Analyse canonique des corrélations (ACC)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de deux ensembles de données quantitatives observées sur les mêmes individus. Représentations graphiques des individus, des variables et simultanée. Lien avec la régression multivariée et les tests associés.

Rerour au plan du cours.

1 Introduction

L'analyse canonique (A.C.) est une méthode de statistique descriptive multidimensionnelle qui présente des analogies à la fois avec l'analyse en composantes principales (A.C.P.), pour la construction et l'interprétation de graphiques, et avec la régression linéaire, pour la nature des données. L'objectif général de l'A.C. est d'explorer les relations pouvant exister entre deux groupes de variables quantitatives observées sur le même ensemble d'individus. L'étude des relations entre deux groupes de variables constitue la principale particularité de l'A.C. par rapport à l'A.C.P. De ce point de vue, l'A.C. est d'avantage proche de la régression linéaire multiple (explication d'une variable quantitative par un ensemble d'autres variables quantitatives), méthode dont elle constitue, d'ailleurs, une généralisation (on retrouve la régression lorsque l'un des deux groupes de l'A.C. ne comporte qu'une seule variable).

En fait, l'analyse canonique est, sur le plan théorique, la méthode centrale de la statistique descriptive multidimensionnelle, dans la mesure où elle généralise diverses autres méthodes. Outre la régression linéaire, l'A.C. redonne en effet l'analyse factorielle discriminante lorsque l'un des deux groupes de variables est remplacé par les indicatrices d'une variable qualitative. Elle redonne également l'analyse factorielle des correspondances lorsque chacun des deux groupes est remplacé par les indicatrices d'une variable qualitative. Signalons également qu'il existe certaines généralisations de l'A.C. à plus de deux groupes de variables quantitatives et qu'elles permettent de retrouver l'analyse des correspondances multiples (en remplaçant chaque groupe par les indicatrices d'une variable qualitative), ainsi que l'A.C.P. (en ne mettant qu'une seule variable quantitative dans chaque groupe). Nous ne nous intéresserons ici qu'à l'A.C. classique, entre deux groupes de variables quantitatives.

En dépit de sa place centrale au sein des méthodes de statistique multidimensionnelle, pendant longtemps, l'A.C. n'était pas (ou très peu) enseignée dans ces cursus, compte tenu du petit nombre d'applications auxquelles elle donnait lieu. Les choses ont changé, d'abord vers le milieu des années 1990, avec le développement de la régression P.L.S. (*partial least squares*), méthode assez voisine de l'A.C., ensuite, plus récemment, avec l'apparition des données de biopuces, dont certaines relèvent typiquement de l'A.C. quant à leur traitement.

Le logiciel statistique SAS dispose d'une procédure assez complète dédiée à l'A.C. : `CANCORR`. Divers développements de ce chapitre ont pour objectif de mieux saisir la signification de certaines sorties de cette procédure. Les commandes R permettant de mettre en œuvre l'A.C., telles qu'elles seront présentées dans les T.P., ont été quelque peu calquées sur le principe de la procédure `CANCORR`.

2 Approche élémentaire

2.1 Exemple : nutrition chez la souris

C'est encore l'exemple de la nutrition chez la souris qui sera utilisé pour illustrer l'A.C. Nous disposons donc des 40 souris sur lesquelles on s'intéresse maintenant à deux catégories de mesures (de variables) : les expressions des 120 gènes considérés et les proportions de 21 acides gras hépatiques. La question qui va être abordée ici est celle des relations entre ces deux ensembles de variables : certains acides gras sont-ils plus présents lorsque certains gènes sont surexprimés, ou le contraire... La réponse sera essentiellement fournie par les graphiques produits par l'A.C. et dans lesquels seront simultanément représentés gènes et acides gras : il s'agira donc de graphiques relatifs aux variables.

Notons tout de suite qu'il n'est pas très courant de représenter les individus en A.C. Toutefois, compte tenu des particularités de l'exemple considéré ici (petit nombre d'observations et structuration de ces observations selon les fac-

teurs “génotype” et “régime”), nous réaliserons ces graphiques et nous verrons quel est leur intérêt.

2.2 Notations

Dans toute la suite de ce chapitre, on notera n le nombre d'individus considérés (autrement dit, la taille de l'échantillon observé, ici 40), p le nombre de variables (quantitatives) du premier groupe (les gènes) et q le nombre de variables (également quantitatives) du second groupe (les acides gras). On désignera par $\mathbf{X}$ la matrice, de dimension $n \times p$, contenant les observations relatives au premier groupe de variables et par $\mathbf{Y}$ la matrice, de dimension $n \times q$, contenant celles relatives au second groupe. La j-ième colonne de $\mathbf{X}$ ($j = 1, \ldots, p$) contient donc les observations x_i^j de la j-ième variable du premier groupe (notée X^j, il s'agit de l'expression du j-ième gène retenu) sur les n individus considérés ($i = 1, \ldots, n$). De même, la k-ième colonne de $\mathbf{Y}$ ($k = 1, \ldots, q$) contient les observations y_i^k de la k-ième variable du second groupe (notée Y^k, il s'agit du pourcentage relatif au k-ième acide gras retenu).

