Université Claude Bernard Lyon 1

Analyse Factorielle (Projet n°6)

Auteurs
Anaïs Rodrigues
Teddy Chaix

Encadrant
Pr. Gabriela Ciuperca

26 janvier 2021

Table des matières

1 Introduction

Notre projet d'analyse factorielle portera sur la mise en œuvre des différentes méthodes d'analyse vues durant ce cours, à savoir :

- Analyse en Composante Principale (ACP)
- Analyse Factorielle Discriminante (AFD)
- Analyse Factorielle des Correspondances (AFC)
- Analyse des Corrélations Canoniques (ACC)

Pour cela nous verrons quatre problèmes, ayant chacun leur propre jeu de données et une méthode associée. Nous commencerons par faire une description de l'expérience qui a permis d'obtenir les données. Ensuite, nous réaliserons une étude descriptive (avec interprétations) pour chaque problème.

Les problèmes vus au cours de ce projet utilisent des données provenant du site Center for Machine Learning and Intelligent Systems at the University of California [1] de l'Université de Californie aux États Unis.

1. adresse web du site : http ://archive.ics.uci.edu/ml/datasets.php

2 Problème 1 : Dermatologie

2.1 Présentation des données

Pour montrer l'utilité de l'application de l'analyse factorielle des correspondances et de son interprétation, nous avons besoin d'avoir à notre disposition un jeu de données composé de plusieurs variables qualitatives, c'est-à-dire de variables composées de modalités nous indiquant par exemple si notre individu est un homme ou une femme.

Pour cela nous utiliserons les données Dermatology [2]. Cet ensemble de données est composé de 366 patients et de 34 attributs : 12 attributs cliniques et 22 attributs histopathologiques, ainsi qu'un attribut indiquant le type de maladie de la peau.

Ces données ont été assemblées suite aux problèmes de diagnostic différentiel des maladies érythémato-squameuses. En effet elles partagent toutes les caractéristiques cliniques de l'érythème et de la desquamation, avec très peu de différences.

Habituellement, une biopsie est nécessaire pour le diagnostic mais malheureusement ces maladies partagent également de nombreuses caractéristiques histopathologiques. Une autre difficulté pour le diagnostic différentiel est que la maladie peut montrer les caractéristiques d'une autre maladie dans les premiers stades et peuvent présenter les caractéristiques d'une autre aux stades suivants. Les patients ont donc d'abord été évalués cliniquement sur nos 12 caractéristiques. Puis, des échantillons de la peau ont été prélevés pour l'évaluation des 22 attributs histopathologiques qui sont déterminés par une analyse des échantillons au microscope.

La variable de l'historique familial prend la valeur 1 pour indiquer des antécédents familiaux et 0 sinon. Toutes les autres caractéristiques (cliniques et histopathologiques) se sont vu attribuées des valeurs indiquant le degré de sévérité ou de présence compris entre 0 et 3. Ici, 0 indique que la caractéristique n'était pas présente, 3 indique la plus grande quantité possible et 1, 2 indique les valeurs intermédiaires.

Pour notre étude nous n'étudierons que les attributs cliniques, excepté l'âge, ainsi que le type de maladie de la peau. Les variables histopathologiques ne seront pas traitées dans la suite.
Les différentes maladies de la peau ainsi que leurs effectifs sont présentés dans le tableau 1, de même pour les variables étudiées, c'est-à-dire les attributs cliniques dans le tableau 2.

Maladie de la peau	Traduction et fréquence	
psoriasis	psoriasis	112
seboreic dermatitis	dermatite séborrhéique	61
lichen planus	lichen plan	72
pityriasis rosea	pityriasis rosea	49
chronic dermatitis	dermatite chronique	52
pityriasis rubra pilaris	pityriasis rubra pilaris	20

TABLE 1 – Types de maladies de la peau

2. lien vers le jeu de données : http ://archive.ics.uci.edu/ml/datasets/Dermatology

Nom des attributs cliniques	Description/Traduction
erythema	rougeur de la peau ou des muqueuses
scaling	masses stratifiées sèches ou grasses de kératine
definite borders	frontières bien définies
itching	niveau de démangeaison
koebner phenomenon	apparition de nouvelles lésions cutanées
polygonal papules	papules polygonales
follicular papules	papules folliculaires
oral mucosal involvement	atteinte de la muqueuse buccale
knee and elbow involvement	atteinte du genou et du coude
scalp involvement	atteinte du cuir chevelu
family history	présence d'antécédents familiaux

TABLE 2 – Attributs Cliniques

2.2 Analyse préliminaire

Dans la suite nous noterons les variables des attributs cliniques par la lettre X, avec X_1 le niveau d'érythème (erythema), X_2 le niveau de pellicule de peau (scaling), ainsi de suite. De même nous noterons Y la variable spécifiant la maladie dont souffre le patient.

La première chose à noter dans nos données est la grande disparité des effectifs pour chaque modalité des attributs cliniques. En effet, on peut observer que la sévérité des démangeaisons (1) semble plus ou moins bien distribuée selon les maladies, de même pour le phénomène de koebner (2) en dehors des exceptions que sont la dermatite chronique et le pityriasis rubra pilaris.
Cependant cette répartition des modalités est bien différente lorsque l'on considère d'autres attributs, comme par exemple le niveau de présence de papules polygonales (3) et l'atteinte de la muqueuse buccale (4), où on observe une absence de ceux-ci pour toutes les maladies excepté le lichen plan.

Maladies / itching	psoriasis	seboreiclichen	rosea	cronic	rubra
0	55	9	2	33	8	11
1	22	16	9	10	7	8
2	21	25	28	5	20	1
3	14	11	33	1	17	

Fig. 1 – Table itching/maladie

Maladies / koebner	psoriasis	seboreiclichen	rosea	cronic	rubra
0	63	60	20	9	52	20
1	27		18	25		
2	18	1	23	12		
3	4		11	3		

Fig. 2 – Table koebner/maladie

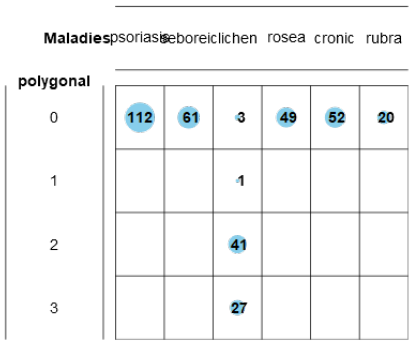

Fig. 3 – Table polygonal papules/maladie

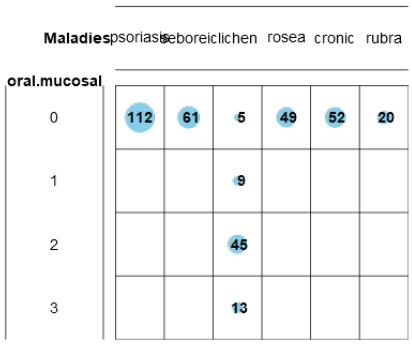

Fig. 4 – Table oral mucosal inv/maladie

Il semble donc évident que certaines caractéristiques cliniques seront plus efficaces que d'autres pour caractériser un type de maladie en particulier, et que nous sommes en présence de dépendance entre nos variables. Un test du χ^2 nous permettra de tester s'il y a indépendance entre nos X_i et Y. Dans le cas où une variable serait indépendante, celle-ci ne serait pas considérée dans la suite du problème.

Il est important de noter que le test du χ^2 n'est pas forcément le test le plus adapté à nos données du fait que nous disposons d'effectifs théoriques inférieurs à 5 (parfois jusqu'à 75% des effectifs le sont). Nous aurions pu envisager un test du χ^2 de Mantel-Haenszel étant donné que nos variables X_i sont ordinales, mais là aussi les faibles effectifs peuvent mettre en défaut notre test. Des solutions comme le test exact de Fisher sont envisageables, mais demandent un effort de computation élevé.

Nous couplerons donc le test du χ^2 à d'autres indicateurs tels que les coefficients λ asymétrie/-symétrie et U d'incertitude. Ces coefficients représentent une mesure du pourcentage d'amélioration du pronostic de Y apporté par la connaissance des X_i.

Ainsi, pour toutes nos variables on rejette l'hypothèse d'indépendance avec le type de maladie Y au niveau de risque $\alpha = 0.01$ d'après les p-value obtenues lors de nos tests et des coefficients λ et U. Néanmoins nous pouvons émettre des doutes sur les deux variables X_1 et X_{11}, c'est-à-dire les variables *erythema* et *family history*. En effet, bien que nos tests d'indépendance du χ^2 renvoient des p-value significatives, le fait est que 33% des effectifs sont inférieurs à 5 observations et nous avons des coefficients λ asymétriques et d'incertitude très faibles. Nous garderons toutefois tous les attributs cliniques pour la réalisation de notre analyse des correspondances multiples.

2.3 Analyse Factorielle des Correspondances Multiples

Nous allons désormais réaliser notre AFCM sur notre jeu de données complet. Pour cela nous utiliserons le logiciel R [3] et les packages *FactoMineR* [4] et *factoextra* [5].

Regardons dans un premier temps la proportion de la variance expliquée par nos dimensions (figure 5). On remarque que nous obtenons des résultats assez mitigés avec des pourcentages de variances expliquées par dimension très faibles. Notamment, nous devons utiliser 18 axes afin de reconstituer plus de 75% de la variance. Cependant ceci est un résultat attendu du fait que nous réalisons une AFC, nous avons donc des données sous forme de table de Burt, donc symétriques, ce qui crée beaucoup de redondance. La variance expliquée est donc sous estimée sur nos axes. Pour simplifier l'analyse, nous nous concentrerons sur l'interprétation des résultats sur les deux premiers axes qui représentent 11.21% et 8.66% de la variance expliquée bien qu'ils ne représentent qu'une faible part de la variabilité contenue dans l'ensemble du jeu de données (On pourrait aussi ignorer toutes les valeurs propres inférieures à 1/12 et donc se limiter à 14 axes, soit 65.56% de la variance).

