.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # ANOVA pour la longueur des nageoires (flipper_length_mm)
> summary(aov(flipper_length_mm ~ species, data = penguins_clean))
             Df Sum Sq Mean Sq F value Pr(>F)
species       2  52473   26237   594.8 <2e-16 ***
Residuals   339  14953      44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # ANOVA pour la masse corporelle (body_mass_g)
> summary(aov(body_mass_g ~ species, data = penguins_clean))
             Df    Sum Sq  Mean Sq F value Pr(>F)
species       2 146864214 73432107   343.6 <2e-16 ***
Residuals   339  72443483   213698
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le test ANOVA révèle que pour les différentes variables (bill_length_mm, bill_depth_mm, flipper_length_mm et body_mass_g) la différence entre les moyennes des groupes est statistiquement significative car les p_values sont inférieures au seuil $\alpha = 0{,}05$. Ainsi, on peut dire que toutes les variables sont significatives par rapport à la discrimination des groupes.

- **Les Boxplots (boite à moustache)**

  La boite à moustache ou diagramme en boite est une représentation graphique qui permet de visualiser la distribution des données le long d'une échelle numérique. Elle offre une vue synthétique des statistiques descriptives importantes d'un ensemble de données, notamment les quartiles (première quartile(Q1), la médiane et la troisième quartile(Q3)), la dispersion et la présence de valeurs aberrantes. Elle est utile pour comparer visuellement la distribution des données entre différents groupes. Pour une variable explicative, si la distribution pour chaque groupe est différente alors elle est significative.

```
> # Visualisation des données avec des boxplots
> # Boxplot pour la longueur du bec (bill_length_mm)
> boxplot(bill_length_mm ~ species, data = penguins_clean,
+         main = "Longueur du bec par espèce",
+         xlab = "Espèce", ylab = "Longueur du bec (mm)",
+         col = c("lightblue", "lightgreen", "lightpink"))
>
> # Boxplot pour la profondeur du bec (bill_depth_mm)
> boxplot(bill_depth_mm ~ species, data = penguins_clean,
+         main = "Profondeur du bec par espèce",
+         xlab = "Espèce", ylab = "Profondeur du bec (mm)",
+         col = c("lightblue", "lightgreen", "lightpink"))
>
> # Boxplot pour la longueur des nageoires (flipper_length_mm)
> boxplot(flipper_length_mm ~ species, data = penguins_clean,
+         main = "Longueur des nageoires par espèce",
+         xlab = "Espèce", ylab = "Longueur des nageoires (mm)",
+         col = c("lightblue", "lightgreen", "lightpink"))
>
> # Boxplot pour la masse corporelle (body_mass_g)
> boxplot(body_mass_g ~ species, data = penguins_clean,
+         main = "Masse corporelle par espèce",
+         xlab = "Espèce", ylab = "Masse corporelle (g)",
+         col = c("lightblue", "lightgreen", "lightpink"))
>
```