En A.C., il est nécessaire d'avoir $p \leq n$, $q \leq n$, $\mathbf{X}$ de rang p et $\mathbf{Y}$ de rang q. Par conséquent, dans l'exemple considéré, il a été nécessaire de faire une sélection des gènes et de ne retenir que les plus importants (ceux dont le rôle prépondérant a préalablement été mis en évidence au moyen des techniques exploratoires). Bien que ce ne soit pas imposé par la théorie, nous avons également fait, pour être cohérents, une sélection des acides gras. Finalement, nous avons sélectionné 10 gènes et 11 acides gras hépatiques.

Les gènes sont les suivants :

```
PMDCI THIOL CYP3A11 CYP4A10 CYP4A14 Lpin Lpin1 GSTmu GSTp12 S14.
```

Les acides gras sont les suivants :

```
C16_0 C18_0 C18_1n_7 C18_1n_9 C18_2n_6 C18_3n_3
C20_4n_6 C20_5n_3 C22_5n_3 C22_5n_6 C22_6n_3.
```

Remarque. — On notera que la notation habituelle des acides gras est un peu différente de celle ci-dessus ; ainsi `C18_1n_7` correspond à `C18:1n-7` ; la notation adoptée est nécessaire pour la lecture par le logiciel SAS.

Enfin, sans perte de généralité, on suppose également $p \leq q$ (on désigne donc par premier groupe celui qui comporte le moins de variables). Finalement, nous avons ici : $n = 40$; $p = 10$; $q = 11$.

2.3 Principe général de la méthode

Chaque variable de chacun des deux groupes (les 10 gènes et les 11 acides gras) sont mesurées sur les n individus ($n = 40$). On peut donc associer à chacune un ensemble de 40 valeurs, autrement dit un vecteur de $\mathbb{R}^{40}$ (espace vectoriel que l'on a préalablement muni d'une base adéquate et d'une métrique appropriée). C'est dans cet espace ($\mathbb{R}^{40}$) que l'on peut définir la méthode : elle consiste à rechercher le couple de vecteurs, l'un lié aux gènes, l'autres aux acides, les plus corrélés possible. Ensuite, on recommence en cherchant un second couple de vecteurs non corrélés aux vecteurs du premier et le plus corrélés entre eux, et ainsi de suite. La démarche est donc similaire à celle utilisée en A.C.P. ou en analyse factorielle discriminante. La représentation graphique des variables se fait soit par rapport aux vecteurs liés aux gènes, soit par rapport à ceux liés aux acides (en général, les deux sont équivalentes, au moins pour ce qui est de leur interprétation). Ces vecteurs, obtenus dans chaque espace associé à chacun des deux groupes de variables, sont analogues aux facteurs de l'A.C.P. et sont ici appelés *variables canoniques*. Comme en A.C.P., on peut tracer le cercle des corrélations sur le graphique des variables, ce qui en facilite l'interprétation (dont le principe est le même que pour le graphique des variables en A.C.P.). Des considérations techniques permettent de faire également un graphique pour les individus.

Appelons d le nombre de couples de variables canoniques jugés intéressants, autrement dit la dimension retenue pour les représentations graphiques. On a nécessairement $1 \leq d \leq p$, et on choisit en général d entre 2 et 4. Nous noterons (V^s, W^s) ($s = 1, \ldots, d$) les couples de variables canoniques retenus ; on posera $\rho_s = \text{Cor}(V^s, W^s)$ et on appellera *corrélations canoniques* les coefficients ρ_s qui sont, par construction, décroissants.

3 Approche mathématique

Dans ce paragraphe, nous reprenons, plus en détail et avec plus de rigueur mathématique, les éléments présentés dans le paragraphe précédent. Le lecteur biologiste peu familiarisé avec ces notions de mathématiques pourra donc le parcourir très rapidement et se contenter d'aller y chercher quelques résultats, lorsque nécessaire.

3.1 Représentations vectorielles des données

Comme en A.C.P., on peut considérer plusieurs espaces vectoriels réels associés aux observations.

Tout d'abord, l'espace des variables ; c'est $F = \mathbb{R}^n$, muni de la base canonique et d'une certaine métrique, en général l'identité. À chaque variable X^j est associé un vecteur unique x^j de F dont les coordonnées sur la base canonique sont les x_i^j $(i = 1, \ldots, n)$. De même, à chaque variable Y^k est associé un vecteur unique y^k de F, de coordonnées les y_i^k. On peut ainsi définir dans F deux sous-espaces vectoriels : F_X, engendré par les vecteurs x^j $(j = 1, \ldots, p)$, en général de dimension p, et F_Y, engendré par les vecteurs y^k $(k = 1, \ldots, q)$, en général de dimension q.

Remarque. — Il est courant de munir l'espace vectoriel F de la métrique dite "des poids", définie, relativement à la base canonique, par la matrice diag $(p_1, \ldots, p_n)$, où les p_i $(i = 1, \ldots, n)$ sont des poids (positifs et de somme égale à 1) associés aux individus observés. Lorsque tous ces poids sont égaux, ils valent nécessairement $\frac{1}{n}$ et la matrice définissant la métrique des poids vaut $\frac{1}{n}\mathbf{I}_n$, où $\mathbf{I}_n$ est la matrice identité d'ordre n. Dans ce cas, il est équivalent d'utiliser la métrique identité, ce que nous ferons par la suite, dans la mesure où les individus seront systématiquement équipondérés.