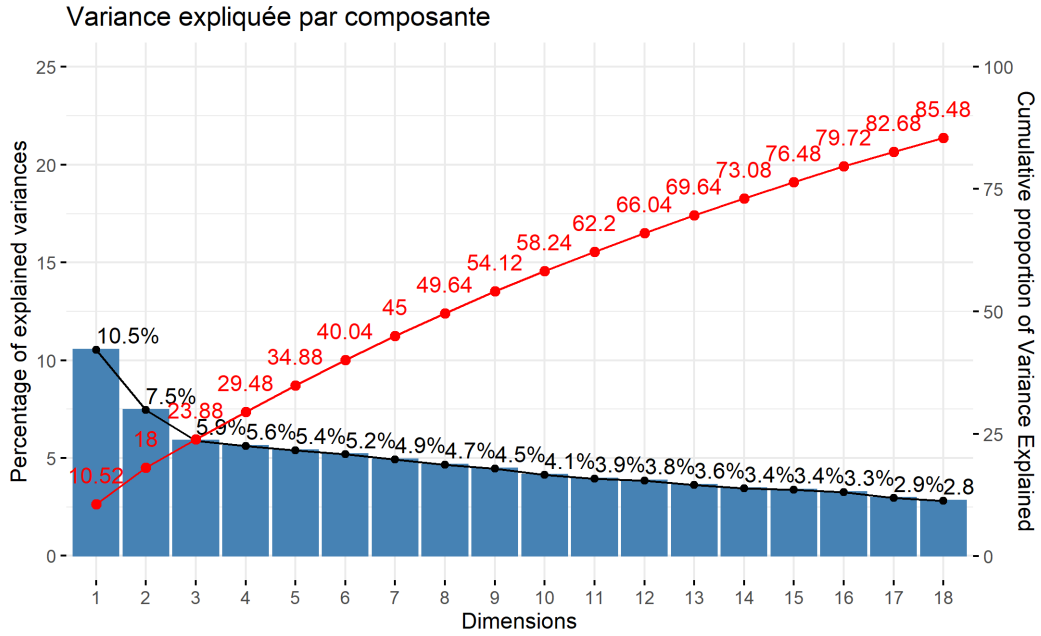

Fig. 5 – PB1 - Variance expliquée

Il semble que nous ayons du mal a expliquer la variabilité du problème, cela peut être dû en partie à la corrélation de nos attributs cliniques entre eux. Il pourra être intéressant de retirer certaines variables de l'analyse par la suite.

Nous pouvons ensuite regarder quelles variables sont les plus corrélées avec nos axes principaux. Pour cela nous pouvons projeter dans le plan factoriel nos variables en utilisant leurs corrélations au carré comme coordonnées (figure 6). Ainsi, on observe que les variables les plus corrélées avec

3. https ://www.r-project.org/
4. https ://cran.r-project.org/web/packages/FactoMineR/index.html
5. https ://cran.r-project.org/web/packages/factoextra/index.html

l'axe 1 sont les maladies de la peau, les papules polygonales et les dommages de la muqueuse buccale. Celles les plus corrélées avec l'axe 2 sont les maladies et la variable *definite borders*.

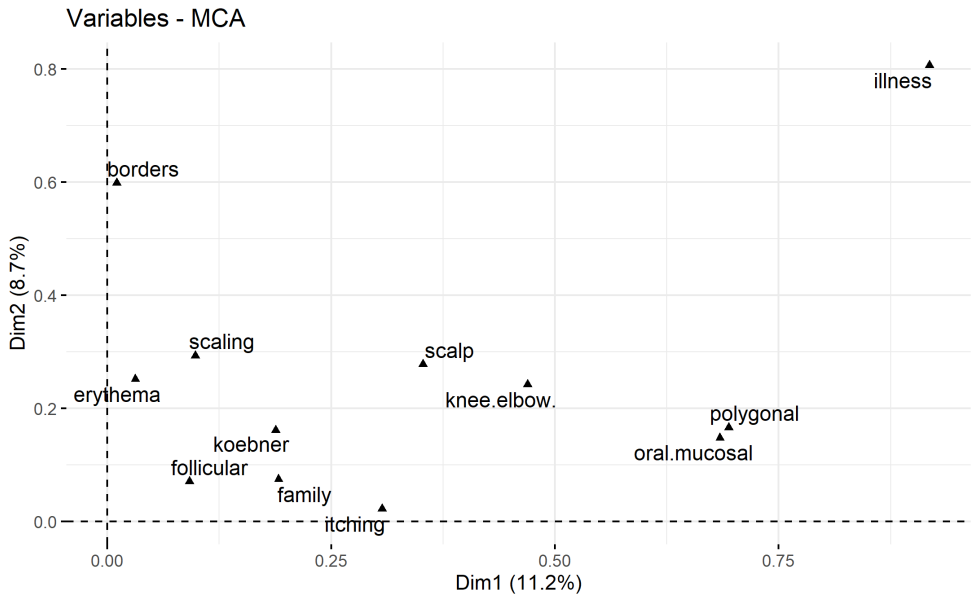

Fig. 6 – PB1 - Corrélation avec les axes principaux 1 et 2

Intéressons nous maintenant aux modalités qui contribuent le plus à la définition des axes 1 et 2. Pour cela nous pouvons regarder le diagramme en barre des contributions de chaque modalités avec nos deux axes (figures 7 et 8).

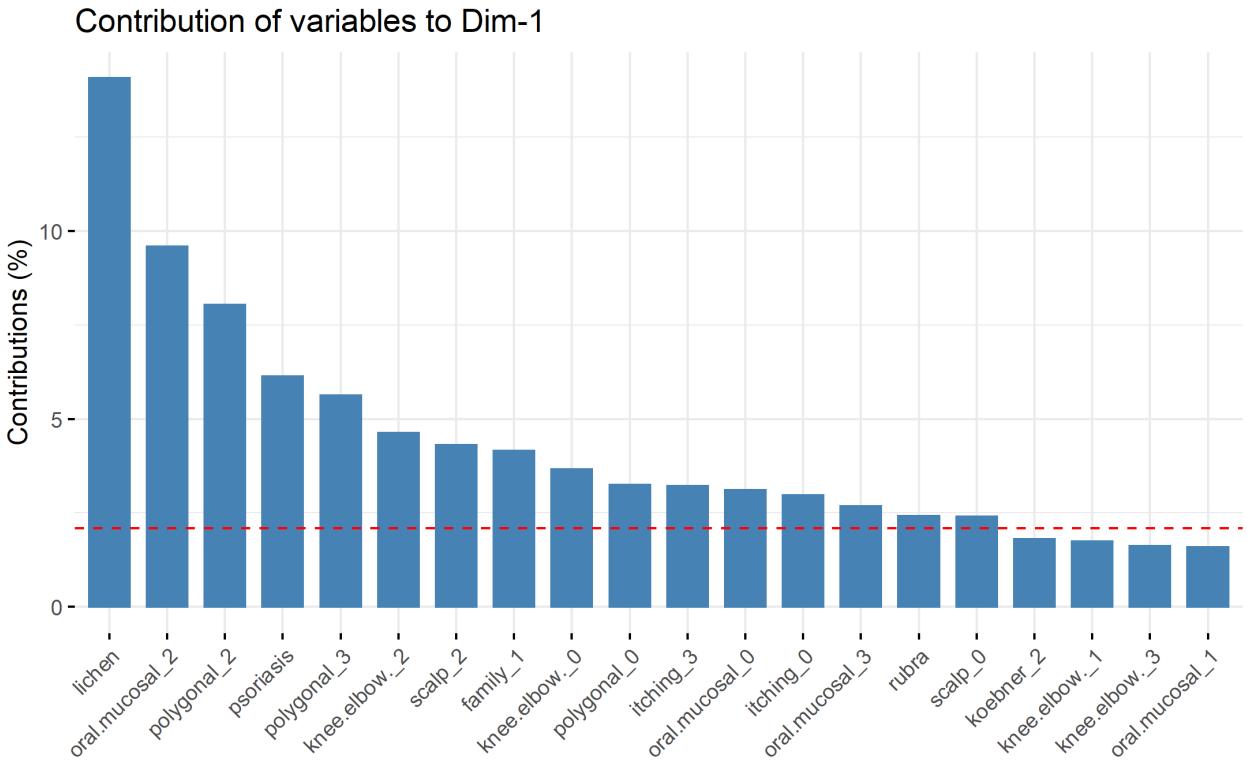

Fig. 7 – Contribution à l'axe 1

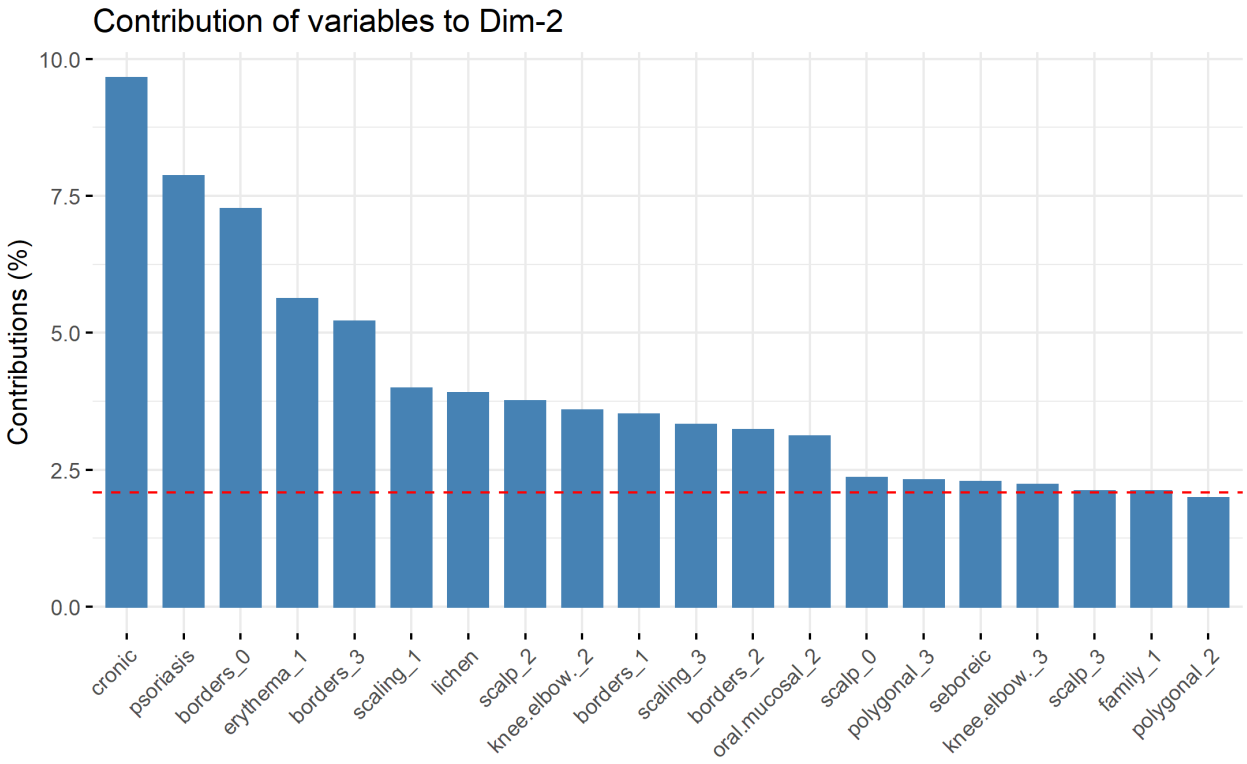

Fig. 8 – Contribution à l'axe 2

On observe que les modalités qui contribuent le plus à l'axe 1 sont :

- lichen
- psoriasis
- pityriasis rubra pilaris
- oral musocal involvement : niveaux 0,2 et 3
- polygonal papules : niveaux 0,2 et 3
- knee and elbow involvement : niveaux 0 et 2
- sclap : niveaux 0 et 2
- family history : niveau 1
- itching : niveaux 0 et 3

Pour l'axe 2 :

- chronic dermatitis
- psoriasis
- lichen
- seboreic dermatitis
- definite borders : niveaux 0,1 et 3
- erythema : niveau 1
- sclap : niveaux 0,2 et 3

- knee and elbow involvement : niveaux 2 et 3
- scaling : niveau 3
- oral musocal involvement : niveaux 2
- family history : niveau 1
- polygonal papules : niveau 3

Nous pouvons désormais nous intéresser aux corrélations des différentes modalités avec nos axes principaux. Pour cela nous réaliserons une projection des modalités dans le plan factoriel défini par les axes 1 et 2 de notre AFCM (figure 9). Les modalités seront également coloriées en fonction de leur valeur du cos^2, c'est-à-dire de leur qualité de représentation sur ces axes.