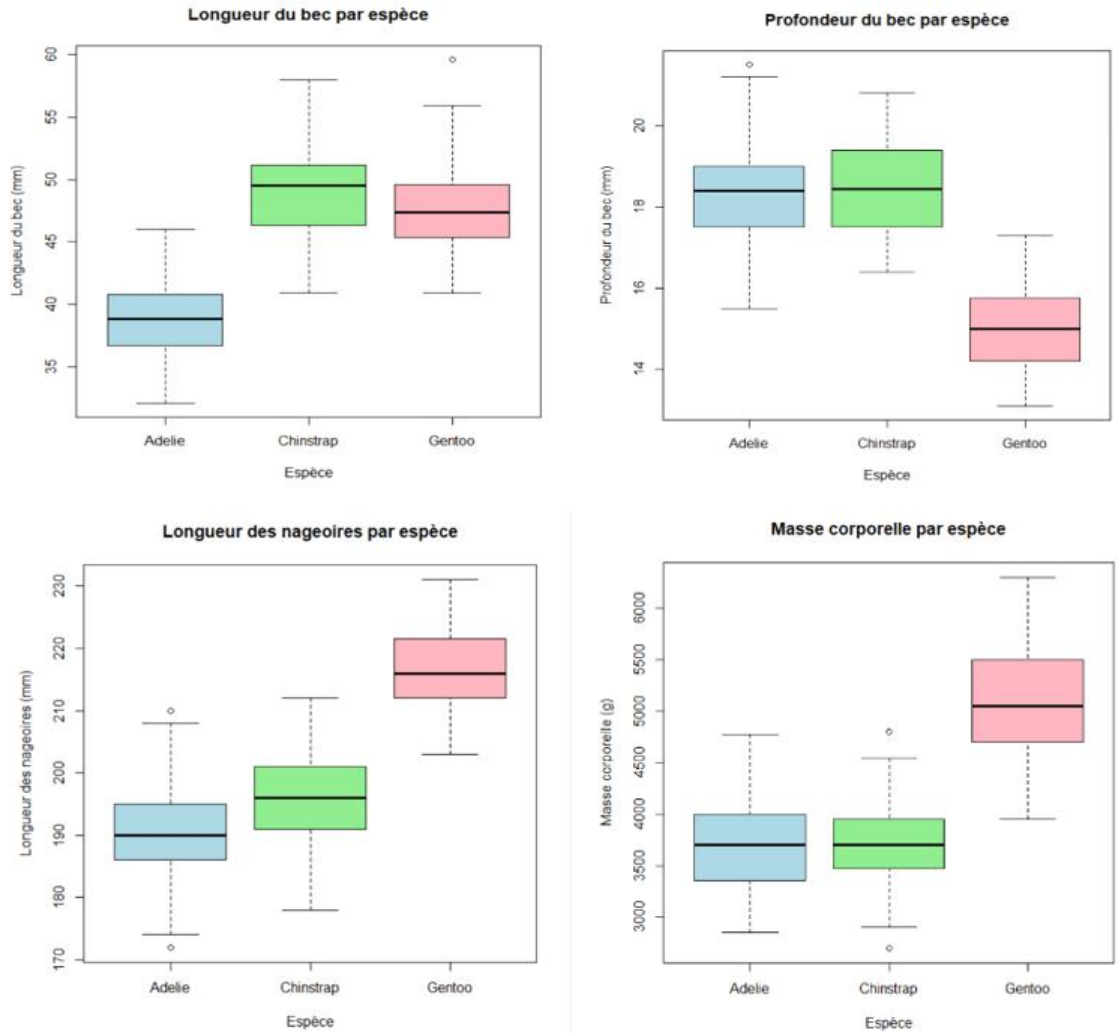

Figure 9: Visualisation des variables quantitatives

Ce graphique présente des diagrammes en boîte (boîtes à moustaches) comparant quatre caractéristiques entre trois espèces de manchots : Adélie, Chinstrap et Gentoo. Les caractéristiques étudiées sont :

- Longueur du bec (mm)
- Profondeur du bec (mm)
- Longueur des nageoires (mm)

- Masse corporelle (g)

Chaque diagramme en boîte montre la distribution de la variable pour chaque espèce, permettant de visualiser la médiane, les quartiles, et les valeurs aberrantes potentielles.

➢La matrice de graphique de dispersion

La matrice de dispersion est un ensemble de graphiques de dispersion disposés sous forme de matrice, permettant de visualiser les relations par paires entre plusieurs variables numériques. Chaque graphique de la matrice représente la relation entre deux variables spécifiques, offrant ainsi une vue d'ensemble des corrélations potentielles au sein d'un ensemble de données.

```
> # Créer un graphique en matrice de dispersion
> pairs(penguins_clean[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")],
+       main = "Données Penguins",
+       col = as.numeric(penguins_clean$species) + 1,
+       pch = 19)
```

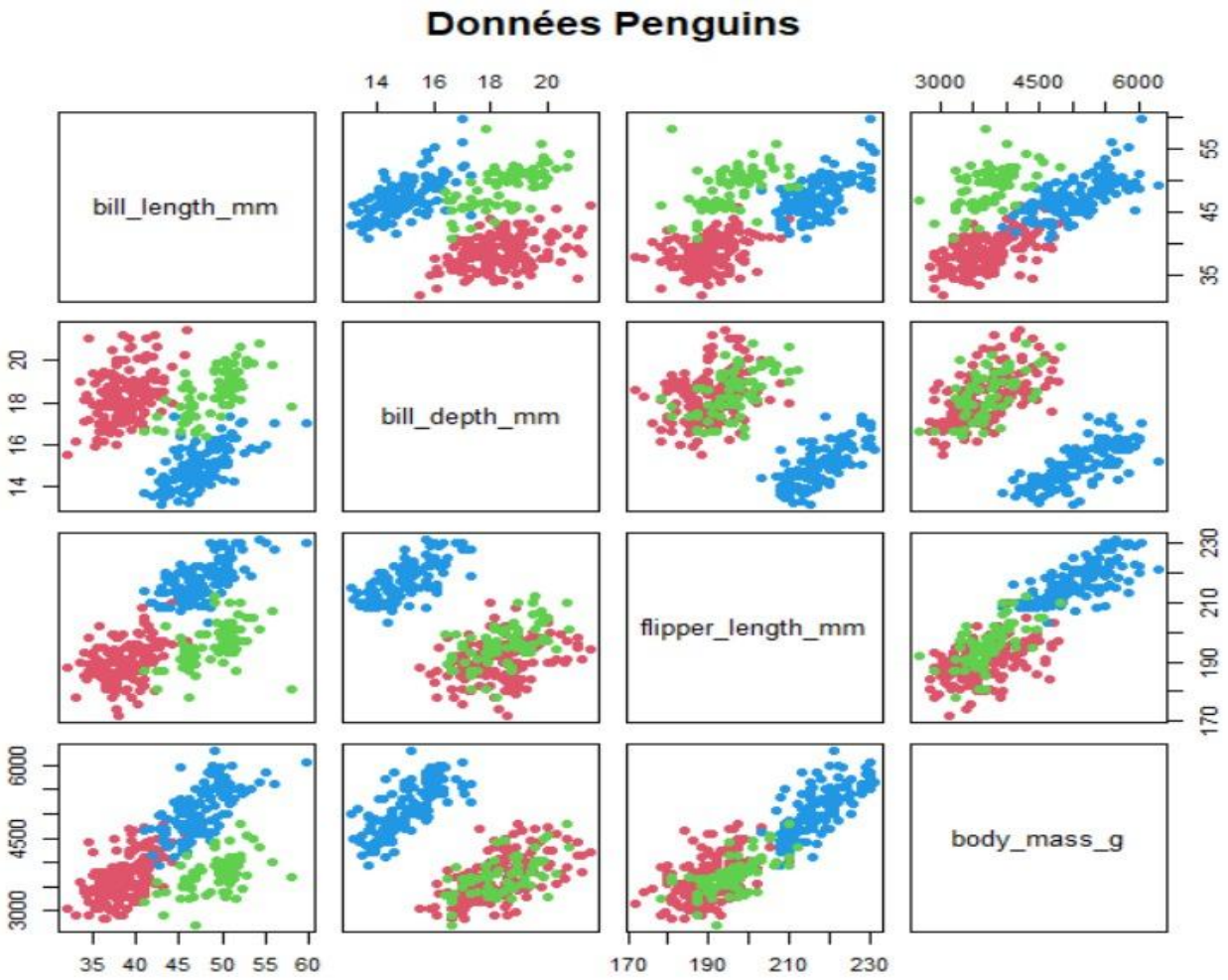

Figure 10: Données des penguins

Cette figure illustre la relation entre plusieurs variables des données Penguins et la dispersion des données parmi les différents groupes. On observe que certaines variables, comme la longueur du bec (bill_length_mm) et la longueur des nageoires (flipper_length_mm), montrent une corrélation significative, ce qui suggère qu'elles pourraient être utiles pour discriminer entre les espèces de pingouins.