On peut ensuite considérer deux espaces vectoriels pour les individus, $E_X = \mathbb{R}^p$ et $E_Y = \mathbb{R}^q$, eux aussi munis de leur base canonique et d'une certaine métrique. Dans E_X, chaque individu i est représenté par le vecteur x_i, de coordonnées x_i^j $(j = 1, \ldots, p)$ sur la base canonique. De même, dans E_Y, l'individu i est représenté par le vecteur y_i, de coordonnées les y_i^k.

En fait, c'est surtout l'espace F que nous considérerons par la suite, la définition de l'A.C. y étant plus naturelle.

3.2 Retour sur le principe de la méthode

Le principe général de l'A.C. est décrit ci-dessous, dans l'espace des variables F.

Dans un premier temps, on cherche un couple de variables (V^1, W^1), V^1 étant une combinaison linéaire des variables X^j (donc un élément de F_X), normée, et W^1 une combinaison linéaire des variables Y^k (donc un élément de F_Y), normée, telles que V^1 et W^1 soient le plus corrélées possible.

Ensuite, on cherche le couple normé (V^2, W^2), V^2 combinaison linéaire des X^j non corrélée à V^1 et W^2 combinaison linéaire des Y^k non corrélée à W^1, telles que V^2 et W^2 soient le plus corrélées possible. Et ainsi de suite...

Remarque. — Dans la mesure où l'A.C. consiste à maximiser des corrélations, quantités invariantes par translation et par homothétie de rapport positif sur les variables, on peut centrer et réduire les variables initiales X^j et Y^k sans modifier les résultats de l'analyse. Pour des raisons de commodité, on le fera systématiquement. Par conséquent, les matrices $\mathbf{X}$ et $\mathbf{Y}$ seront désormais supposées centrées et réduites (en colonnes).

L'A.C. produit ainsi une suite de p couples de variables (V^s, W^s), $s = 1, \ldots, p$. Les variables V^s constituent une base orthonormée de F_X (les V^s, combinaisons linéaires de variables centrées, sont centrées ; comme elles sont non corrélées, elles sont donc orthogonales pour la métrique identité). Les variables W^s constituent, de même, un système orthonormé de F_Y (ils n'en constituent une base que si $q = p$). Les couples (V^s, W^s), et plus particulièrement les premiers d'entre eux, rendent compte des liaisons linéaires entre les deux groupes de variables initiales. Les variables V^s et W^s sont appelées les *variables canoniques*. Leurs corrélations successives (décroissantes) sont appelées les *coefficients de corrélation canonique* (ou *corrélations canoniques*) et notées ρ_s $(1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_p \geq 0)$.

Remarque. — Toute variable canonique V^{s_0} est, par construction, non corrélée (donc orthogonale) avec les autres variables canoniques V^s, $s \neq s_0$. On peut également montrer que V^{s_0} est non corrélée avec W^s, si $s \neq s_0$ (la même propriété est bien sûr vraie pour toute variable W^{s_0} avec les variables V^s, $s \neq s_0$).

Remarque. — Si nécessaire, on peut compléter le système des variables W^s $(s = 1, \ldots, p)$ pour obtenir une base orthonormée de F_Y dans laquelle les dernières variables W^s $(s = p + 1, \ldots, q)$ sont associées à des coefficients de corrélation canonique nuls ($\rho_s = 0$, pour $s = p + 1, \ldots, q$).

3.3 Propriété

La propriété donnée ici permet, dans la pratique, de déterminer les variables canoniques V^s et W^s en utilisant un algorithme standard de recherche des vecteurs propres d'une matrice.

Dans l'espace vectoriel F muni de la métrique identité, notons $\mathbf{P_X}$ et $\mathbf{P_Y}$ les matrices des projecteurs orthogonaux sur les sous-espaces F_X et F_Y définis plus haut. Les formules usuelles de définition des projecteurs permettent d'écrire ($\mathbf{X}'$ désignant la matrice transposée de $\mathbf{X}$) :

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \; ; \; \mathbf{P_Y} = \mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'.$$

On peut alors montrer la propriété ci-dessous.

PROPOSITION 1. — *Les vecteurs V^s sont les vecteurs propres normés de la matrice $\mathbf{P_X P_Y}$ respectivement associés aux valeurs propres λ_s rangées par ordre décroissant (on peut vérifier que ces valeurs propres sont comprises entre 1 et 0). De même, les vecteurs W^s sont les vecteurs propres normés de la matrice $\mathbf{P_Y P_X}$ respectivement associés aux mêmes valeurs propres λ_s. De plus, les coefficients de corrélation canonique ρ_s sont les racines carrées positives de ces valeurs propres : $\rho_s = \sqrt{\lambda_s}$, $s = 1, \ldots, p$ (le logiciel SAS fournit les corrélations canoniques ρ_s ainsi que leurs carrés λ_s).*

3.4 Retour sur les représentations graphiques

Comme en A.C.P., les représentations graphiques des résultats d'une A.C. se font en dimension réduite (souvent 2 ou 3). Nous noterons d cette dimension, avec : $1 \leq d \leq p$. Plusieurs représentations sont envisageables, à la fois pour les variables et pour les individus.