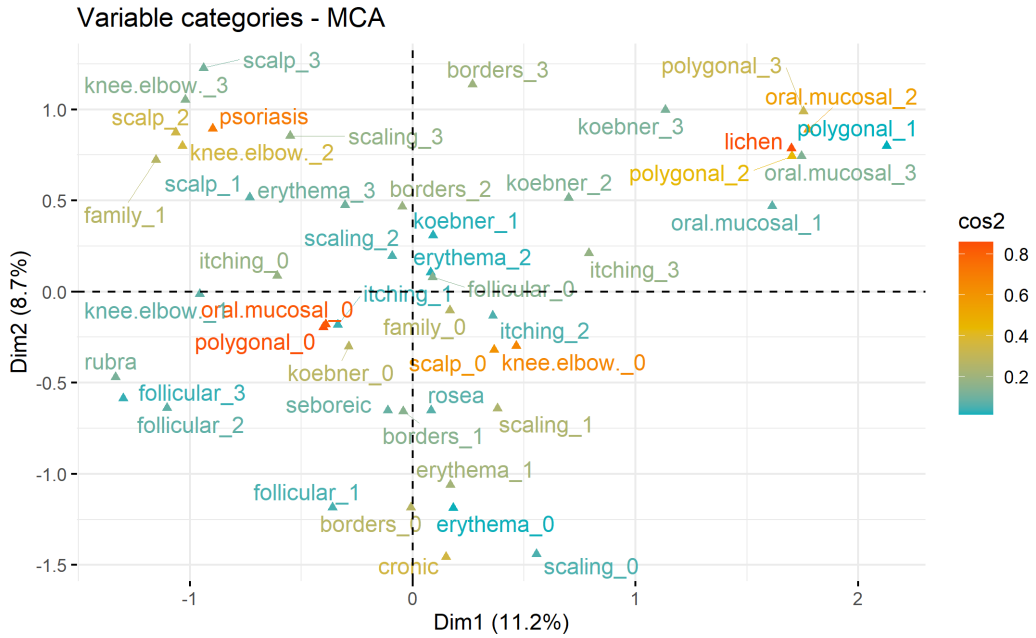

Fig. 9 – PB1 - Corrélation avec les axes principaux 1 et 2

On observe que les modalités :

- Lichen
- polygonal papules : niveaux 1,2 et 3
- oral and mucosal involvement : niveaux 1,2 et 3

sont proches et sont opposées à :

- Psoriasis
- sclap involvement : niveaux 2 et 3
- knee and elbow involvement : niveaux 2 et 3
- family history : 1

qui sont également opposées à :

- pityriasis rubra pilaris
- folluclar papilus : niveaux 2 et 3

et enfin :

- chronic dermatisis
- scaling : niveau 0
- border : niveau 0
- erythema : niveau 0

C'est-à-dire que les individus présentant des papules polygonales et des dommages à la muqueuse buccale sont sujets à la présence de lichen plan. Et à l'opposé, les individus qui présentent des problèmes de peaux au niveau du crâne, des genoux et des coudes assez sévères et des antécédents familiaux sont sujets à du psoriasis. Ceux ayant des papules folliculaires sont sujets à du pityriasis rubra pilaris. Et enfin les patients qui présentent une absence d'érythème, de croutes, et de frontières bien définies pour les lésions, sont corrélés avec la maladie de dermatite chronique.
Nous pouvons finir par représenter les individus colorés en fonction de la maladie dont ils souffrent dans le plan factoriel afin de voir si ceux-ci sont bien caractérisés. (figure 10)

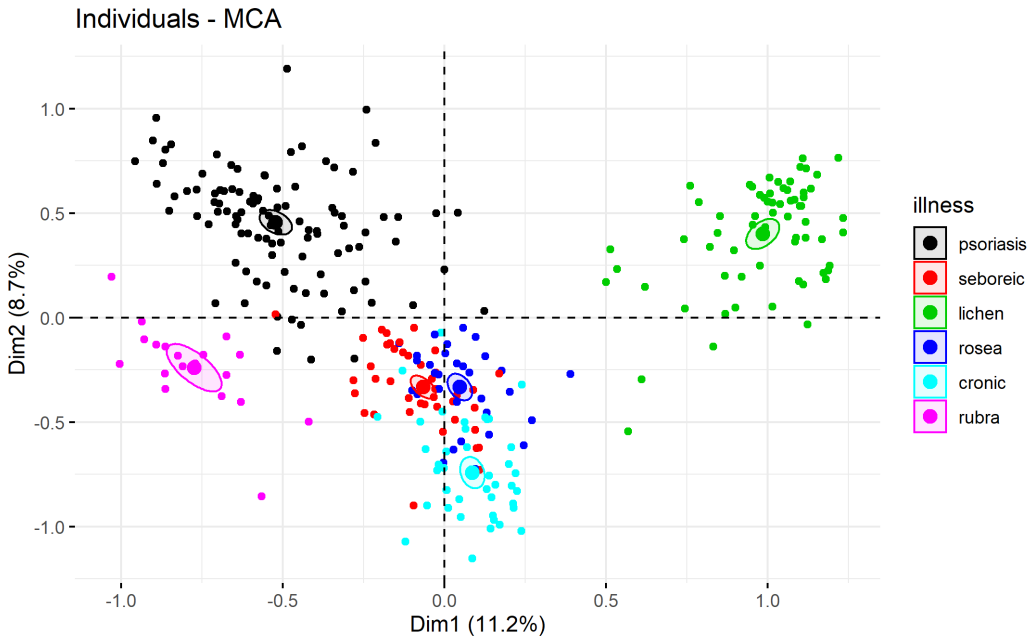

Fig. 10 – PB1 - Représentation des individus dans le plan

On remarque que les maladies citées plus haut sont bien espacées dans le plan, notamment les trois premières. Par contre il reste difficile de caractériser la dermatite séborrhéique et le pityriasis rosea avec les modalités des attributs cliniques.

3 Problème 2 : Espèces Anuran

Pour montrer l'utilité de l'application de l'analyse en composante principale et de son interprétation, nous avons besoin d'un jeu de données composé de plusieurs variables quantitatives.

Pour cela nous utiliserons les données Anuran [6]. Cet ensemble de données est composé de 7195 syllabes et de 26 attributs : 3 attributs catégoriques et 22 attributs Mel-frequency cepstral coefficients [7] (MFCCs) ainsi qu'un attribut indiquant l'identifiant de l'enregistrement audio.

Fig. 11 – PB2 - Illustration d'une dendrobates azureus - wikipédia

Ces données ont été créées à partir d'enregistrement audio collectés in situ dans des conditions de bruit réel (le bruit de fond) dans le but de détecter le type de grenouille ou de crapaud à partir de leurs cris. Les syllabes ont été réalisées en segmentant 60 enregistrements audio appartenant à 4 familles différentes, 8 genres et 10 espèces.

Chaque clip audio correspond à un spécimen (une grenouille ou un crapaud). La méthode de cluster binaire basée sur l'entropie spectrale a été utilisée afin de détecter les syllabes des enregistrements. Le logiciel Matlab a été quand à lui utilisé pour la segmentation et l'extraction des caractéristiques. À partir de chaque syllabe extraite, 22 MFCCs ont été calculés en utilisant 44 filtres triangulaires, et étant donné que chaque syllabe a une longueur d'onde différente, chaque colonne (i) a été normalisée par : $\frac{MFCCs_i}{max(|MFCCs_i|)}$ car les valeurs des MFCCs ne sont pas très robustes en présence de bruit il est donc courant de normaliser leurs valeurs, par exemple dans les systèmes de reconnaissance vocale, pour en réduire l'influence.

Pour notre étude nous étudierons principalement les attributs MFCCs afin de voir si nous pouvons réduire la dimension de notre espace de variable (on posera $X_i = MFCCs_i$) pour faciliter le traitement et l'étude des données. Puis nous regarderons si nous pouvons trouver des relations entre nos variables.

6. lien vers le jeu de données : http ://archive.ics.uci.edu/ml/datasets/Anuran+Calls+(MFCCs)
7. lien vers wikipédia pour la description des MFCCs : https ://en.wikipedia.org/wiki/Mel-frequency_cepstrum

Les différentes variables catégoriques et leurs effectifs sont présentés dans le tableau 3. Elles seront utilisées plus tard afin de voir si notre espace de dimension réduite permet de les caractériser.