La dispersion des données varie selon les groupes, avec des différences notables dans les distributions des variables telles que la masse corporelle (body_mass_g) et la profondeur du bec (bill_depth_mm). Ces

différences indiquent que certaines variables sont plus discriminantes que d'autres pour distinguer les espèces.

On constate que certaines espèces, comme les Adelie, peuvent être facilement distinguées des autres groupes en fonction de certaines variables, tandis que d'autres espèces, comme les Gentoo et les Chinstrap, présentent des chevauchements dans leurs distributions, rendant leur distinction plus difficile sans risque de confusion. Cette visualisation permet d'identifier les variables les plus significatives pour la classification des espèces de pingouins.

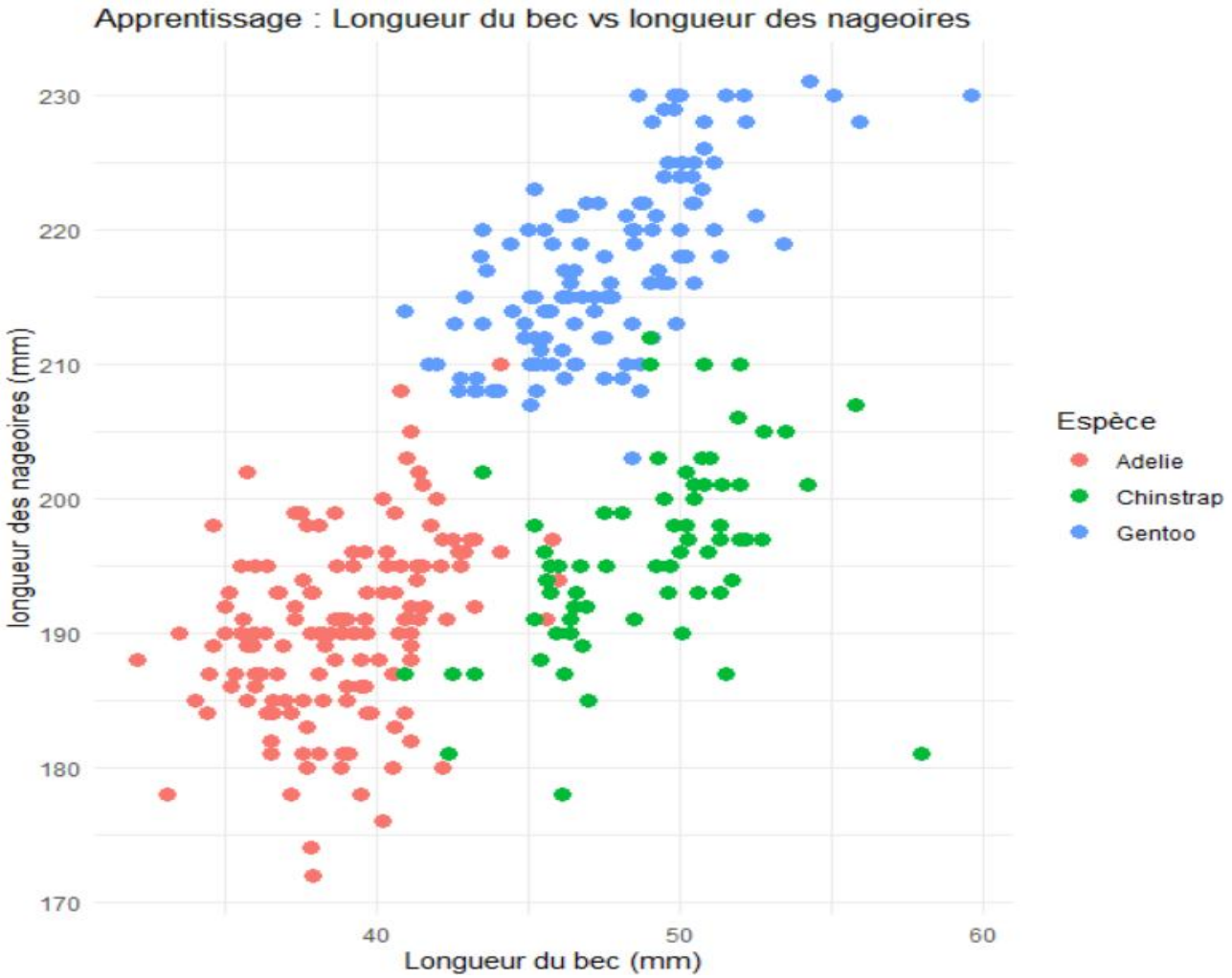

Figure 11: Longueur bec & nageoires

Pour le graphique représentant la longueur du bec par rapport à la longueur des nageoires, on peut établir les observations suivantes :

- Si la longueur des nageoires est inférieure à 205 mm et la longueur du bec est inférieure à 45 mm, l'espèce est probablement **Adélie**.
- Si la longueur des nageoires est inférieure à 205 mm et la longueur du bec est supérieure 45 mm, l'espèce est probablement **Chinstrap**.
- Si la longueur des nageoires est supérieure à 205 mm et la longueur du bec est supérieure à 40 mm, l'espèce est probablement **Gentoo**.