Représentation des variables dans le sous-espace F_X

Désignons par v^s et w^s les vecteurs de F_X et F_Y respectivement associés aux variables canoniques V^s et W^s.

Dans F_X, on considère la base orthonormée $(v^1, \ldots, v^p)$ que l'on restreint à $(v^1, \ldots, v^d)$ pour les représentations graphiques.

On peut tout d'abord représenter chacune des variables initiales X^j au moyen de ses coordonnées sur les v^s. Ces coordonnées s'obtiennent en calculant les produits scalaires $< x^j, v^s >$, $j = 1, \ldots, p$, $s = 1, \ldots, d$. Les variables X^j étant centrées et réduites, les vecteurs x^j sont centrés et normés (et il en va de même pour les vecteurs v^s), de sorte que ces produits scalaires sont égaux aux corrélations entre variables initiales X^j et variables canonique V^s (au coefficient n près, puisqu'on a considéré la métrique identité).

Dans le même espace, on peut également représenter les variables de l'autre groupe, les Y^k, en projetant tout d'abord les vecteurs y^k dans F_X, au moyen de $\mathbf{P_X}$, puis en prenant le produit scalaire de ces projections avec les vecteurs v^s. On doit donc calculer pour cela les produits scalaires

$$< \mathbf{P_X}(y^k), v^s > = < y^k, \mathbf{P_X}(v^s) > = < y^k, v^s >,$$

encore égaux aux corrélations entre les variables initiales Y^k et les variables canoniques V^s.

Dans la mesure où le graphique ainsi obtenu est "bon" (sur ce point, voir plus loin), on peut l'utiliser pour interpréter les relations (proximités, oppositions, éloignements) entre les deux ensembles de variables. Par construction, ce graphique représente les corrélations entre les variables canoniques V^s et les variables initiales X^j et Y^k, corrélations à la base de son interprétation. On peut aussi conforter cette interprétation en utilisant les coefficients de corrélation linéaire entre variables X^j, entre variables Y^k, et entre variables X^j et Y^k. Tous ces coefficients sont en général fournis par les logiciels.

Représentation des variables dans le sous-espace F_Y

De façon symétrique, on restreint le système $(w^1, \ldots, w^p)$ de F_Y aux premières variables $(w^1, \ldots, w^d)$, par rapport auxquelles on représente aussi bien les variables initiales X^j que les Y^k, selon le même principe que celui décrit ci-dessus (les coordonnées sont les corrélations).

Là encore, dans la mesure où ce graphique est "bon", il permet d'interpréter les relations entre les deux ensembles de variables.

Les deux graphiques (dans F_X et dans F_Y) ayant la même qualité et conduisant aux mêmes interprétations, un seul suffit pour interpréter les résultats d'une analyse.

Représentation des individus

Dans chacun des espaces relatifs aux individus (E_X et E_Y), il est encore possible de faire une représentation graphique de ces individus en dimension d, ces deux représentations graphiques étant comparables (d'autant plus comparables que les corrélations canoniques sont élevées).

En fait, on peut vérifier que les coordonnées des individus sur les axes canoniques pour ces deux représentations sont respectivement données par les

lignes des matrices $\mathbf{V_d}$ (dans E_X) et $\mathbf{W_d}$ (dans E_Y), $\mathbf{V_d}$ et $\mathbf{W_d}$ désignant les matrices $n \times d$ dont les colonnes contiennent les coordonnées des d premières variables canoniques sur la base canonique de F.

Choix de la dimension

Comme dans toute méthode factorielle, différents éléments doivent être pris en compte pour le choix de la dimension d dans laquelle on réalise les graphiques (et dans laquelle on interprète les résultats).

- Tout d'abord, il est clair que d doit être choisi petit, l'objectif général de la méthode étant d'obtenir des résultats pertinents dans une dimension réduite ; ainsi, le plus souvent, on choisit d égal à 2, 3 ou 4.
- Plus l'indice de dimension s augmente, plus la corrélation canonique ρ_s diminue ; or, on ne s'intéresse pas aux corrélations canoniques faibles, puisqu'on cherche à expliciter les relations entre les deux groupes de variables ; par conséquent, les dimensions correspondant à des ρ_s faibles peuvent être négligées.
- Le pourcentage que chaque valeur propre λ_s représente par rapport à la somme de toutes les valeurs propres, c'est-à-dire par rapport à la trace de la matrice diagonalisée, facilitent également le choix de d (voir la remarque 5).

4 Compléments : analyse canonique et régression multivariée

L'objectif principal de ce paragraphe est de donner une idée, à l'utilisateur du logiciel SAS, du principe des tests figurant dans la procédure `CANCORR`, celle qui permet de réaliser l'analyse canonique. Accessoirement, ce paragraphe introduit la régression multivariée et fait le lien entre cette technique et l'analyse canonique.

On notera que les tests présentés ici sont des tests statistiques classiques dans le contexte de l'analyse multivariée, que ce soit l'analyse canonique, la régression multivariée, l'analyse de variance multivariée (la MANOVA), ou même l'analyse discriminante. Ils apparaissent ainsi dans toutes les procédures du logiciel SAS permettant de mettre en œuvre ces méthodes.