Nom des attributs catégoriques	Effectifs
Famille (nom commun)	
Bufonida (true toads)	68
Dendrobatidae (poison dart frogs)	542
Hylidae (tree frogs)	2165
Leptodactylidae (tropical frogs)	4420
Genre (nom commun)	
Adenomera (tropical bullfrogs)	4150
Ameerega	542
Dendropsophus	310
Hypsiboas (gladiator frogs)	1593
Leptodactylus (ditch frogs)	270
Osteocephalus (slender-legged tree frogs)	114
Rhinella (Beaked toads)	68
Scinax (snouted tree frogs)	148
Espèces	
Adenomera Andreae	672
Adenomera Hylaedactyla	3478
Ameerega Trivittata	542
Hyla Minuta	310
Hypsiboas Cinerascens	472
Hypsiboas Cordobae	472
Hypsiboas Cinerascens	1121
Leptodactylus Fuscus	270
Osteocephalus Oophagus	114
Rhinella Granulosa	68
Scinax Ruber	148

TABLE 3 – Répartition des observations selon la famille, genre et espèces

3.1 Étude de la corrélation

Avant de nous lancer dans une ACP, intéressons-nous aux liaisons qu'il peut y avoir entre nos variables. En effet, si nos variables sont indépendantes ou très peu corrélées entre elles, la réduction de la dimension risque d'être faible voir inenvisageable. Pour cela nous pouvons jeter un œil au tableau 12 qui représente les coefficients de corrélation entre chaque $MFCCs_i$ ainsi que la significativité de la p-value associée au test d'indépendance. Les cases blanches indiquent une faible corrélation entre les variables ligne et colonne associées, bleues une forte corrélation positive et rouges une forte corrélation négative.
On remarque que nos données sont fortement corrélées entre elles, avec notamment une forte association négative entre les variables $MFCCs_i$ et $MFCCs_{i+2}$ et une association positive entre $MFCCs_i$ et $MFCCs_{i+4}$. L'ACP s'avère donc intéressante car elle est particulièrement

utile lorsque les variables sont fortement corrélées, en effet cela indique qu'il y a redondance dans les données. On s'attend donc a pouvoir réduire assez efficacement la dimension de notre espace. On remarque également que les variables $MFCCs_1$, $MFCCs_2$, $MFCCs_{18}$, $MFCCs_{21}$ sont globalement moins corrélées avec le reste des données que les autres variables.

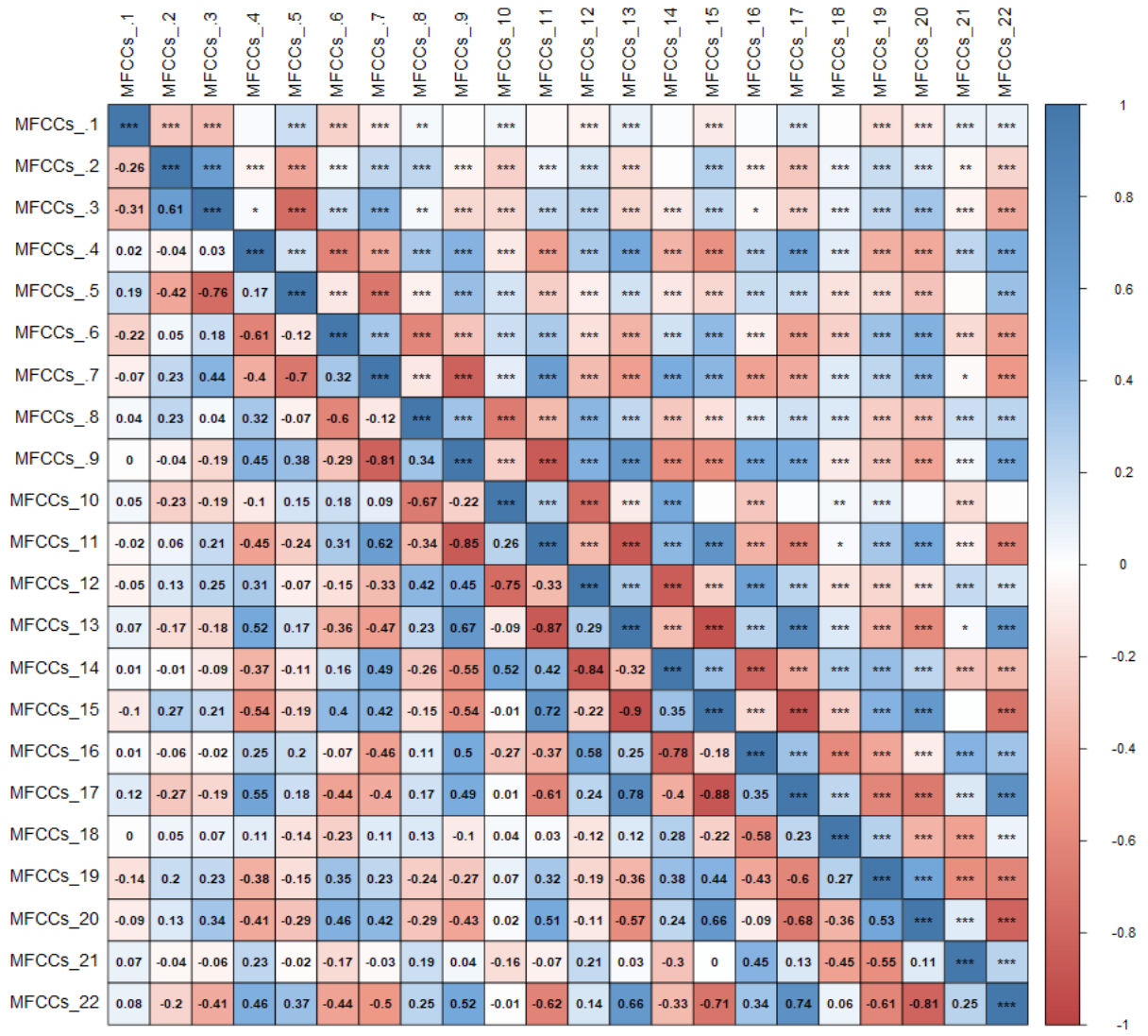

Fig. 12 – PB2 - Corrélations des MFCCs et significativités

3.2 Analyse en Composantes Principales

Nous allons maintenant réaliser notre ACP sur les variables quantitatives de notre jeu de données, les variables qualitatives (Family, Genus, Species) seront utilisées pour colorier nos individus par la suite. Comme pour le problème précédent nous réaliserons notre étude à l'aide du logiciel R et des packages *FactoMineR* et *factoextra*. De plus, bien que nos données soient

déjà normalisées pour les rendre comparables, nous les normaliserons de nouveau par l'intermédiaire de notre fonction réalisant l'ACP afin que leur écart-type soit égal à 1 pour faciliter l'interprétation.

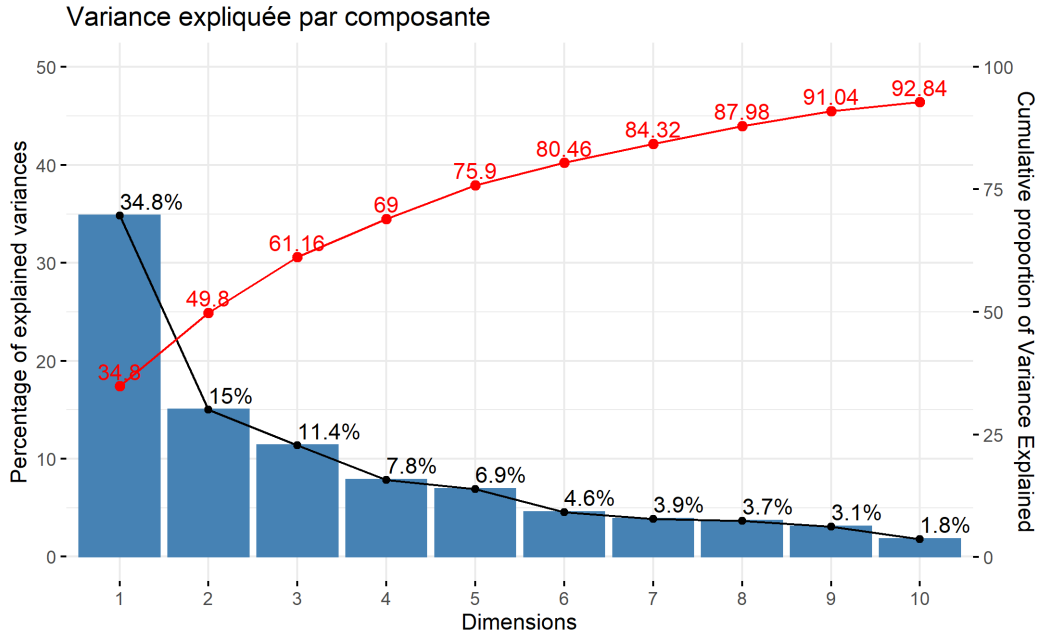

Fig. 13 – PB2 - Variance expliquée

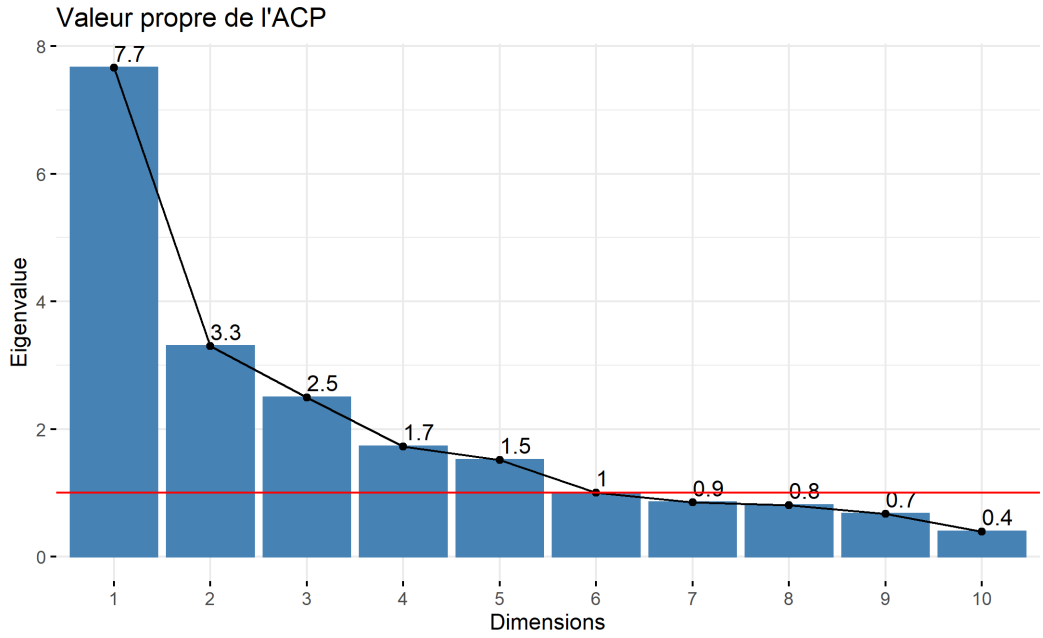

Fig. 14 – PB2 - Valeurs propres de l'ACP

On a représenté sur la figure 23 la part de la variance expliquée par nos composantes principales. Si nous souhaitons expliquer au moins 80% de la variabilité des données nous devons recourir à 6 axes principaux. Nous pouvons également regarder les valeurs propres associées à nos axes et se limiter à celles strictement supérieures à 1 (Kaiser-Guttman sur données normalisées), et ici aussi on retiendra 6 axes (figure 14). On peut donc réduire nos données à 6