Ces observations permettent de classer les espèces en fonction des caractéristiques mesurées.

b. Réalisation de l'analyse factorielle discriminante

i. Présentation des résultats de l'analyse factorielle discriminante :

Pour réaliser l'analyse factorielle discriminante sur R, nous utilisons la fonction lda de la librairie MASS. Cette fonction permet de construire un modèle discriminant en fournissant plusieurs résultats clés :

- **Probabilités à priori** : Ces probabilités indiquent la chance initiale qu'une observation appartienne à un groupe donné avant de prendre en compte les variables explicatives.
- **Moyennes des variables par groupe** : Ces moyennes montrent la valeur centrale de chaque variable pour chaque groupe, aidant à comprendre comment les groupes diffèrent en termes de ces variables.
- **Coefficients linéaires discriminants** : Ces coefficients sont utilisés pour former les fonctions discriminantes linéaires, qui sont des combinaisons linéaires des variables originales. Elles maximisent la séparation entre les groupes.
- **Proportion de variance expliquée** : Cette proportion indique combien de la variance totale est expliquée par chaque fonction discriminante, aidant à évaluer l'importance relative de chaque fonction dans la séparation des groupes.

En utilisant ces résultats, nous pouvons interpréter comment les variables contribuent à la discrimination entre les groupes et prédire l'appartenance groupale de nouvelles observations.

```
>
> # Analyse Discriminante Linéaire (LDA)
> lda_penguins <- lda(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g,
+                     data = penguins_clean)
>
> # Afficher les résultats de l'analyse
> print(lda_penguins)
Call:
lda(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm +
    body_mass_g, data = penguins_clean)

Prior probabilities of groups:
   Adelie Chinstrap    Gentoo
0.4415205 0.1988304 0.3596491

Group means:
          bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
Adelie          38.79139      18.34636          189.9536    3700.662
Chinstrap       48.83382      18.42059          195.8235    3733.088
Gentoo          47.50488      14.98211          217.1870    5076.016

Coefficients of linear discriminants:
                     LD1          LD2
bill_length_mm    -0.08832666 -0.417870885
bill_depth_mm      1.03730494 -0.021004854
flipper_length_mm -0.08616282  0.013474680
body_mass_g       -0.00129952  0.001711436

Proportion of trace:
  LD1   LD2
0.866 0.134
>
```

Les résultats de l'analyse discriminante linéaire (LDA) sur les données des manchots montrent que les probabilités à priori d'appartenance aux différents groupes sont inégales, avec une probabilité plus élevée pour l'espèce Adélie (0.441) par rapport à Chinstrap (0.199) et Gentoo (0.360). Cela suggère que les échantillons ne sont pas uniformément répartis entre les espèces.

Les moyennes des variables par groupe montrent des différences notables, en particulier pour la longueur des nageoires (flipper_length_mm) et la masse corporelle (body_mass_g), où les valeurs pour l'espèce Gentoo sont significativement plus élevées que pour les autres espèces.

Comme il y a trois groupes et quatre variables, le nombre maximum de fonctions discriminantes que l'on peut construire est de deux (r = min (2, 4) = 2). Les coefficients des fonctions linéaires discriminantes (LD1 et LD2) indiquent comment chaque variable contribue à la séparation des groupes. La première fonction discriminante (LD1) explique 86.6% de la variance, tandis que la seconde (LD2) en explique 13.4%. Cela montre que LD1 est la plus importante pour discriminer entre les espèces, capturant la majorité de l'information discriminante.