Le lecteur peu familiarisé avec les méthodes multivariées pourra néanmoins sauter ce paragraphe.

4.1 Introduction

Ouvrages et logiciels anglo-saxons de statistique présentent souvent l'analyse canonique parallèlement à la régression linéaire multivariée (régression d'un ensemble de variables Y^k, à expliquer, sur un autre ensemble de variables X^j, explicatives). Cette approche est, en fait, assez naturelle, dans la mesure où les données sont de même nature dans les deux méthodes et où l'on cherche, dans l'une comme dans l'autre, des relations linéaires entre variables.

Il convient toutefois de noter les deux différences fondamentales entre les deux méthodes : contrairement à ce qu'il se passe en A.C., les deux ensembles de variables X^j et Y^k ne sont pas symétriques en régression, puisqu'il s'agit d'expliquer les variables Y^k au moyen des variables X^j ; d'autre part, toujours en régression, on suppose la normalité des variables réponses Y^k, alors qu'aucune hypothèse de cette nature n'est nécessaire en A.C. L'avantage de cette hypothèse (lorsqu'elle est "raisonnable") est de permettre de réaliser des tests dans le modèle de régression.

4.2 Le modèle de régression multivariée

Le modèle de régression multivariée des variables Y^k sur les variables X^j s'écrit :

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} \; ;$$

les matrices $\mathbf{Y}$, $n \times q$, et $\mathbf{X}$, $n \times p$, sont celles introduites en A.C. ; $\mathbf{B}$ est la matrice $p \times q$ des paramètres inconnus, à estimer (les coefficients de régression) ; $\mathbf{U}$ est la matrice $n \times q$ des erreurs du modèle. Chaque ligne U_i de $\mathbf{U}$ est un vecteur aléatoire de $\mathbb{R}^q$ supposé $\mathcal{N}_q(0, \Sigma)$, les U_i étant indépendants (Σ est une matrice inconnue, à estimer, supposée constante en i).

L'estimation maximum de vraisemblance de $\mathbf{B}$ conduit à la solution :

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

On appelle alors *valeurs prédites* (de $\mathbf{Y}$ par le modèle) les quantités :

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{P_X}\mathbf{Y} \; ;$$

d'autre part, on appelle *résidus* les quantités :

$$\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y}$$

(dans l'écriture ci-dessus, $\mathbf{P}_{\mathbf{X}}^{\perp}$ désigne, dans $\mathbb{R}^n$, le projecteur orthogonal sur le sous-espace supplémentaire orthogonal à F_X dans $\mathbb{R}^n$; on sait que ce projecteur s'écrit : $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}$).

4.3 Matrices intervenant dans les tests

Dans le cadre du modèle gaussien, on peut tester la significativité du modèle en généralisant le test de Fisher, bien connu dans le cas unidimensionnel. Au numérateur de la statistique de Fisher figure la norme carrée du vecteur $\hat{y} - \overline{y}$, ici remplacée par $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ (cette matrice est centrée). Au dénominateur figure la norme carrée des résidus, ici remplacée par $\hat{\mathbf{U}}'\hat{\mathbf{U}}$ (on néglige, pour l'instant, les degrés de liberté de ces quantités). La statistique de Fisher est donc remplacée par le produit matriciel $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}(\hat{\mathbf{U}}'\hat{\mathbf{U}})^{-1}$. Comme on a $\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$, il vient : $\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{H}$ (la notation $\mathbf{H}$ est standard, car cette quantité est liée à l'hypothèse nulle testée). D'autre part, $\hat{\mathbf{U}} = \mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y}$ entraîne : $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y} = \mathbf{E}$ (il s'agit encore d'une notation standard, cette matrice représentant les erreurs du modèle). Les tests multidimensionnels de significativité du modèle sont ainsi basés sur l'étude des valeurs propres soit du produit matriciel

$$\mathbf{H}\mathbf{E}^{-1} = (\mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y})(\mathbf{Y}'\mathbf{P}_{\mathbf{X}}^{\perp}\mathbf{Y})^{-1},$$

soit encore du produit $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$, les valeurs propres de ces deux matrices se déduisant les unes des autres. Développons le second produit matriciel :

$$\mathbf{H} + \mathbf{E} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y} + \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\mathbf{Y} = \mathbf{Y}'\mathbf{Y};$$

d'où :

$$\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \mathbf{Y}'\mathbf{P}_{\mathbf{X}}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1},$$

matrice ayant les mêmes valeurs propres que

$$\mathbf{P}_{\mathbf{X}}\mathbf{Y}(\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}' = \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Y}},$$

c'est-à-dire les λ_s ($s = 1, \ldots, p$), carrés des corrélations canoniques.

Remarque. — On peut vérifier (le résultat est classique) que les valeurs propres de la matrice $\mathbf{H}\mathbf{E}^{-1}$ valent $\dfrac{\lambda_s}{1-\lambda_s}$. Ces valeurs propres sont fournies par le logiciel SAS, ainsi que les pourcentages (et les pourcentages cumulés) qu'elles représentent par rapport à leur somme, trace de la matrice $\mathbf{H}\mathbf{E}^{-1}$.