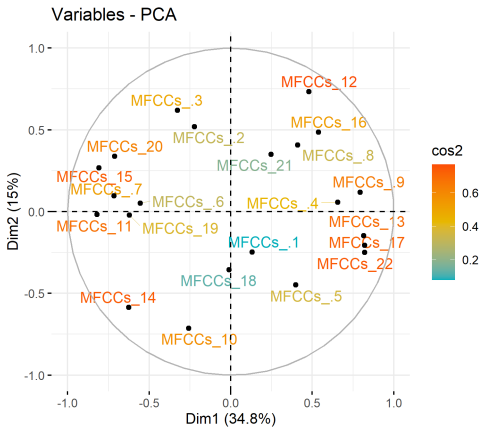

Fig. 15 – Biplot des variables - axes 1 et 2

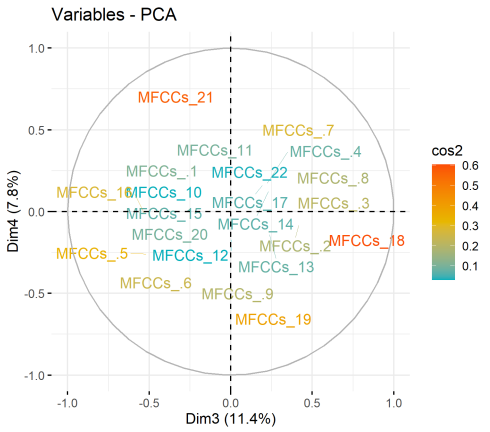

Fig. 16 – Biplot des variables - axes 3 et 4

axes principaux expliquant 80.46% de la variance du modèle. Pour des facilités d'analyse, nous nous réduirons cependant à l'étude des 4 premiers axes avec donc 69% de la variance du modèle.

Nous pouvons maintenant projeter nos variables dans le plan factoriel, plus particulièrement sur les axes 1 et 2 (figure 15) puis 3 et 4 (figure 16) afin de visualiser les variables formant des groupes. La coloration des variables dans les graphiques indique la qualité de représentation de chaque variable sur les composantes considérées. Les variables ayant un score de *cos2* proche de 1 sont bien représentées par ces axes (rouge) à l'inverse celles proche de 0 (bleu).

On peut ainsi classer les variables selon leurs corrélations avec les axes, par ordre de la plus à la moins corrélée.

Pour l'axe 1, les variables corrélées positivement sont :

- $MFCCs_{22}$
- $MFCCs_{17}$
- $MFCCs_{13}$
- $MFCCs_{9}$
- $MFCCs_{4}$
- $MFCCs_{16}$

Pour l'axe 2 les variables corrélées positivement sont :

- $MFCCs_{12}$
- $MFCCs_{3}$
- $MFCCs_{2}$

Pour l'axe 1, les variables corrélées négativement sont :

- $MFCCs_{11}$
- $MFCCs_{15}$
- $MFCCs_{7}$
- $MFCCs_{20}$
- $MFCCs_{14}$
- $MFCCs_{19}$
- $MFCCs_{6}$

Pour l'axe 2 les variables corrélées négativement sont :

- $MFCCs_{10}$
- $MFCCs_{14}$

Pour l'axe 3 les variables corrélées positivement sont :

- $MFCCs_{18}$

Pour l'axe 4 les variables corrélées positivement sont :

- $MFCCs_{21}$

Pour l'axe 3 les variables corrélées négativement sont :

- $MFCCs_{16}$
- $MFCCs_{5}$

Pour l'axe 4 les variables corrélées négativement sont :

- $MFCCs_{19}$

On observe donc une opposition des coefficients cepstraux corrélés positivement avec ceux corrélés négativement sur les axes principaux. On remarque également que le premier axe oppose les deux groupes de corrélation (figure 12), c'est-à-dire deux groupes de variables corrélées entre elles et qui s'opposent. De même le second, troisième et quatrième axe sont des axes d'opposition.

On s'intéresse maintenant aux variables qui contribuent le plus aux axes, c'est-à-dire celles qui contribuent au dessus du seuil de $100/NbVars = 4.54\%$ qui représente la contribution moyenne attendue. Pour la première composante, les variables qui contribuent le plus, classées de la plus à la moins importante sont :

- $MFCCs_{22}$, $MFCCs_{17}$, $MFCCs_{11}$, $MFCCs_{13}$, $MFCCs_{15}$, $MFCCs_{9}$, $MFCCs_{7}$, $MFCCs_{20}$, $MFCCs_{4}$, $MFCCs_{14}$, $MFCCs_{19}$

Pour la seconde composante :

- $MFCCs_{12}$, $MFCCs_{10}$, $MFCCs_{3}$, $MFCCs_{14}$, $MFCCs_{2}$, $MFCCs_{16}$, $MFCCs_{5}$, $MFCCs_{8}$

Pour la troisième composante :

- $MFCCs_{18}$, $MFCCs_{16}$, $MFCCs_{5}$, $MFCCs_{3}$, $MFCCs_{2}$, $MFCCs_{21}$, $MFCCs_{8}$

Pour la quatrième composante :

- $MFCCs_{21}$, $MFCCs_{19}$, $MFCCs_{9}$, $MFCCs_{7}$, $MFCCs_{6}$, $MFCCs_{11}$

Nous avons donc sur l'axe 1 qui représente les individus qui ont de fortes valeurs de $MFCCs_i$ pour $i = \{4, 9, 13, 16, 17, 22\}$ et de faibles valeurs pour les $MFCCs_j$ pour $j = \{7, 11, 14, 15, 19, 20\}$
L'axe 2 qui représente les individus qui ont de fortes valeurs de $MFCCs_i$ pour $i = \{2, 3, 12\}$ et de faibles valeurs pour les $MFCCs_j$ pour $j = \{10, 14\}$
L'axe 3 qui représente les individus qui ont de fortes valeurs de $MFCCs_i$ pour $i = \{2, 3, 12\}$ et de faibles valeurs pour les $MFCCs_j$ pour $j = \{10, 14\}$
L'axe 4 qui représente les individus qui ont de fortes valeurs de $MFCCs_i$ pour $i = \{2, 3, 12\}$ et de faibles valeurs pour les $MFCCs_j$ pour $j = \{10, 14\}$

Nous avons représenter les graphes de nos individus coloriés en fonction des 3 labels : Famille, Genre, Espèce sur les figures 17 à 22

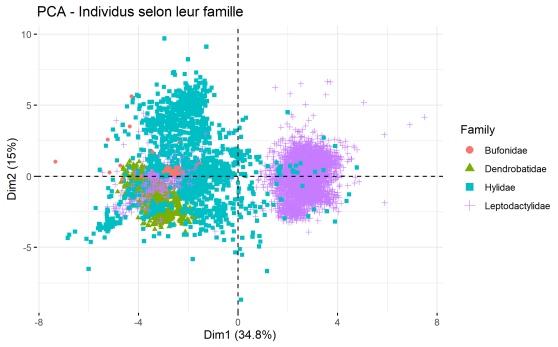

Fig. 17 – Individus selon Family - axes 1 et 2

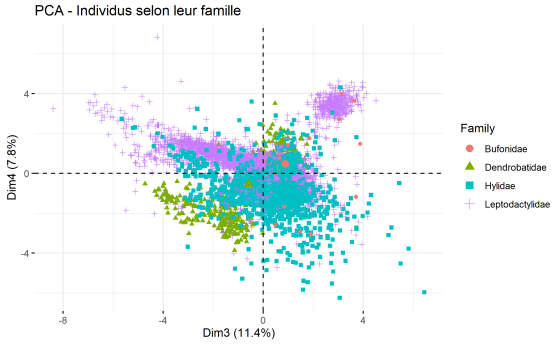

Fig. 18 – Individus par Family - axes 3 et 4

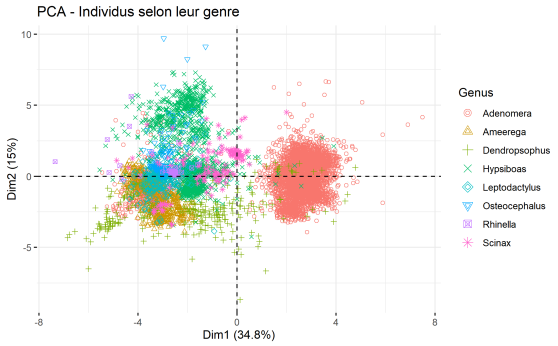

Fig. 19 – Individus par Genus - axes 1 et 2

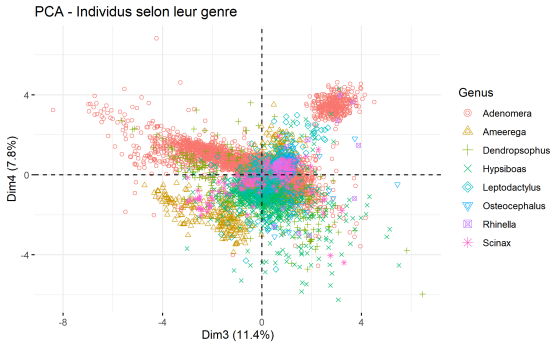

Fig. 20 – Individus par Genus - axes 3 et 4

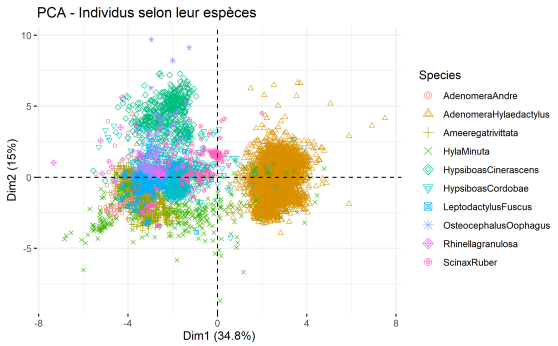

Fig. 21 – Individus par Species - axes 1 et 2

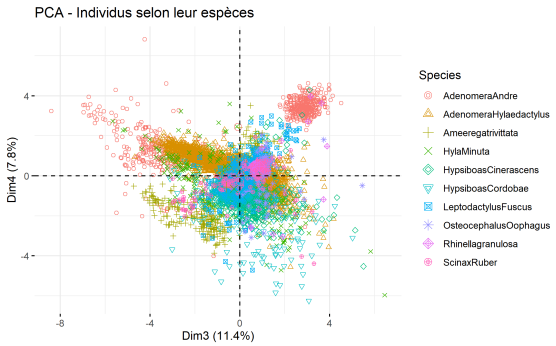

Fig. 22 – Individus par Species - axes 3 et 4

On observe que les familles Dendrobalidae, Hylidae et Bufonidae sont plutôt bien représentées sur notre 1, ce sont donc des familles avec de faibles valeurs pour $i = \{4, 9, 13, 16, 17, 22\}$ et de fortes valeurs pour $j = \{7, 11, 14, 15, 19, 20\}$. On note également que la famille des Leptodactylidae se scinde en deux groupes distincts, l'un très bien représenté sur l'axe 1 et isolé des autres familles, et l'autre confondu avec les trois premières. Notre ACP permet donc de bien différencier les grenouilles membres de la famille des Leptodactylidae qui sont situées à droite sur notre graphique. En regardant maintenant les genres, on observe que ce sont surtout les grenouilles du genre Adenomera qui sont bien caractérisées par notre ACP, alors que le genre Ameerega est confondu avec les autres espèces. On remarque également que le groupe des genres Dendropsophus et Hypsiboa qui s'opposent sur l'axe 2. Et en regardant les espèces, on remarque que ce sont les grenouilles Adenomera Hylaedactyla qui sont bien représentées par l'axe 1 et se démarquent des autres grenouilles et ce sont les grenouilles Hypsiboa Cinerascens qui s'opposent aux grenouilles Hyla Minuta sur l'axe 2.

