

Cours analyse canonique des corrélations

Table des matières

Introduction.....	2
I. Pourquoi utiliser l'ACC ?.....	2
II. Hypothèses et conditions d'utilisations	2
III. Explication de la méthode	3
1. Approche élémentaire.....	3
2. Démarche	4
a) Calcul des matrice variance co-variance	4
b) Combinaison linéaire.....	5
c) Définition des variables canoniques.....	5
d) Calcul du coefficient de corrélation canonique.....	5
IV. Démonstration dans R.....	6
1. Importation du jeu de données.....	6
2. Analyse des résultats	7
Premier graphique : Graphique de dispersion avec dimensions canoniques	7
Deuxième graphique : Graphique de dispersion des individus dans l'espace canonique .	8
V. Comparaison avec d'autres méthodes	9
VI. Limites de l'ACC.....	9
Conclusion	10

Introduction

Analyse canonique = méthode descriptive multi-dimensionnelle qui présente des analogies à la fois avec l'**ACP** (construction et interprétation des graphiques) et la **régression linéaire** (nature des données). Elle a été mise au point par Harold Hotelling en 1936.

Objectif : Explorer les relations pouvant exister entre **deux groupes de variables quantitatives** observées sur le même ensemble d'individus. Contrairement à la corrélation simple, qui mesure la relation entre deux variables, l'ACC est utilisée pour **examiner les associations linéaires entre deux groupes de variables**.

AC n'est que très peu appliquée mais elle connaît un essor dans les années 1990 avec le développement de la régression PLS, et plus récemment avec l'apparition des données de biopuces.

I. Pourquoi utiliser l'ACC ?

L'analyse Canonique des Corrélations est une méthode statistique multidimensionnelle qui permet d'explorer les relations entre deux groupes de variables quantitatives observées sur les mêmes individus. Il est pertinent d'utiliser cette méthode lorsque les deux groupes de variables peuvent influencer de manière réciproque les résultats comme dans les études où il faut interpréter des interactions complexes entre plusieurs variables.

Donc, l'ACC est particulièrement utile dans les études où ni l'un ni l'autre des groupes n'est considéré comme dépendant ou explicatif et donc que la symétrie entre deux groupes de variables est essentielle. Par conséquent, l'ACC trouve son application dans des domaines variés, tels que la biologie, la psychologie et l'économie, où les chercheurs explorent souvent les influences réciproques entre groupes de variables observées dans les mêmes conditions.

II. Hypothèses et conditions d'utilisations

L'ACC repose sur l'hypothèse de linéarité et, idéalement, de normalité des variables dans chaque groupe.

Conditions d'utilisation :

- Variables quantitatives → calcul de corrélation linéaire
 - Deux groupes de variables minimum: avec un nombre de variables \leq au nombre d'individus
- Indépendance inter groupes

- Colinéarité intra groupes → maximiser les corrélations entre les combinaisons linéaires des deux groupes

III. Explication de la méthode

1. Approche élémentaire

Nous allons découvrir l'ACC à travers l'exemple suivant. Nous disposons de 40 souris sur lesquelles on s'intéresse à deux catégories de mesures : les expressions des 120 gènes considérés et les proportions de 21 acides gras hépatiques. On cherche à savoir si certains acides gras sont plus présents lorsque certains gènes sont surexprimés.

Deux groupes :


- Gènes : Y_1, Y_2, \dots, Y_q
- Acides gras : X_1, X_2, \dots, X_p

On désigne Y la matrice de dimension $n \times q$ contenant les observations relatives au premier groupe (gènes) et X la matrice de dimension $n \times p$ contenant les observations relatives au deuxième groupe (acides gras). En AC, il est nécessaire d'avoir un **nombre de variables inférieure au nombre d'individus** dans chaque groupe. Par conséquent, dans l'exemple considéré, il est nécessaire de faire une sélection des gènes et de ne retenir que les plus importants. Pour la suite de l'exercice, on en retient 10 et on retient 11 acides gras.