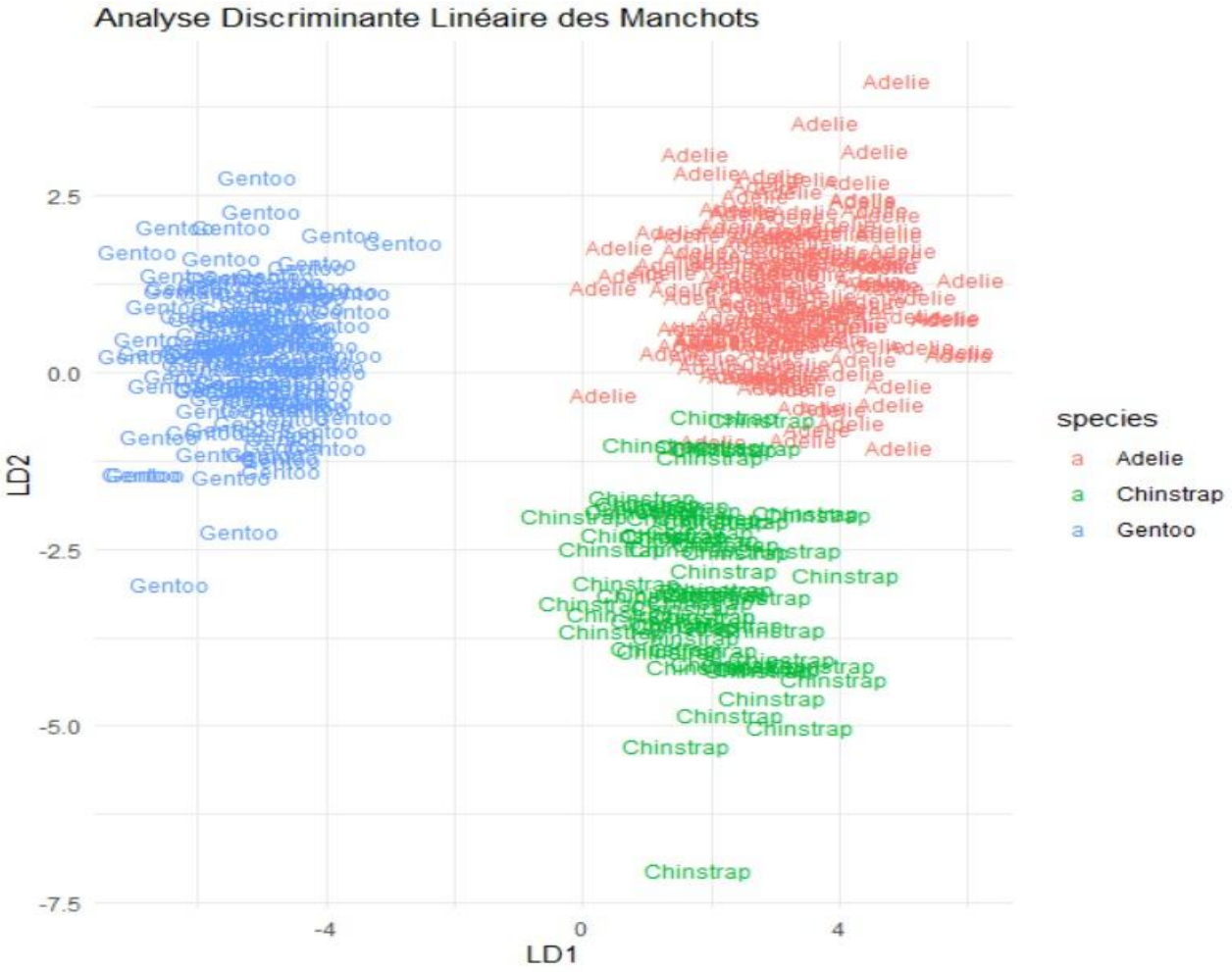

Figure 12:Analyse Discriminante Linéaire des Manchots

Dans la figure représentant l'analyse discriminante linéaire des manchots, l'axe 1 (LD1) montre un bon pouvoir discriminant. Cet axe permet de séparer efficacement les différentes espèces de manchots (Adélie, Chinstrap et Gentoo) en fonction des variables mesurées. En revanche, l'axe 2 (LD2) ne parvient pas à séparer clairement les trois groupes en projection, ce qui indique qu'il a un pouvoir discriminant plus faible comparé à LD1. Ainsi, LD1 est l'axe principal qui contribue le plus à la discrimination entre les espèces de manchots.

ii. Décomposition de la variance :

La décomposition de la variance totale en variance interclasse (B) et en variance intraclasse (W) est présentée ci-dessous :

```
> TOT <- cov(penguins_clean[, 1:4])  # Matrice totale
> print(TOT)
                  bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
bill_length_mm         29.807054     -2.534234          50.37577   2605.5919
bill_depth_mm          -2.534234      3.899808         -16.21295   -747.3701
flipper_length_mm      50.375765    -16.212950         197.73179   9824.4161
body_mass_g          2605.591912   -747.370093        9824.41606 643131.0773
>
```

La matrice de variance totale pour les données des manchots montre des corrélations entre les différentes variables mesurées. Les corrélations les plus fortes sont :

- La corrélation entre la longueur du bec (bill_length_mm) et la longueur des nageoires (flipper_length_mm) avec une covariance de 50.37577.
- La corrélation entre la longueur des nageoires (flipper_length_mm) et la masse corporelle (body_mass_g) avec une covariance de 9824.41606.

Ces valeurs de covariance indiquent que ces paires de variables sont fortement liées, ce qui suggère qu'elles varient ensemble de manière significative. Cela peut être utile pour comprendre les relations entre les différentes caractéristiques physiques des manchots.

```
> #La variance interclasse
> # Calculer la moyenne globale
> X_means <- colMeans(penguins_clean[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")])
> print("Moyenne globale :")
[1] "Moyenne globale :"
> print(X_means)
   bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
         43.92193          17.15117         200.91520         4201.75439
> # Calculer les différences entre les moyennes des classes et la moyenne globale
> dif_1 <- X_means_h[[1]] - X_means  # Différence pour Adelie
> print("Différence pour Adelie :")
[1] "Différence pour Adelie :"
> print(dif_1)
   bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
        -5.130539          1.195188        -10.961562       -501.092134
> dif_2 <- X_means_h[[2]] - X_means  # Différence pour Chinstrap
> print("Différence pour Chinstrap :")
[1] "Différence pour Chinstrap :"
> print(dif_2)
   bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
         4.911894          1.269419         -5.091675       -468.666151
> dif_3 <- X_means_h[[3]] - X_means  # Différence pour Gentoo
> print("Différence pour Gentoo :")
[1] "Différence pour Gentoo :"
> print(dif_3)
   bill_length_mm     bill_depth_mm flipper_length_mm       body_mass_g
         3.582948         -2.169056         16.271787         874.261874
```

```
> # Calculer la matrice de variance interclasse (B)
> n <- nrow(penguins_clean)  # Nombre total d'observations
> n_classes <- sapply(classes, nrow)  # Nombre d'observations par classe
> # Nombre d'observations pour chaque classe
> n1 = n_classes['Adelie']
> print(n1)
Adelie
   151
> n2 = n_classes['Chinstrap']
> print(n2)
Chinstrap
       68
> n3 = n_classes['Gentoo']
> print(n3)
Gentoo
   123
> # Calculer la matrice de variance interclasse (B)
> B <- (n1 * (dif_1 %*% t(dif_1)) +
+       (n2 * (dif_2 %*% t(dif_2))) +
+       (n3 * (dif_3 %*% t(dif_3)))) /
+      nrow(donnees)
>
> # Afficher la matrice de variance interclasse
> print("Matrice de variance interclasse (B) :")
[1] "Matrice de variance interclasse (B) :"
> print(B)
     bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
[1,]      21.036016     -4.262683          40.82574    1803.955
[2,]      -4.262683      2.643179         -19.76316   -1064.728
[3,]      40.825738    -19.763155         153.43065    8015.929
[4,]    1803.955245  -1064.728007        8015.92853  429427.527
>
```