Pour les axes 3 et 4 il est très difficile de caractériser nos groupes comme précédemment. L'ACP n'est donc pas la meilleure approche si l'objectif est de caractériser nos grenouilles selon leurs labels. Mais elle permet de mettre en avant les syllabes communes.

4 Problème 3 : Graines

4.1 Présentation des données

Pour notre troisième problème, nous souhaitons illustrer l'Analyse Factorielle Discriminante, et plus particulièrement son approche décisionnelle. Nous utiliserons le jeu de données Seeds [8].

Ce jeu de données est composé de 210 observations et de 8 variables : 7 quantitatives et 1 qualitative. Celles-ci sont présentées dans la table 4.

Nom des variables	Description
Area (A)	aire du grain de blé
Perimeter (P)	Périmètre du grain
Compactness ($C = 4\pi A/P^2$	Compacité
Length of Kernel (LK)	Longueur du grain
Width of Kernel (WK)	Largeur du grain
Asymetry Coefficient (AC)	Coefficient d'asymétrie
Length of Kernel groove (LKG)	Longueur de la rainure du grain
Variety	variété : 1 (Kama), 2 (Rosa) and 3 (Canadian)

TABLE 4 – Répartition des observations selon la famille, genre et espèces

Nos 210 observations ont été récupérées par une sélection aléatoire de grains, avec 70 grains par variétés. Les variables quantitatives proviennent de mesures des caractéristiques géométriques des noyaux des trois espèces de blé : Kama, Rosa et Canadian. Elles ont été obtenues à partir d'une méthode de visionnage par rayon-X doux, c'est-à-dire des rayon-X ayant une longueur d'onde de $1nm$, une fréquence de $300PHz$ et une énergie par photon de $1.24KeV$. Cette méthode est non destructive et considérablement moins chère que les autres techniques d'imagerie et pourrait potentiellement servir de moyen pour identifier les variétés de blé. Les images ont été ensuite enregistrées sur des plaques KODAK à rayon-X de 13 par 18cm.

8. lien vers le jeu de données : http ://archive.ics.uci.edu/ml/datasets/Seeds

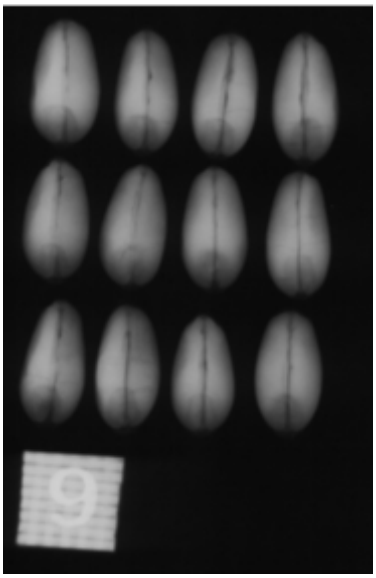

Fig. 23 – PB3 - Illustration d'une plaque KODAK et des grains

Nous pouvons voir dans la figure 24 une répartition des 3 variétés de graines une fois les observations projetées dans le plan principal d'une ACP. On observe une distinction assez nette des classes.

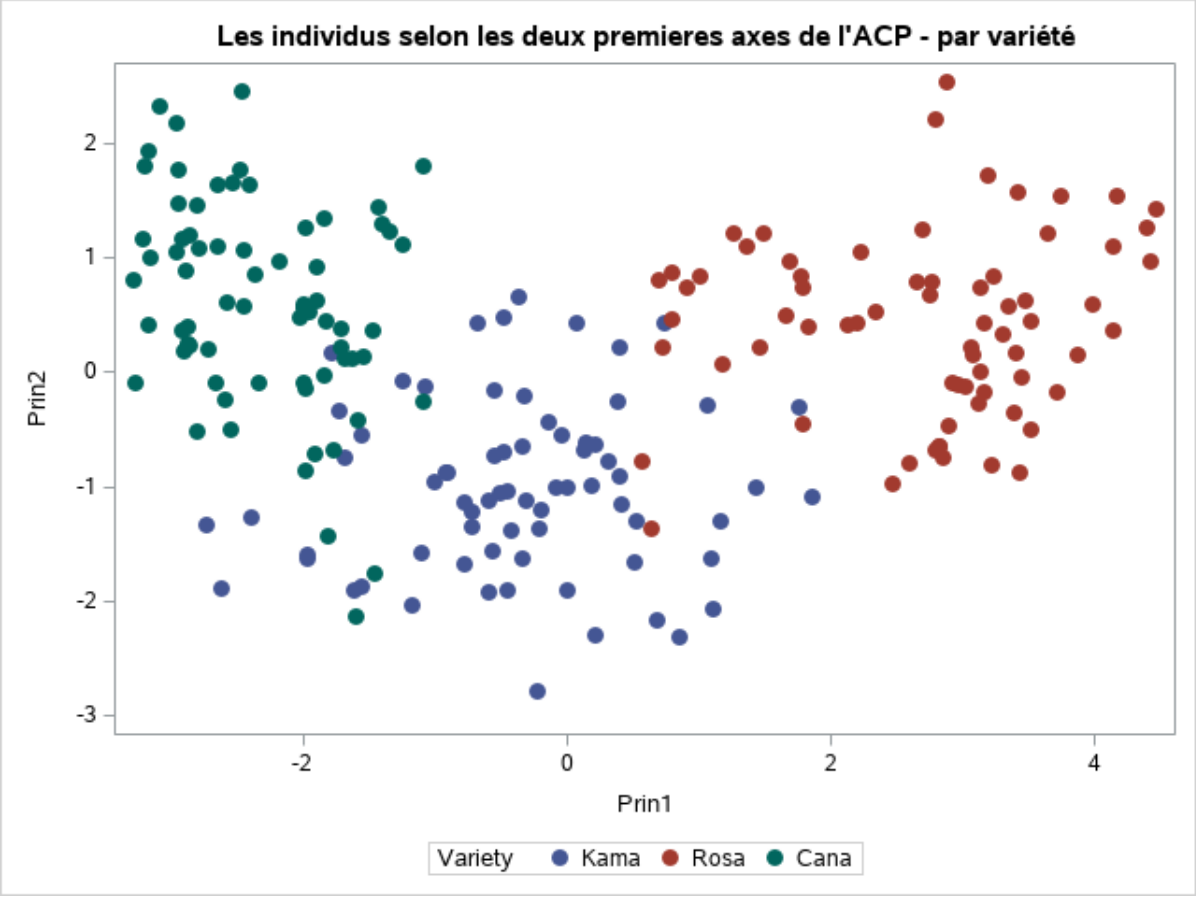

Fig. 24 – PB3 - Classification des variétés après ACP sur 2 axes

4.2 Analyse Factorielle Discriminante - Décisionnelle

Oublions pour la suite notre ACP et utilisons les variables dans leur espace d'origine. Nous considèrerons 80% des données comme jeu d'apprentissage. Les 20% restants serviront pour le passage du test et pour conclure notre analyse.

On obtient les répartitions suivantes dans nos jeux de données d'entrainement et de test (table5 et 6) :

Variety	Fréquence	Pourcentage
Kama	48	28.57
Rosa	63	37.50
Canadian	57	33.93

TABLE 5 – Fréquence - jeu d'entrainement

Variety	Fréquence	Pourcentage
Kama	22	52.38
Rosa	7	16.67
Canadian	13	30.95

TABLE 6 – Fréquence - jeu de test

On remarque que notre jeu de test a des proportions très différente de celui d'entrainement et du jeu d'origine.

Maintenant, intéressons nous aux variables pouvant discriminer nos grains de blé. Pour cela, on commence premièrement par réaliser un test de significativité du modèle, ici c'est-à-dire que les clusters de nos différentes classes ne sont pas superposés. Pour cela, on aura recours à la statistique du Lambda de Wilks pour tester la séparabilité de nos données avec une MANOVA. On obtient la statistique observée de 0.03725756 associée à une p.value < 0.0001, on rejette donc l'hypothèse que nos clusters sont inséparables (ce qui nous semble pertinent au vu de la figure 24 obtenue lors d'une ACP). On peut coupler ce test à une ANOVA et des méthodes de sélection dites step-wise afin de sélectionner les variables pour notre modèle. Ici, toutes les variables sont significatives. On inclura donc les 7 variables quantitatives dans le modèle.

On souhaite désormais trouver la meilleure méthode pour prédire nos variétés de blé. Pour cela nous avons deux approches : la méthode paramétrique et la non paramétrique.

Pour la première approche, on suppose que nos données sont de loi gaussienne relativement aux classes. Cette hypothèse peut se tester facilement en vérifiant la normalité de chacune de nos variables. Un test de Shapiro-Wilk nous permet de rejeter l'hypothèse de normalité pour chacune de nos variables au niveau de risque $\alpha = 0.05$. Cependant, cette approche est assez robuste elle peut donc être mise en place.

La seconde approche, dite non-paramétrique, est d'utiliser par exemple des estimateurs à noyaux (par exemple gaussien ou uniforme) ou encore la méthode des K - plus proches voisins. L'avantage de ces méthodes est qu'on ne suppose pas la loi de nos données, mais il peut être compliqué de trouver les paramètres de réglages adéquats (taille de la fenêtre pour les estimateurs à noyaux, nombre de voisins pour les KNN).