L'ACC peut très bien être utilisé quand le jeu de données est composé de plus de deux groupes de variables et que le nombre de variables par groupe est inégal.

Comment calculer le coefficient de corrélation entre les deux groupes de variable ?

	Gène 1	...	Gène 10	AG 1	...	AG 11
1						
2						
3						
4						



X



Y

Chaque groupe est réduit à une seule variable qui prend la forme d'une **combinaison linéaire**. Ce sont les **variables canoniques** :

- $X_1 = a_1 \text{Gène}_1 + \dots + a_{10} \text{Gène}_{10}$
- $Y_1 = b_1 \text{AG}_1 + \dots + b_{11} \text{AG}_{11}$

Ces combinaisons linéaires sont choisies pour que la **corrélation** entre les variables canoniques soit **maximisée**.

L'ACC cherche ensuite le couple X_2 et Y_2 , avec X_2 une combinaison linéaire des Gènes non corrélée à X_1 et Y_2 une combinaison linéaire de AG non corrélée à Y_1 , telle que X_2 et Y_2 soient le plus corrélées possible. Et ainsi de suite... L'objectif est donc de trouver plusieurs paires de variables canoniques, où chaque paire est orthogonale (non corrélée) avec les autres, tout en maximisant la corrélation entre chaque couple de variables canoniques.

L'AC produit ainsi une suite de p couples de variables (X_s, Y_s) avec $1 < s < p$. Leurs corrélations successives sont appelées les **coefficients de corrélation canonique** et sont notées ρ_s . Les valeurs des corrélation canoniques sont décroissantes : $1 \geq \rho_1 \geq \rho_2 \geq \dots \geq 0$.

2. Démarche

Étapes pour déterminer la corrélation entre les deux groupes :

a) Calcul des matrice variance co-variance

Matrices de covariance intra-groupe + matrices intergroupes

On peut alors calculer les matrices R_v et R_w :

- $R_X = R_{XX}^{-1} R_{XY} R_{YY}^{-1} R_{YX}$
- $R_Y = R_{YY}^{-1} R_{YX} R_{XX}^{-1} R_{XY}$

Où :

- R_{XX} est la matrice de variance-covariance 10×10 des variables du groupe X (Gène1, ..., Gène10),
- R_{YY} est la matrice de variance-covariance 11×11 des variables du groupe Y (AG1, ..., AG11),
- R_{XY} est la matrice de covariance entre les variables de X et Y, c'est une matrice 10×11 (entre X_1, X_2, \dots, X_{10} et Y_1, Y_2, \dots, Y_{11}),
- R_{YX} est simplement la transposée de R_{XY} , c'est une matrice 11×10.

b) Combinaison linéaire

Il est alors possible de calculer les valeurs propres des deux matrices R_x et R_y . Les valeurs propres dans la première colonne sont alors utilisées comme valeur poids dans la combinaison linéaire de chaque groupe. Si un gène a un poids plus élevé qu'un autre cela signifie qu'il est plus fortement appliqué dans la corrélation entre les deux ensembles de variables.

c) Définition des variables canoniques

Chaque groupe est réduit à une seule variable qui prend la forme d'une combinaison linéaire. Ce sont les variables canoniques :

- $X1 = a1Gène1 + \dots + a10Gène10$
- $Y1 = b1AG1 + \dots + b11AG11$

Ces combinaisons linéaires sont choisies pour que la corrélation entre les variables canoniques soit maximisée.

L'ACC cherche ensuite le couple $X2$ et $Y2$, avec $X2$ une combinaison linéaire des Gènes non corrélée à $X1$ et $Y2$ une combinaison linéaire de AG non corrélée à $Y1$, telle que $X2$ et $Y2$ soient le plus corrélées possible. Et ainsi de suite... L'objectif est donc de trouver plusieurs paires de variables canoniques, où chaque paire est orthogonale (non corrélée) avec les autres, tout en maximisant la corrélation entre chaque couple de variables canoniques.

d) Calcul du coefficient de corrélation canonique

L'AC produit ainsi une suite de p couples de variables (X_s , Y_s) avec $1 < s < p$. Leurs corrélations successives sont appelées les coefficients de corrélation canonique et sont notées r_s . Les valeurs des corrélation canoniques sont décroissantes : $1 \geq r_1 \geq r_2 \geq \dots \geq 0$.