En interprétant ces pourcentages comme la part d'inertie globale du nuage des individus restituée par les différents axes canoniques (ce qu'elles sont, par exemple, en analyse factorielle discriminante), ces quantités facilitent le choix de la dimension d retenue pour les graphiques et les interprétations.

4.4 Tests

Il existe plusieurs tests de significativité du modèle de régression multivariée, en général équivalents (au moins au niveau des décisions qu'ils entraînent). Ces tests sont les généralisations classiques du test de Fisher au cas multivarié (on les retrouve, par exemple, en analyse de variance multivariée) et sont des tests asymptotiques. Le logiciel SAS fournit les trois premiers ci-dessous, mais pas le quatrième. Il fournit également le test de Roy, basé sur la plus grande valeurs propre de la matrice $\mathbf{H}\mathbf{E}^{-1}$, soit $\dfrac{\lambda_1}{1-\lambda_1}$, mais ce test est à déconseiller.

- Le test de Wilks, adaptation du test du rapport des vraisemblances, est basé sur la statistique

$$\Lambda = \prod_{s=1}^{p}(1-\lambda_s) = \prod_{s=1}^{p}(1-\rho_s^2).$$

- Le test de la trace de Pillai est basé sur la statistique

$$Z = \text{trace } \mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \sum_{s=1}^{p}\lambda_s.$$

- Le test de la trace de Lawley-Hotelling est basé sur la statistique

$$T^2 = \text{trace } \mathbf{H}\mathbf{E}^{-1} = \sum_{s=1}^{p}\frac{\lambda_s}{1-\lambda_s}.$$

- Le test du khi-deux est basé sur la statistique

$$K = -[(n-1) - \frac{1}{2}(p+q+1)]\ln\prod_{s=1}^{p}(1-\lambda_s).$$

Le test du khi-deux présente l'avantage d'être directement utilisable, puisqu'on compare la statistique K à une loi de khi-deux à pq degrés de libertés (il s'agit d'un test approché).

Dans les trois autres tests ci-dessus, on doit transformer la statistique (Λ, Z ou T^2) pour obtenir un test de Fisher approché, les transformations étant assez compliquées à expliciter (toutefois, SAS les réalise automatiquement).

Remarque. — Dans un article de 1951, Rao a montré que, dans la plupart des cas, l'approximation de Fisher du test de Wilks est la meilleure. C'est donc le test que nous conseillerons.

Si le modèle de régression est significatif (il en va alors de même pour l'analyse canonique), on peut tester la significativité d'une dimension et de l'ensemble des suivantes, en particulier pour guider le choix de la dimension en A.C. Ainsi, supposons que les corrélations canoniques soient significatives depuis la première jusqu'à la k-ième ($1 \leq k \leq p$). On peut alors tester l'hypothèse nulle

$$\{H_0 : \rho_{k+1} = \cdots = \rho_p = 0\} \ (\Longleftrightarrow \{H_0 : d = k\})$$

contre l'alternative

$$\{H_1 : \rho_{k+1} > 0\} \ (\Longleftrightarrow \{H_1 : d > k\}).$$

Pour cela, il faut adapter soit le test de Wilks, soit le test du khi-deux.

Pour le test de Wilks, il suffit de faire le produit des quantités $(1 - \lambda_s)$ de l'indice $k + 1$ à l'indice p et d'adapter la transformation en fonction des nouvelles dimensions. SAS le fait automatiquement. Pour le test du khi-deux, il faut considérer la statistique

$$K_k = -[(n - 1 - k) - \frac{1}{2}(p + q + 1) + \sum_{s=1}^{k} \frac{1}{\lambda_s}] \ln \prod_{s=k+1}^{p} (1 - \lambda_s)$$

et la comparer à une loi de khi-deux à $(p - k)(q - k)$ degrés de liberté.

Remarque. — Dans l'utilisation de ces tests, il convient de ne pas perdre de vue d'une part qu'il s'agit de tests asymptotiques (d'autant meilleurs que la taille de l'échantillon, n, est grande), d'autre part qu'ils ne sont valables que sous l'hypothèse de normalité des variables Y^k.

5 Exemple : nutrition chez la souris

5.1 Traitements préliminaires

Nous donnons ci-dessous les statistiques élémentaires relatives aux deux groupes de variables. Les corrélations entre gènes se trouvent en Annexe A, celles entre acides en Annexe B.

Variable	N	Mean	Std Dev	Minimum	Maximum
PMDCI	40	-0.7673	0.1861	-1.07	-0.44
THIOL	40	-0.4110	0.2125	-0.90	-0.03
CYP3A11	40	-0.5083	0.2556	-1.02	0.06
CYP4A10	40	-0.9798	0.2237	-1.33	-0.48
CYP4A14	40	-0.9930	0.2460	-1.29	-0.15
Lpin	40	-0.7533	0.1735	-1.13	-0.48
Lpin1	40	-0.7648	0.1638	-1.10	-0.49
GSTmu	40	-0.1190	0.1504	-0.44	0.23
GSTpi2	40	0.2298	0.1422	0	0.55
S14	40	-0.8068	0.2008	-1.05	-0.25