On va donc chercher la méthode, parmi celles citées plus haut, qui minimise les erreurs de prédiction sur notre jeu d'entrainement par cross-validation LOOCV (Left One Out Cross Validation). Nous avons donc procédé à différents tests entre méthodes paramétriques et non paramétriques afin de déterminer la méthode achevant ce résultat. Ainsi la méthode de prévision

que nous avons retenue est celle des **K - plus proches voisins** avec un paramètre $K = 13$. Les résultats obtenus peuvent être visualisés dans la table 7. Cinq erreurs ont été commises au total, 2 Canadian ont été confondu par des Kana et inversement, 3 Kana ont été confondu par des Canadian.

	Canadian	Kama	Rosa	Total
Taux (en %)	4.17	4.76	0	2.98
A priori	0.2857	0.3750	0.3393	

TABLE 7 – Résultats entrainement

Nous sommes désormais prêt à passer notre test avec notre classifieur réglé afin de voir l'erreur commise. Nous obtenons des résultats de bonne qualité sur nos données avec un taux d'erreur de 1.30% sur notre jeu de test. Ces erreurs ont là aussi été commises entre les variétés Kama et Canadian, où une graine de la première a été confondue avec la seconde. Ces bons résultats viennent principalement du fait que nous avons des clusters bien séparés en terme de distance (distances inter groupes très élevées).

On peut conclure que ces méthodes sont simples à mettre en place et donnent de très bons résultats lorsque nos données sont bien séparables. Dans des cas plus compliqué, c'est-à-dire avec un nombre de variable plus important, il serait intéressant de réaliser en premier lieu une AFD descriptive afin de réduire la dimension de nos données tout en maximisant la représentation des classes. Puis de fit notre modèle sur les axes principaux et d'y projeter notre jeu de test pour pouvoir calculer les erreurs commises.

5 Problème 4 : Nutrimouse

5.1 Présentation des données

Pour notre dernier problème nous allons mettre en place notre dernier outil d'analyse : ACC. Pour cela nous utiliserons les données qui se trouvent dans le package *CCA*[9] du logiciel R. Plus précisément, il s'agit du tableau de données *nutrimouse*.
(Source : P. Martin, H. Guillou, F. Lasserre, S. Déjean, A. Lan, J-M. Pascussi, M. San Cristobal, P. Legrand, P. Besse, T. Pineau - Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. Hepatology, in press, 2007.)

Ces données sont issues d'une étude de nutrition chez les souris. Pour 40 souris, on dispose des groupes de données suivants :

- L'expression de 120 gènes potentiellement liés à des problèmes nutritionnels.
- Des mesures sur la concentration de 21 acides gras hépatiques.

Les 40 souris sont réparties en groupe selon deux facteurs :

- Génotype (2 modalités) : Sauvages et génétiquement modifiées (PPARalpha -/-)
- Régime (diet, 5 modalités) :
  - Huile de maïs et colza (50/50), le régime de référence (REF)
  - Huile de coco hydrogénée, pour un régime en gras saturés (COC)
  - Huile de tournesol, pour un régime riche en acide gras oméga6 (SUN)
  - Huile de lin, pour un régime riche en oméga3 (LIN)
  - Huile de maïs/colza/poisson enrichies (43%/43%/14%) (FISH)

Nous avons 4 souris de chaque type qui sont soumises à chaque régime alimentaire.

Nous souhaitons donc utiliser l'Analyse des Corrélations Canoniques afin d'étudier le lien entre l'expression des gènes et les lipides (acide gras hépatique). Pour ce faire nous sélectionnerons deux groupes de variables que nous noterons X_i et Y_j :

- X_1 à X_{15} : les gènes de 76 à 90 (*PLTP* à *SHP1*)
- Y_1 à Y_{10} : les 10 premiers acides gras (*C14.* à *C18 2n 6*)

Ainsi que les deux variables qualitatives *Diet* et *Genotype* qui seront utilisées à des fins descriptives, c'est-à-dire à séparer nos souris en sous-groupes.

5.2 Analyse préliminaire

Nous commencerons par regarder si nos données sont corrélées entre elles. Pour cela nous représenterons graphiquement les corrélations de nos variables dans la figure 25 ainsi que la significativité de la p-value associée au test d'indépendance. Les cases blanches indiquent une faible corrélation entre les variables ligne et colonne associées, bleues une forte corrélation positive et rouges une forte corrélation négative.

9. https ://cran.r-project.org/web/packages/CCA/index.html

On remarque que nos variables X_i ont une forte corrélation positive entre elles et que nous avons de nombreuses dépendances au sein du groupe, et nous sommes probablement face à de la multicolinéarité. On a assez peu de dépendance entre nos variables X et Y. Celle-ci est principalement caractérisée par les variables $PLTP$, $PMDCI$, les RAR et les RxR du jeu X avec les variables de Y. Le lien entre nos deux jeux de données semble donc assez faible.

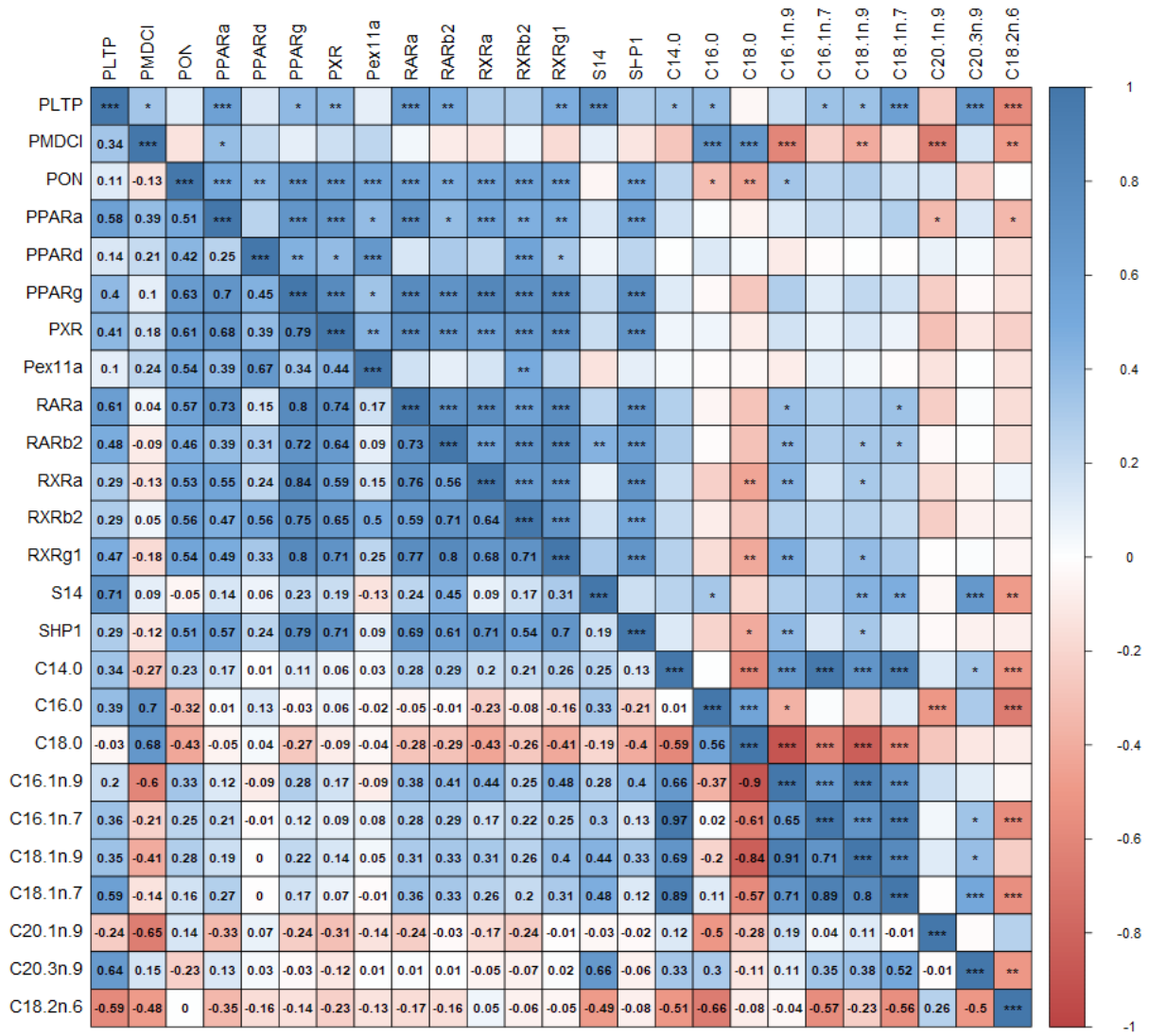

Fig. 25 – PB4 - Corrélation des données nutrimouse

5.3 Analyse des Corrélations Canoniques

Dans le tableau 8 nous donnons une partie des résultats des coefficients de corrélation canoniques pour les 4 premiers axes.

	Corrélation Canonique	Pr > F
1	0.990946	<.0001
2	0.949490	<.0001
3	0.902735	0.0042
4	0.843526	0.0645

TABLE 8 – PB4 - Axes canoniques

Un test sur la statistique du Lambda de Wilks, qui permet de tester si X et Y sont indépendants, nous indique que notre modèle est significatif ($(Pr > F) < .0001$). Nous retiendrons seulement les trois premiers axes pour notre étude, étant donné que seuls ces derniers sont significatifs ($(Pr > F) < 0.01$). De plus leurs valeurs de corrélation canonique sont assez élevées, il semble donc y avoir une forte relation entre les groupes de variables X et Y. Nous pouvons donc pour le moment réduire nos $15 \times 10 = 150$ relations entre les variables originales à 3 relations entre les variables canoniques.

Nous pouvons déjà voir si nous pouvons observer des groupements à partir de nos coefficients canoniques normalisés. Ils définissent la relation linéaire entre les variables du groupe X (resp. Y) et les variables canoniques. On peut les interpréter comme les coefficients d'une régression linéaire, en prenant comme variable à expliquer la variable canonique.

Intéressons nous maintenant à la corrélation entre les variables d'origine et les axes canoniques. Pour cela on peut représenter graphiquement les biplot (figure 26) sur les premières variables canoniques des gènes (respectivement des lipides pour lesquels nous aurions une représentation très semblable).