IV. Démonstration dans R

1. Importation du jeu de données

Reprenons l'exemple précédent concernant les souris. Nous avons des variables d'expression de gènes et des variables de proportions des lipides.

On importe le package CAA qui réalise les analyses canoniques des corrélations et le jeu de données concernant les souris.

```
library(CAA)
```

```
data(nutrimouse)
```

On extrait les matrices des variables gene et lipid pour réaliser les calculs

```
X=as.matrix(nutrimouse$gene)
```

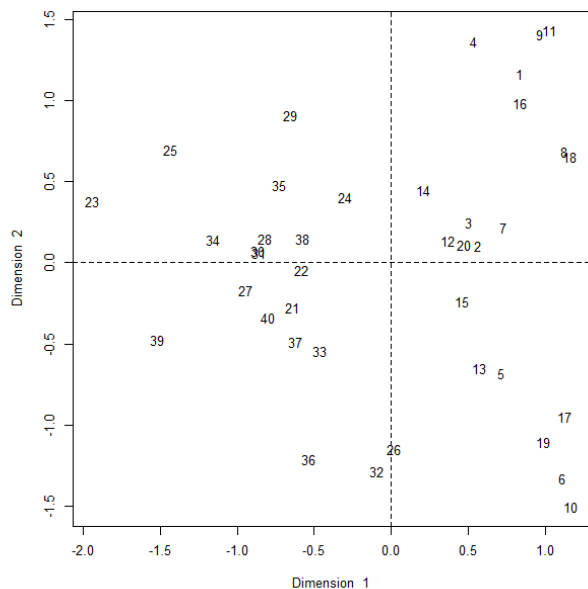
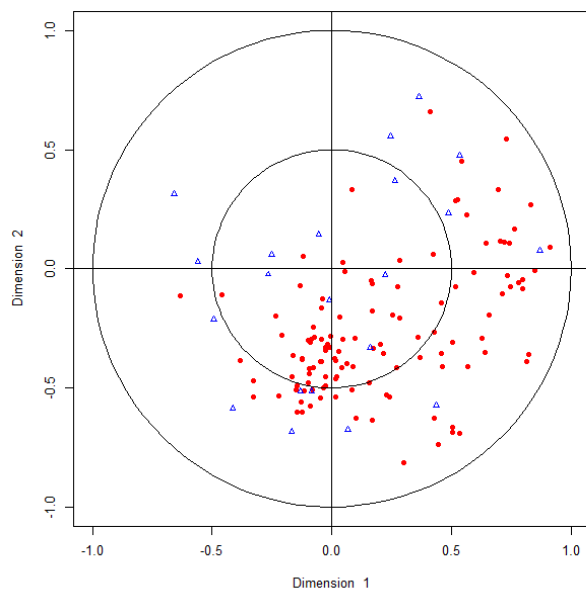
```
Y=as.matrix(nutrimouse$lipid)
```

On réalise l'analyse canonique des corrélations. C'est la fonction rcc qui réalise l'ACC. Il prend en entrée les deux matrices x et y et deux coefficients qui sont lambda 1 et lambda 2 (0.1 et 0.2). Ces paramètres permettent de réduire la complexité du modèle. C'est vraiment utile lorsqu'il y a beaucoup de variables par rapport au nombre d'individus.