Variable	N	Mean	Std Dev	Minimum	Maximum
C16_0	40	23.03	3.57	14.65	29.72
C18_0	40	6.75	2.64	1.68	10.97
C18_1n_7	40	4.43	3.38	1.53	15.03
C18_1n_9	40	25.27	7.34	14.69	41.23
C18_2n_6	40	15.28	8.76	2.31	40.02
C18_3n_3	40	2.89	5.83	0	21.62
C20_4n_6	40	5.28	4.46	0.75	15.76
C20_5n_3	40	1.79	2.59	0	9.48
C22_5n_3	40	0.87	0.86	0	2.58
C22_5n_6	40	0.44	0.66	0	2.52
C22_6n_3	40	5.91	5.33	0.28	17.35

Remarque. — Les valeurs ci-dessus sont relatives aux variables brutes (aux données initiales). Comme indiqué dans la remarque 3, ces variables ont ensuite été centrées et réduites avant la réalisation de l'A.C.

5.2 Analyse canonique

Généralités

Les premiers résultats fournis par une A.C. sont les corrélations croisées entre les deux groupes de variables. Nous donnons ces corrélations dans l'an-

nexe C.

Ensuite sont données les corrélations canoniques reproduites ci-dessous.

	Canonical Correlation
1	0.96
2	0.93
3	0.91
4	0.86
5	0.79
6	0.72
7	0.61
8	0.41
9	0.25
10	0.04

On notera que “le plus petit” groupe ne comportant que 10 variables, on ne peut déterminer que 10 corrélations canoniques. L’objectif principal de l’A.C. étant d’étudier les relations entre variables des deux groupes, on peut noter ici qu’il existe effectivement des relations fortes entre ces deux groupes, puisque les premiers coefficients canoniques sont très élevés. Compte tenu des valeurs importantes des premiers coefficients, on peut raisonnablement se contenter de deux ou trois dimensions pour étudier les résultats fournis par la méthode et nous avons choisi ici seulement deux dimensions, compte tenu qu’il s’agit essentiellement d’une illustration.

Remarque. — Les valeurs propres de la matrice $\mathbf{HE}^{-1}$ et les pourcentages d’inertie restitués par les différentes dimensions sont les suivants :

	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
	Eigenvalue	Difference	Proportion	Cumulative
1	12.7583	6.1471	0.4167	0.4167
2	6.6111	1.7001	0.2159	0.6326
3	4.9111	2.1433	0.1604	0.7930
4	2.7678	1.1107	0.0904	0.8833
5	1.6571	0.5892	0.0541	0.9375
6	1.0679	0.4877	0.0349	0.9723
7	0.5802	0.3792	0.0189	0.9913
8	0.2010	0.1369	0.0066	0.9978
9	0.0641	0.0623	0.0021	0.9999
10	0.0018		0.0001	1.0000

Par ailleurs, les tests de Wilks, de significativité de chaque dimension, sont les suivants :

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.00003857	4.08	110	155.53	<.0001
2	0.00053068	3.31	90	145.91	<.0001
3	0.00403909	2.77	72	135.32	<.0001
4	0.02387531	2.21	56	123.78	0.0001
5	0.08995724	1.78	42	111.33	0.0090
6	0.23902627	1.41	30	98	0.1087
7	0.49427788	0.99	20	83.865	0.4795
8	0.78104952	0.56	12	69.081	0.8636
9	0.93806320	0.29	6	54	0.9380
10	0.99819295	0.03	2	28	0.9750

On voit que le choix optimal de la dimension serait probablement $d = 4$ (ne pas oublier que ces tests sont asymptotiques et que nous avons $n = 40$). Pour simplifier, nous ne présentons, par la suite, que les graphiques selon les deux premières dimensions.

Graphique des individus

Dans un premier temps, nous avons réalisé le graphique des individus (les 40 souris) relativement aux deux premiers axes canoniques de l’espace des gènes E_X (Fig. 1). En général, dans une A.C., ce graphique sert seulement à contrôler l’homogénéité de l’ensemble des individus (absence d’individus atypiques par exemple). Ici, dans la mesure où les individus proviennent d’un plan d’expériences à deux facteurs croisés (le génotype et le régime), il est intéressant de regarder si l’on retrouve la structure de ce plan. On notera que cela est très net en ce qui concerne le génotype et encore assez net pour ce qui est du régime (en fait, la sélection des gènes a été réalisée de telle sorte que ceux retenus soient le plus structurant possible pour ces deux facteurs ; le résultat, s’il est rassurant, n’a donc rien d’extraordinaire).

Signalons pour terminer qu’on a également réalisé le graphique des individus relativement aux deux premiers axes de l’autre espace (espace des acides gras, E_Y) et qu’il est très semblable à celui-ci.