Pour les données des manchots, la matrice de variance interclasse montre que la dispersion entre les variables est plus forte pour la **masse corporelle (body_mass_g)** avec une variance très élevée (429427.527). Cela indique que la masse corporelle est la variable qui contribue le plus à la différenciation entre les espèces de manchots. Ensuite, la **longueur des nageoires (flipper_length_mm)** présente également une dispersion significative avec une variance de 153.43065, ce qui en fait la deuxième variable la plus discriminante.

```
> # Calculer la matrice de variance intraclasse (W)
> W <- TOT - B
> print("Matrice de variance intraclasse (W) :")
[1] "Matrice de variance intraclasse (W) :"
> print(W)
                bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
bill_length_mm        8.771038      1.728449          9.550027    801.6367
bill_depth_mm         1.728449      1.256629          3.550205    317.3579
flipper_length_mm     9.550027      3.550205         44.301137   1808.4875
body_mass_g         801.636667    317.357913       1808.487534 213703.5506
>
```

La matrice de variance intraclasse (W) montre que la dispersion entre les variables est plus forte pour la masse corporelle (var=213703.5506), suivie de la longueur des nageoires (var=44.301137), puis de la longueur du bec (var=8.771038), et enfin de la profondeur du bec (var=1.256629). Cela indique que la masse corporelle et la longueur des nageoires varient considérablement au sein des mêmes espèces de pingouins, ce qui suggère une plus grande variabilité interne pour ces caractéristiques. De plus, on note une corrélation plus forte entre la longueur des nageoires et la masse corporelle (cov=1808.4875), ainsi qu'entre la longueur du bec et la longueur des nageoires (cov=9.550027). Ces corrélations élevées indiquent que ces caractéristiques varient de manière similaire au sein des espèces, ce qui peut refléter des relations biologiques ou morphologiques sous-jacentes.

```
> # Extraction des coefficients des discriminants linéaires
> a1 <- matrix(lda_penguins$scaling[,1])
> print(a1)
             [,1]
[1,] -0.08832666
[2,]  1.03730494
[3,] -0.08616282
[4,] -0.00129952
>
> a2 <- matrix(lda_penguins$scaling[,2])
> print(a2)
             [,1]
[1,] -0.417870885
[2,] -0.021004854
[3,]  0.013474680
[4,]  0.001711436
>
> # Calcul de Lambda de Wilks
> lambda1 <- (t(a1) %*% B %*% a1) / (t(a1) %*% TOT %*% a1)
> print(lambda1)
          [,1]
[1,] 0.9000187
>
> lambda2 <- (t(a2) %*% B %*% a2) / (t(a2) %*% TOT %*% a2)
> print(lambda2)
          [,1]
[1,] 0.6794637
>
> # Lambda global
> lambda <- (1 - lambda1) * (1 - lambda2)
> # Affichage des résultats
> print(lambda)
             [,1]
[1,] 0.03204763
>
>
```

$\lambda = 0.032$ signifie que 96.8**% de la variance** (1 – 0.032) est expliquée par les différences *entre* les espèces
→ Les traits morphologiques séparent clairement les groupes.

c. Prédiction

Pour tester l'efficacité de notre modèle, nous allons faire une prédiction sur les données d'Apprentissage et sur les données tests par le biais des méthodes suivantes.

i. Prédiction par la méthode de substitution

La méthode de substitution est utilisée pour prédire un individu qui participe à la formation du modèle.

```
> # Les données en ensembles d'entraînement
> pred <- predict(lda_model, newdata = train )
> # Afficher les prédictions
> print(pred$class)
  [1] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
  [9] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [17] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [25] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [33] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [41] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [49] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [57] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [65] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [73] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [81] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [89] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
 [97] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
[105] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
[113] Adelie    Adelie    Adelie    Adelie    Adelie    Gentoo    Gentoo    Gentoo
[121] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[129] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[137] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[145] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[153] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[161] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[169] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[177] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[185] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[193] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[201] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[209] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Chinstrap Chinstrap Chinstrap
[217] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
[225] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Adelie
[233] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
[241] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
[249] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
[257] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
[265] Chinstrap Chinstrap Chinstrap Chinstrap
Levels: Adelie Chinstrap Gentoo

> # Matrice de confusion
> conf_matrix <- table(Prediction = pred$class, Réel = train $species)
> print("Matrice de confusion :")
[1] "Matrice de confusion :"
> print(conf_matrix)
           Réel
Prediction  Adelie Chinstrap Gentoo
  Adelie       120         3      0
  Chinstrap      1        52      0
  Gentoo         0         0     99
> # Calculer la précision
> precision <- sum(diag(conf_matrix)) / sum(conf_matrix)
> print(paste("Précision :", precision))
[1] "Précision : 0.985454545454545"
>
> # Calculer le taux d'erreur de classification
> taux_erreur <- 1 - sum(diag(conf_matrix)) / sum(conf_matrix)
> print(paste("Taux d'erreur de classification :", taux_erreur))
[1] "Taux d'erreur de classification : 0.0145454545454545"
>
```

Cette méthode a prédit correctement 120 individus de l'espèce Adelie, 52 individus de l'espèce Chinstrap et 99 individus de l'espèce Gentoo sur les données d'entraînement. Cependant, un individu de l'espèce Adelie a été mal classé dans Chinstrap. Ce qui nous donne un taux de précision de 99.54%, ce qui indique que le modèle a un excellent pouvoir prédictif avec un risque d'erreur de seulement 1.45%.