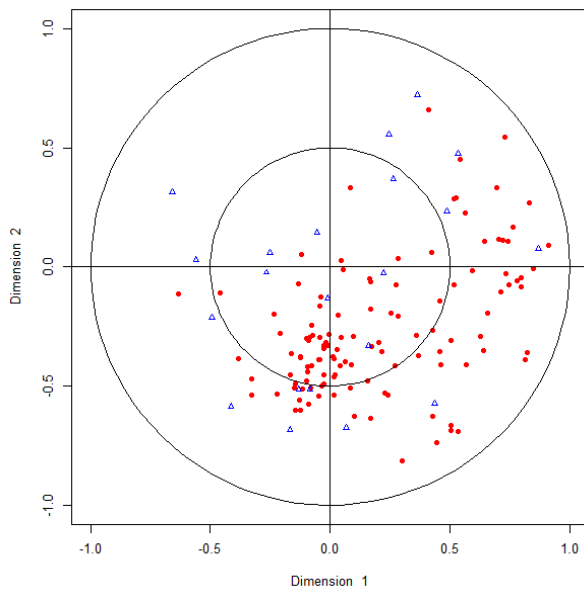
```
res.cc=rcc(X,Y,0.1,0.2)
```

On visualise les résultats sous forme de graphiques

```
plt.cc(res.cc)
```



2. Analyse des résultats



Premier graphique : Graphique de dispersion avec dimensions canoniques

Description du graphique :

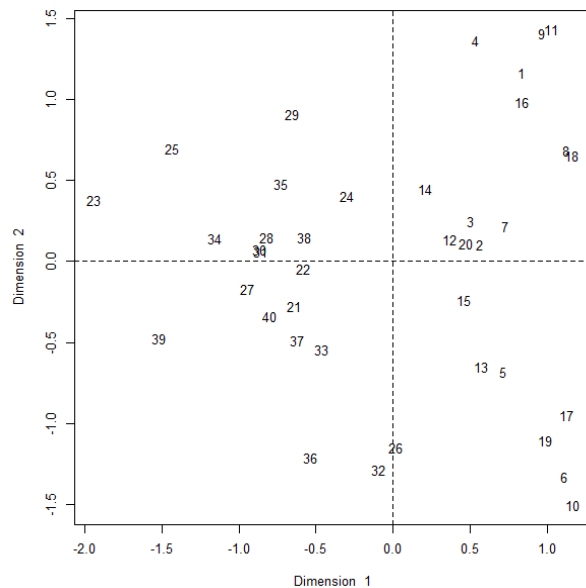
- **Les axes :** Dans ce graphique, les axes sont libellés Dimension 1 et Dimension 2, qui représentent les deux premières variables canoniques. La dimension 1 représente la corrélation linéaire maximale entre les deux groupes de variables. La dimension 2 est calculée de manière à être orthogonale à la dimension 1 tout en maximisant la corrélation entre ces 2 groupes.
- **Cercles concentriques :** Les cercles indiquent les niveaux de corrélation entre les variables canoniques et les variables originales dans chaque groupe. Les points plus proches du centre sont faiblement corrélés, tandis que les points situés sur les cercles extérieurs indiquent des corrélations plus fortes avec les dimensions canoniques.
- **Triangles bleus :** Ils représentent les variables du groupe X donc les gènes.
- **Ronds rouges :** Ils représentent les variables du groupe Y donc les acides gras.
- **La position de ces points** montre comment les variables des deux groupes se distribuent par rapport aux dimensions canoniques principales. Lorsque des points bleus et rouges sont proches les uns des autres, cela suggère que ces variables des deux groupes sont fortement associées dans la dimension canonique.

Interprétation :

- **Proximité aux cercles :** Les points situés plus près du cercle extérieur indiquent des variables qui sont fortement impliquées dans la corrélation canonique entre les deux groupes (gènes et lipides).

- **Placement des points** : Les points bleus et rouges proches montrent des variables qui contribuent ensemble à la corrélation sur ces dimensions canoniques. Par exemple, si plusieurs triangles bleus et ronds rouges sont dans le même quadrant, cela suggère qu'un certain sous-ensemble de gènes est associé de manière significative avec un sous-ensemble d'acides gras dans la dimension représentée.