Graphique des variables

Pour la représentation des variables, nous avons considéré le sous-espace F_X, engendré par les 10 gènes, et nous avons représenté à la fois les gènes et les acides gras relativement aux deux premières variables canoniques, V^1 et V^2 (Fig. 2). Comme indiqué en 3.4, les coordonnées des variables initiales

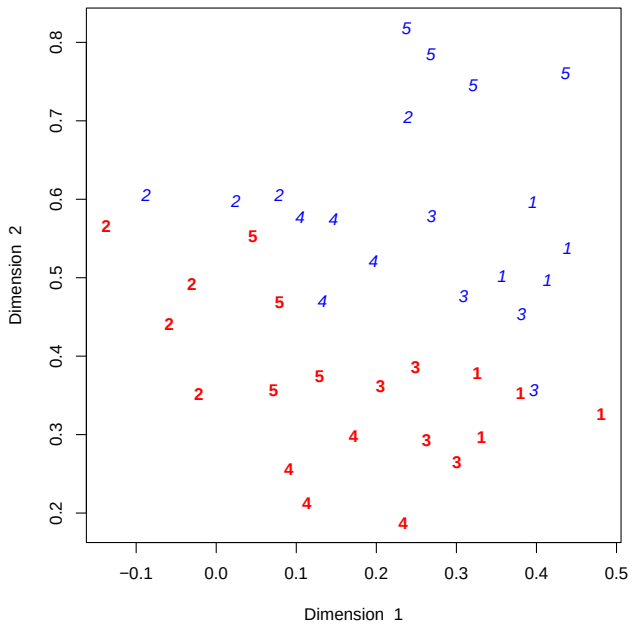

FIGURE 1 – *Souris : représentation des individus (souris) dans l'espace des gènes. Les WT sont en rouge-gras et les PPAR en bleu-italique ; les numéros correspondent aux régimes.*

sont fournies par leur corrélations avec les variables canoniques.

Certaines associations entre gènes et acides gras, en particulier celles correspondant à des points éloignés de l'origine, sont intéressantes à noter. Ainsi peut-on observer que la séparation des génotypes est principalement liée d'une part à l'accumulation préférentielle de l'acide gras `C18_2n_6` chez les souris `PPAR`, au détriment de `C16_0`, de `C18_0` et des acides gras longs polyinsaturés `C20_5n_3` et `C22_6n_3` (les *oméga 3*), d'autre part à la plus forte expression des gènes `THIOL`, `PMDCI`, `CYP3A11` et `GSTpi2` chez les souris `WT` par rapport aux souris `PPAR`. On peut également noter les proximités entre le `C16_0` et le gène `THIOL`, ainsi que les proximités entre `CYP3A11` et `GSTpi2` et les acides gras `C18_0` et `C22_6n_3`. Par ailleurs, l'opposition entre le régime `2-efad` et les régimes `1-dha` et `3-lin` est liée, sous régime `efad`, à l'accumulation d'acides gras monoinsaturés (`C18_1n_9` et `C18_1n_7`) chez les souris des deux génotypes (mais plus marquée chez les souris `PPAR`), accompagnée de la sur-expression du gène `S14` presque exclusivement chez les souris `WT`. Sous régime riche en *Oméga 3* (`1-dha` et `3-lin`), on observe une accumulation préférentielle des acides gras `C20_5n_3` (surtout pour le régime `lin`), `C22_6n_3` (surtout pour le régime `dha`) et `C18_0` accompagnée de régulations positives des gènes `GSTpi2`, `CYP3A11` et des `CYP4A` qui, cependant, se révèlent moins marquées, voire absentes, chez les souris `PPAR`. Enfin, remarquons que la position particulière du régime `5-tsol` chez les souris `PPAR` est liée à l'accumulation extrêmement marquée de `C18_2n_6` dans le foie de ces souris sous le régime `tsol` (sous ce régime, la proportion de `C18_2n_6` est presque deux fois plus importante chez les souris `PPAR` que chez les souris `WT`), soulignant ainsi le rôle primordial de PPARα dans la prise en charge de cet acide gras, que ce soit pour sa dégradation ou pour son utilisation pour la biosynthèse des acides gras longs polyinsaturés de la famille *Oméga 6*.

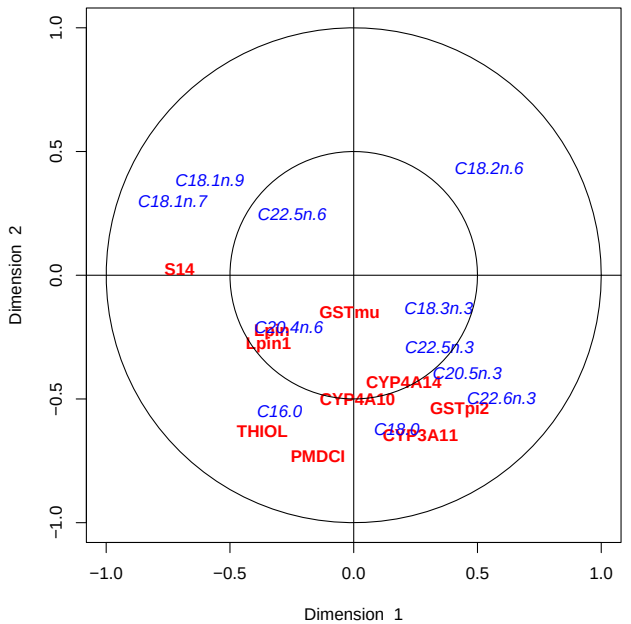

FIGURE 2 – *Souris : représentation des gènes (en rouge-gras) et des acides (en bleu-italique) dans le sous-espace des gènes.*