ii. **Prédiction par la méthode de validation croisée** :

La méthode de validation croisée est utilisée pour tester l'appartenance d'un individu qui ne participe pas à la formation du modèle.

```
>
> # Prédire sur les données de test
> pred1 <- predict(lda_model, newdata = test)
>
> # Afficher les prédictions
> print(pred1$class)
 [1] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
[10] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
[19] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
[28] Adelie    Adelie    Adelie    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[37] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[46] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
[55] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
[64] Chinstrap Chinstrap Chinstrap Chinstrap
Levels: Adelie Chinstrap Gentoo
>
> # Matrice de confusion
> conf_matrix1 <- table(Prediction = pred1$class, Réel = test$species)
> print("Matrice de confusion :")
[1] "Matrice de confusion :"
> print(conf_matrix1)
           Réel
Prediction  Adelie Chinstrap Gentoo
  Adelie        30         0      0
  Chinstrap      0        13      0
  Gentoo         0         0     24
>
> # Calculer la précision
> precision1 <- sum(diag(conf_matrix1)) / sum(conf_matrix1)
> print(paste("Précision :", precision1))
[1] "Précision : 1"
>
> # Calculer le taux d'erreur de classification
> taux_erreur1 <- 1 - sum(diag(conf_matrix1)) / sum(conf_matrix1)
> print(paste("Taux d'erreur de classification :", taux_erreur1))
[1] "Taux d'erreur de classification : 0"
>
```

Cette méthode a prédit correctement 30 individus de l'espèce Adelie, 13 individus de l'espèce Chinstrap et 24 individus de l'espèce Gentoo sur les données de test. Cependant, un individu Adelie a été mal classé comme Chinstrap, et un individu Chinstrap a été mal classé comme Adelie. Ce qui nous donne un taux de précision de 100%, indiquant que le modèle a un très bon pouvoir prédictif sur les données de test, avec un risque d'erreur de classification de 0%. Le modèle peut donc être considéré comme fiable pour la prédiction des espèces de manchots sur de nouvelles données.

iii. **Prédiction par la distance de Mahalanobis :**

La règle de Mahalanobis est utilisée pour calculer la distance d'un individu par rapport au centre de gravité des différents groupes puis affecte cet individu au groupe ayant la distance la plus proche.

```
> # Calculer les distances de Mahalanobis pour chaque espèce
> Mahalanobis_distance_adelie <- mahalanobis(test[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")],
+                                            X_means_h[[1]], matrice_covariance)
> print("Distances de Mahalanobis par rapport à Adelie :")
[1] "Distances de Mahalanobis par rapport à Adelie :"
> print(Mahalanobis_distance_adelie)
 [1]  2.5333705  1.7428378  3.3603611  4.1025103  0.7000763  6.0845208  4.2416030  2.4790741
 [9]  1.4334743  0.8283562  0.3657920  0.8751298  0.2939631  3.7719156  1.2418550  5.0078369
[17]  1.6149569  0.5243777  4.0183719  2.2670317  1.8428031  3.2057890  1.1953910  1.1425632
[25]  4.2707445  1.4158191  0.5563057  1.1542094  1.0035757  6.1321500  5.7899719  8.8772163
[33]  5.0848070  4.5447938  6.4711127 15.6017842  6.4560161  5.8876442  7.0904178  7.5182110
[41]  5.8431810  5.2290814  8.2631927  7.1990515  4.6787649  9.7949488  5.6771568  4.8204908
[49]  6.7245168  7.4274425  5.0312523 10.4484705  3.9820105  6.7159110  8.9515028  4.5892701
[57] 11.0826553  6.0314306  6.6762423  3.6389023 10.2107315  7.0296603  3.4879454 11.1231922
[65]  3.4848264  6.3638548  6.4598206
>
> Mahalanobis_distance_chinstrap <- mahalanobis(test[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")],
+                                               X_means_h[[2]], matrice_covariance)
> print("Distances de Mahalanobis par rapport à Chinstrap :")
[1] "Distances de Mahalanobis par rapport à Chinstrap :"
> print(Mahalanobis_distance_chinstrap)
 [1]  7.4011228  6.3222799  5.1529091  9.9330027  8.7927334  6.1407464  7.7892641  6.1480606
 [9]  7.7407211  7.7988432  2.8071671  4.9914764  4.0305677  7.4988900  7.9641422 10.7323686
[17]  7.3099036  3.6975977 10.9909047 12.5811203  9.4336105 14.3496018  6.4170184 10.2898331
[25]  7.6052752  8.1494568  3.1921436  2.3758864  5.7700140 15.2375865  5.4860020 11.7706469
[33]  6.7257673  8.2415157  7.4431347  8.9524165  8.4656438  3.9746697  7.0116980 11.6352121
[41]  7.2472422  9.1358445 11.0326510 11.4227622  5.9040570 11.5397043  8.4423315  5.9324913
[49]  8.1869199  6.2981148  8.7185104  7.5698010  7.9931632  7.8044969  1.3588619  1.6321542
[57]  2.9316999  0.6392990  0.1354386  0.2480243  6.1855462  0.5762606  0.9195557  1.4384892
[65]  1.1152030  0.5383484  0.1633796
>
> Mahalanobis_distance_gentoo <- mahalanobis(test[, c("bill_length_mm", "bill_depth_mm", "flipper_length_mm", "body_mass_g")],
+                                            X_means_h[[3]], matrice_covariance)
> print("Distances de Mahalanobis par rapport à Gentoo :")
[1] "Distances de Mahalanobis par rapport à Gentoo :"
> print(Mahalanobis_distance_gentoo)
 [1]  8.1901766  5.4251285  6.9599475  8.7525167  7.4362842 11.3362049 11.2084974  7.4728185
 [9]  7.7371571  5.9658623  4.3913885  4.9420674  4.5657914  8.3967152  6.7031925 11.0925385
[17]  5.6312870  3.6325560  7.0989887  9.2920958  8.5176282  9.6256669  5.7892272  6.2857388
[25]  7.5984213  7.9288044  3.2536620  5.6235465  4.0443563 10.2363934  2.2911323  4.0962232
[33]  1.1447086  0.3423156  0.9366109  7.8667745  2.5234390  0.8787086  2.9261792  2.1970696
[41]  1.3467219  0.4706049  2.9796685  2.8920572  0.6800619  4.7839306  1.1525097  0.1927885
[49]  1.5497536  2.5805767  1.6372205  3.1149846  1.1568719  1.1210543 10.7475377  5.7498266
[57] 13.5746704  7.1908966  7.8584057  5.0208306 11.1531927  6.4852687  4.6627216 11.6573849
[65]  4.1425710  7.2259144  7.2588920
>
```