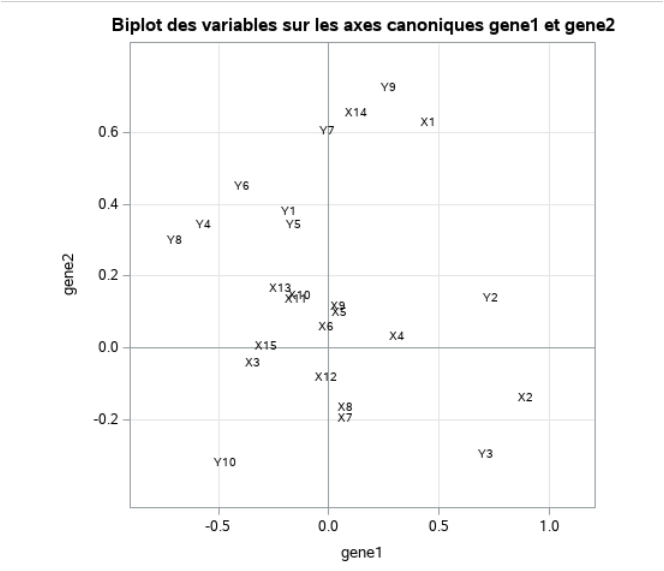

Fig. 26 – PB4 - Biplot selon les axes gene1 et gene2

On peut observer que différents groupes se forment :

- X_1 et X_{14}
- X_7, X_8 et X_{12}
- X_3 et X_{15}
- X_5, X_6 et X_9
- X_{11}, X_{11}, et X_{13}

Ces X_i se comporteront de la même façon par rapport aux Y. De même ces Y_j se comporteront de la même façon par rapport aux X :

- Y_4 et Y_8
- Y_1 et Y_5
- Y_7 et Y_9

De plus pour l'axe canonique gene1 il y a une forte relation positive avec la variable X_2 et de fortes relations négatives avec les variables X_3 et X_{15}. Pour l'axe gene2 il y a de fortes relations positives avec les variables X_1 et X_4. Pour ce qui est de l'axe lipid1, nous avons de fortes relations positives avec les variables Y_2 et Y_3 et de fortes relations négatives avec les variables Y_4, Y_6, Y_8 et Y_{10}. Pour l'axe lipid2, nous avons de fortes relations négatives avec Y_3 et Y_{10} et de fortes relations positives avec les autres, excepté Y_2.

Intéressons nous maintenant à la variance reconstituée par nos axes canoniques.

Leurs propres variables canoniques	Les variables canoniques inverses
Proportion	Proportion
0.0940	0.0923
0.0676	0.0609
0.0253	0.0206

TABLE 9 – PB4 - Proportion de la variance pour les X

Leurs propres variables canoniques	Les variables canoniques inverses
Proportion	Proportion
0.2396	0.2353
0.1969	0.1775
0.0661	0.0538

TABLE 10 – PB4 - Proportion de la variance pour les Y

On remarque que nos variables canoniques ont du mal a reconstituer la variance de nos données initiales, en effet :

- Gene1 reconstitue 9.40% de la variance des X et 23.53% de celle des Y.
- Gene2 reconstitue 6.76% de la variance des X et 17.75% de celle des Y.
- Gene3 reconstitue 2.53% de la variance des X et 5.38% de celle des Y.
- Lipid1 reconstitue 23.96 : % de la variance des Y et 9.23% de celle des X.
- Lipid2 reconstitue 19.69% de la variance des Y et 6.09% de celle des X.

- Lipid3 reconstitue 6.61% de la variance des Y et 2.06% de celle des X.

Et donc,

- 18.69% de la variabilité des X est reconstituée par les variables canoniques des gènes.
- 17.38% de la variabilité des X est reconstituée par les variables canoniques des lipides.
- 46.67% de la variabilité des Y est reconstituée par les variables canoniques des gènes.
- 50.26% de la variabilité des Y est reconstituée par les variables canoniques des lipides.

Notre modèle semble donc bien mieux réussir à reconstituer la variabilité des lipides que celle des gènes. Cela peut être en partie dû à la forte corrélation présente au sein des gènes.

Regardons enfin les coefficients canoniques normalisés :

- L'axe gene1 est principalement défini par de fortes valeurs pour les gènes X_1, X_2, X_6 et de faibles valeurs pour les gènes X_{13}, X_{14} et X_{15}.
- Pour l'axe gene2 de fortes valeurs pour le gène X_1 et de faibles valeurs pour X_7 et X_9.
- Pour l'axe gene3 ce sont principalement de fortes valeurs pour X_4 et des faibles pour X_6.
- Pour lipid1, de fortes valeurs pour Y_7 et des faibles pour Y_4, Y_5 et Y_8.
- Pour lipid2, de fortes valeurs pour Y_7 et des faibles pour Y_3, Y_5 et Y_6.
- Pour lipid3, de fortes valeurs pour Y_3 et Y_5 et des faibles pour Y_1 et Y_2.

Enfin, intéressons nous aux projections des points dans l'espace des axes canoniques. Nous colorons les souris par leurs modalités : régime et génotype.

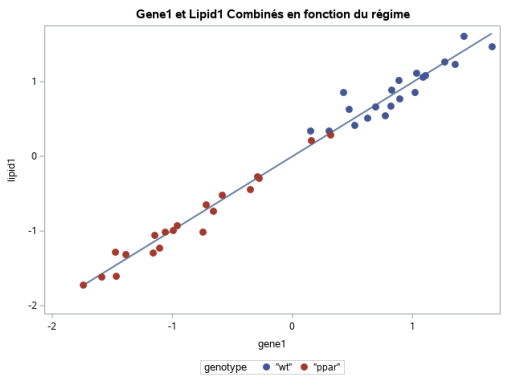

Fig. 27 – Individus par génotype - g1/l1

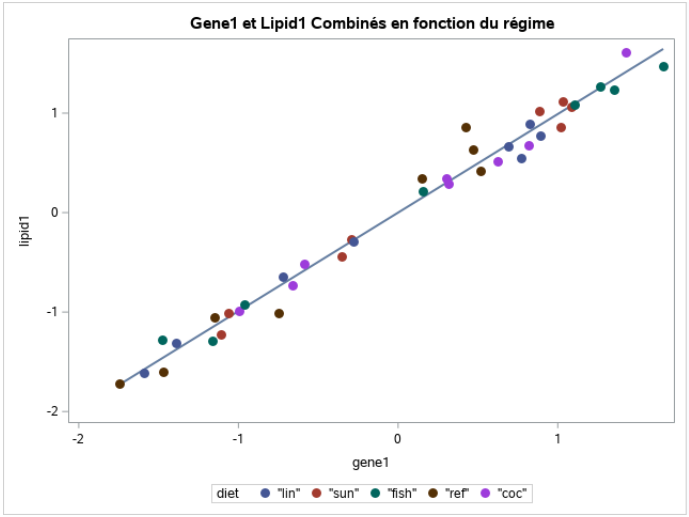

Fig. 28 – Individus par régime - g1/l1

On remarque dans la figure 27 qu'on distingue très bien nos souris sauvages de celles PPAR. Les secondes ont de fortes valeurs de gene1 et lipid1, c'est-à-dire qu'elles ont de fortes valeurs pour les gènes X_1, X_2 et X_6 et l'acide gras Y_7, et de faibles valeurs pour les gènes X_{13}, X_{14} et X_{15} et les acides gras Y_4, Y_5 et Y_8. Et la situation inverse pour les souris sauvages.
On remarque également que parmi les souris $PPAR$, celles qui suivent un régime de type $FISH$ ont des valeurs de gene1 et lipid1 très élevées et se différencient des souris du même génotype.

Si on regarde la figure 29 on observe qu'une partie des souris sous régime coco hydrogénée (62.5%) ont de fortes valeurs sur l'axe gene1 et lipid2, c'est-à-dire de fortes valeurs pour X_1 et Y_7 et de faibles valeurs pour X_7, X_9,Y_3, Y_5 et Y_8.

Il est difficile de distinguer suffisamment les régimes restants pour en dégager un comportement particulier selon les variables X et Y. On pourrait envisager de retirer certaines variables en X fortement corrélées avec les autres et qui le sont peu avec les Y afin de trouver plus facilement des similitudes entre les gènes exprimés et les acides gras hépatiques. De cette façon nous pourrions peut être réussir à caractériser plus facilement les gènes liés aux lipides dans les régimes de mauvaise qualité. Nous pourrions aussi à l'inverse considérer plus de variables en considérant le jeu de données entier pour tenter de déceler d'autres liaisons ou une sélection de variables différentes.

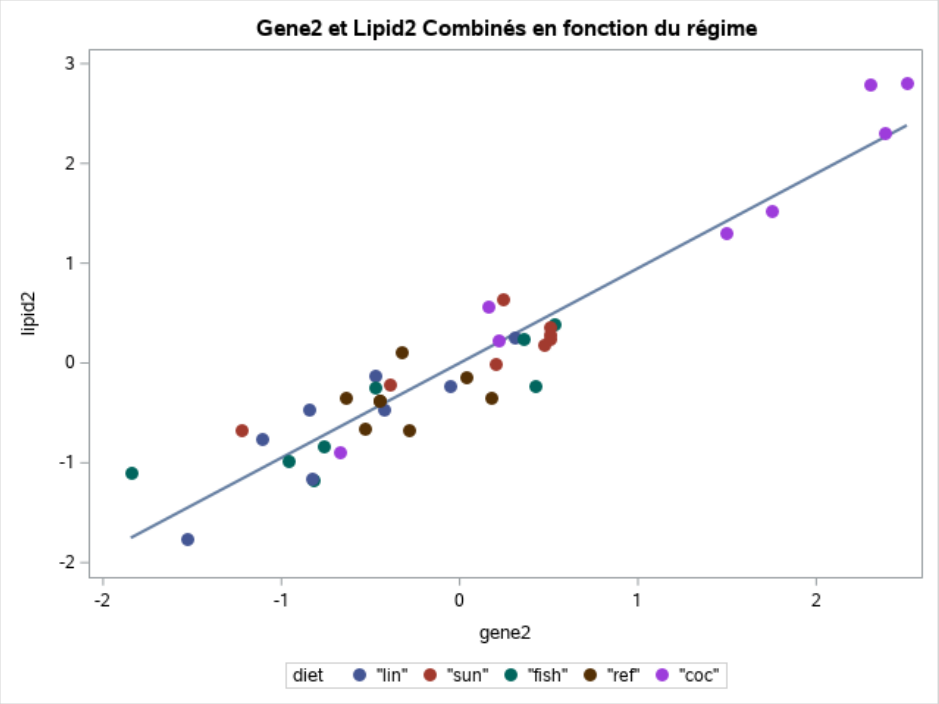

Fig. 29 – Individus par régime - l2/l3