Deuxième graphique : Graphique de dispersion des individus dans l'espace canonique



Description du graphique :

- **Les axes** : les axes sont également les deux premières dimensions canoniques (Dimensions 1 et 2), mais il se concentre sur la position des individus dans cet espace.
- **Points numérotés** : Chaque numéro représente un individu (par exemple, une souris) dans l'étude, positionné en fonction de ses scores pour les premières dimensions canoniques. La position de chaque point indique comment cet individu se situe par rapport aux dimensions canoniques, en fonction des combinaisons linéaires de variables génétiques et lipidiques.

Interprétation :

- **Position des individus** : Les individus proches les uns des autres partagent des profils similaires dans les dimensions canoniques. Par exemple, les individus 17, 19, 6 et 10 sont proches, donc ils ont des profils similaires de gènes et d'acides gras dans les combinaisons linéaires définies par les dimensions 1 et 2

Résumé de l'interprétation

Ensemble, ces graphiques permettent de comprendre :

1. Quels gènes et lipides sont associés dans l'espace canonique (premier graphique).

- Comment les individus (souris) se positionnent par rapport aux deux premières dimensions canoniques (second graphique), en identifiant des groupes d'individus avec des profils similaires de gènes et de lipides.

Ces visualisations fournissent donc des informations tant au niveau des variables (gènes et lipides) qu'au niveau des individus, ce qui permet de dégager des associations significatives dans vos données.

V. Comparaison avec d'autres méthodes

Analyse Canonique des Corrélations	analyse simultanée de deux groupes complets de variables maximise les corrélations entre groupes
Analyse en Composantes Principales	= construction et interprétation des graphiques ≠ focalise sur la structure interne d' <u>un seul groupe</u> de variables
Régression linéaire multiple	= nature des données ≠ étudie les relations entre une variable <u>dépendante</u> et plusieurs variables explicatives
Analyse Factorielle Discriminante	≠ maximise la séparation entre des groupes de <u>variables catégorielles</u>
Partial Least Squares	maximise la corrélation entre deux groupes de variables plus utilisée lorsque un des groupes est fortement influencé par un grand nombre de variables

VI. Limites de l'ACC

L'analyse canonique des corrélations est un outil puissant pour explorer les relations entre deux ensembles de variables. Cependant, cet outil a quelques limites :

- Les hypothèses de linéarité : les relations entre les groupes de variables doivent être linéaires
- Hypothèse de normalité : l'ACC repose sur l'hypothèse que les variables suivent une distribution normale. Si ce n'est pas le cas, les résultats peuvent être biaisés.
- La sensibilité aux valeurs extrêmes : l'ACC utilise la covariance qui est très sensible aux valeurs extrêmes. S'il y a une donnée aberrante, elle peut fortement fausser le résultat.
- Les difficultés d'interprétation : les coefficients de corrélation canonique peuvent être difficiles à interpréter. Cela ne représente pas la corrélation directe entre les variables, mais entre les combinaisons linéaires de ces variables.

Pour pallier certaines de ces limites, il peut être utile d'envisager des méthodes alternatives ou complémentaires, comme les modèles structurels, l'ACP ou l'utilisation de techniques de rééchantillonnage (bootstrap).

Conclusion

L'analyse canonique des corrélations (ACC) permet d'identifier des relations linéaires maximales entre deux groupes de variables quantitatives observés sur un même ensemble d'individus.

Utilisée en biologie, psychologie ou économie, elle révèle des associations complexes que ne capturent pas les corrélations simples.

Toutefois, l'ACC a des limites : elle suppose une relation linéaire et normale entre les variables et est sensible aux valeurs extrêmes, ce qui peut rendre l'interprétation difficile.

Aujourd'hui, l'ACC reste une méthode précieuse pour explorer des relations multidimensionnelles en complément d'autres approches analytiques.