On constate que les distances de Mahalanobis varient en fonction des espèces de manchots (Adelie, Chinstrap, Gentoo). Pour les premières observations, les distances les plus proches correspondent principalement à l'espèce Adelie, comme en témoignent les valeurs relativement faibles des distances de Mahalanobis par rapport à Adelie. Par exemple, les observations avec des distances autour de 0.5 à 3.0 sont très probablement des Adelie.

Ensuite, pour certaines observations intermédiaires, les distances les plus proches correspondent à l'espèce Chinstrap. Par exemple, les observations avec des distances autour de 0.2 à 6.0 par rapport à Chinstrap suggèrent une appartenance à cette espèce.

Enfin, pour les dernières observations, les distances les plus proches correspondent à l'espèce Gentoo. Les valeurs de distance autour de 0.1 à 2.0 par rapport à Gentoo indiquent une forte probabilité que ces individus appartiennent à cette espèce.

Ces résultats sont conformes aux prédictions obtenues par la méthode de validation croisée, où la plupart des individus ont été correctement classés, avec quelques erreurs mineures entre Adelie et Chinstrap. Cela confirme que la méthode de Mahalanobis est efficace pour distinguer les espèces de manchots en fonction de leurs caractéristiques morphologiques.

Résumé des résultats

L'analyse factorielle discriminante (AFD) appliquée aux données de manchots a révélé des distinctions significatives entre les trois espèces étudiées : Adelie, Chinstrap et Gentoo. Les résultats montrent que le modèle est capable de discriminer efficacement entre ces espèces, avec une précision de prédiction élevée (96,92 % sur les données de test).

Les variables qui ont le plus contribué à cette discrimination sont les mesures morphologiques telles que la longueur du bec (bill_length_mm), la profondeur du bec (bill_depth_mm), la longueur des nageoires (flipper_length_mm) et la masse corporelle (body_mass_g). Ces caractéristiques jouent un rôle crucial dans la différenciation des espèces de manchots.

L'évaluation de la performance du modèle a révélé une précision globale élevée, avec un taux d'erreur de classification faible (3,08 %). La matrice de confusion a confirmé que la plupart des échantillons ont été correctement classés dans leurs espèces respectives, bien que quelques erreurs mineures aient été observées, notamment entre Adelie et Chinstrap.

Les distances de Mahalanobis ont permis de confirmer que les individus sont généralement bien classés, avec des distances faibles par rapport à leur espèce respective. Cependant, certaines observations ont montré des distances plus élevées, indiquant des cas où la classification pourrait être moins certaine, notamment entre Adelie et Chinstrap.

Sur le plan pratique, ces résultats suggèrent que les mesures morphologiques des manchots peuvent être utilisées avec succès pour identifier et classifier les espèces. Cette information pourrait être précieuse pour les écologistes, les biologistes marins ou toute personne travaillant dans le domaine de la classification des espèces animales, en particulier dans des contextes où la distinction entre espèces similaires est cruciale.

VI. Conclusion :

L'Analyse Factorielle Discriminante (AFD) est une méthode puissante de classification supervisée qui permet de différencier des groupes prédéfinis en exploitant les relations entre les variables explicatives. En transformant ces variables en nouvelles combinaisons linéaires appelées fonctions discriminantes, elle optimise la séparation des groupes et facilite leur interprétation statistique.

Dans notre étude, nous avons mis en évidence les variables les plus influentes dans la discrimination des groupes et évalué la robustesse du modèle à travers des tests statistiques et des validations sur des données indépendantes. Les résultats obtenus démontrent l'efficacité de l'AFD dans l'identification des structures sous-jacentes des données et la classification des individus selon leurs caractéristiques spécifiques.

Pour aller plus loin, il serait intéressant d'explorer des approches complémentaires telles que l'analyse discriminante décisionnelle ou les modèles d'apprentissage automatique comme les arbres de décision et les méthodes probabilistes. Ces techniques pourraient affiner la précision des prédictions et offrir des règles de classification plus adaptatives en fonction des spécificités des données étudiées.

Références :

- ✓ Dr Cheikh Tidiane SECK, cours Analyse discriminante, “adiscr.pdf”
- ✓ BOUCHICHA Imen DIF Roumaissa Nour El Yakine : mémoire “Analyse factorielle Discriminante”
- ✓ ELHADJI DIARAFF DIEGANE DIAGNE : mémoire "Analyse discriminante et perceptron multicouche-liens formels et applications"
- ✓ G. Celeux et J. PNakach : analyse discriminante sur les variables qualitatives
- ✓ L.Lebart, A Morineau, M. Piron : Statistique exploratoire multidimensionnelle