

Analyse canonique des corrélations

Emma Da Costa Silva, Maud Lesage, Elise Lonchampt

Plan

Introduction

- I. Quand utiliser l'ACC ?
- II. Principe général de l'ACC
- III. Hypothèse et conditions d'utilisation
- IV. Explication de la méthode
- V. Interprétation des résultats
- VI. Comparaison avec d'autres méthodes
- VII. Limites de l'ACC

Conclusion

Introduction



ACC = méthode descriptive multidimensionnelle

→ Analogies avec :

- l'ACP : construction et interprétation des graphiques
- la régression linéaire : nature des données

Méthode mise au point par Harold Hotelling en 1936

Objectif : explorer les relations entre deux groupes de variables

Méthode peu utilisée mais connaît un essor avec le développement des biopuces

I. Quand utiliser l'Analyse Canonique des Corrélations ? (ACC)

- Etude des relations entre **deux groupes de variables quantitatives**
- Lorsque les deux groupes de variables peuvent **influencer de manière réciproque** les résultats
- Lorsque la **symétrie** entre deux groupes de variables est essentielle

Application dans les domaines : biologie, psychologie, économie...

II. Principe général de l'ACC

Réduction des groupes à une combinaison linéaire → **variables canoniques**

$$X = a_1 \text{Gène}_1 + \dots + a_{10} \text{Gène}_{10}$$

$$Y = b_1 \text{AG}_1 + \dots + b_{11} \text{AG}_{11}$$

L'algorithme ACC cherche une suite de p couples de variables canoniques avec p correspondant au nombre minimum de variables dans les deux groupes.

→ Chaque paire **maximise la corrélation** entre les deux groupes tout en étant **orthogonale** à la paire précédente.

Obtention de p **coefficients de corrélation canoniques** : $1 \geq r_1 \geq \dots \geq r_p \geq 0$

On ne prend pas en compte la direction → corrélation exprimée en valeur absolue

III. Hypothèses et conditions d'utilisation

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

IV. Explication de la méthode

Etude sur 40 souris → deux catégories de données

- Gènes : expression de 120 gènes
- AG : proportion de 21 AG hépatiques



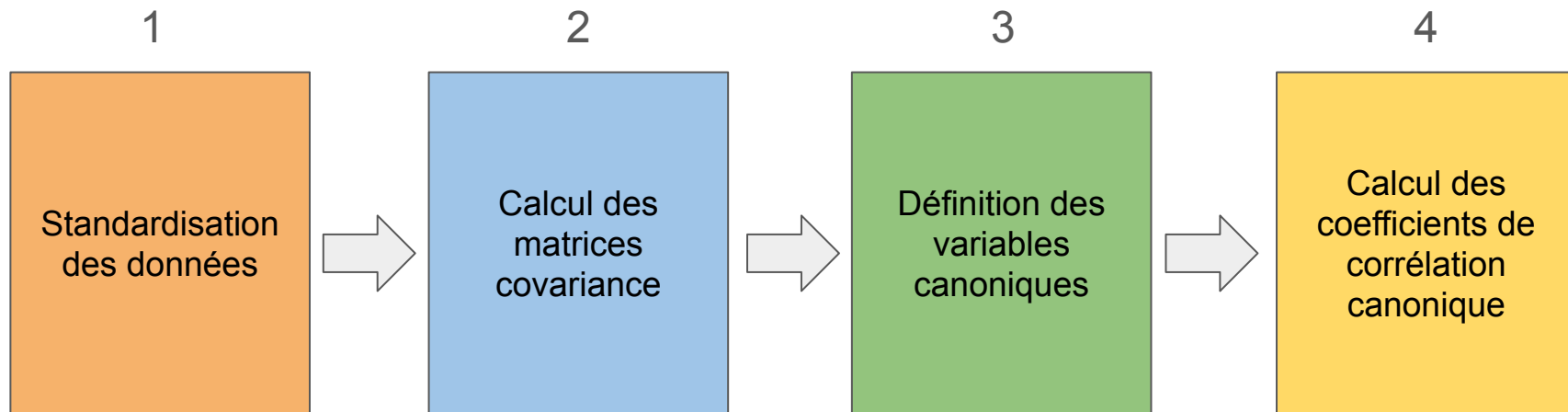
En ACC, il est nécessaire d'avoir un nombre de variables inférieur au nombre d'individus dans le groupe → sélection des gènes les plus importants

	Gène 1	...	Gène 10	AG 1	...	AG 11
1						
2						
...						
40						

X

Y

IV. Explication de la méthode



IV. Explication de la méthode

A. Calcul des matrices de covariance

- R_{XX} est la matrice de variance-covariance 10×10 des variables du groupe X,
- R_{YY} est la matrice de variance-covariance 11×11 des variables du groupe Y,
- R_{XY} est la matrice de covariance entre les variables de X et Y, c'est une matrice 10×11 ,
- R_{YX} est simplement la transposée de R_{XY} , c'est une matrice 11×10 .

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

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

IV. Explication de la méthode

B. Définition des variables canoniques

$$X = a_1 \text{Gène}_1 + \dots + a_{10} \text{Gène}_{10}$$

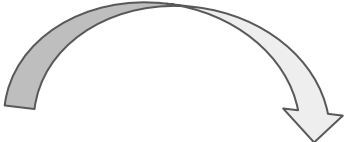
$$Y = b_1 \text{AG}_1 + \dots + b_{11} \text{AG}_{11}$$

Poids canoniques déterminés à partir des **vecteurs propres** des matrices R_X et R_Y **associés** **au valeurs propres** λ_s ($s = 1, \dots, 10$).

Plus une variable a un poids élevé, plus elle est fortement appliquée dans la corrélation entre les deux ensemble de variables.

IV. Explication de la méthode

C. Calcul du coefficient de corrélation



	Gène 1	...	Gène 10	AG 1	...	AG 11	X	Y
1								
2								
...								
40								

X
Y
Variables canoniques

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

Dans R

```
# on importe le package et le jeu de données
```

```
library(CAA)
```

```
data(nutrimouse)
```

```
# on extrait les matrices des variables gene et lipid
```

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

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

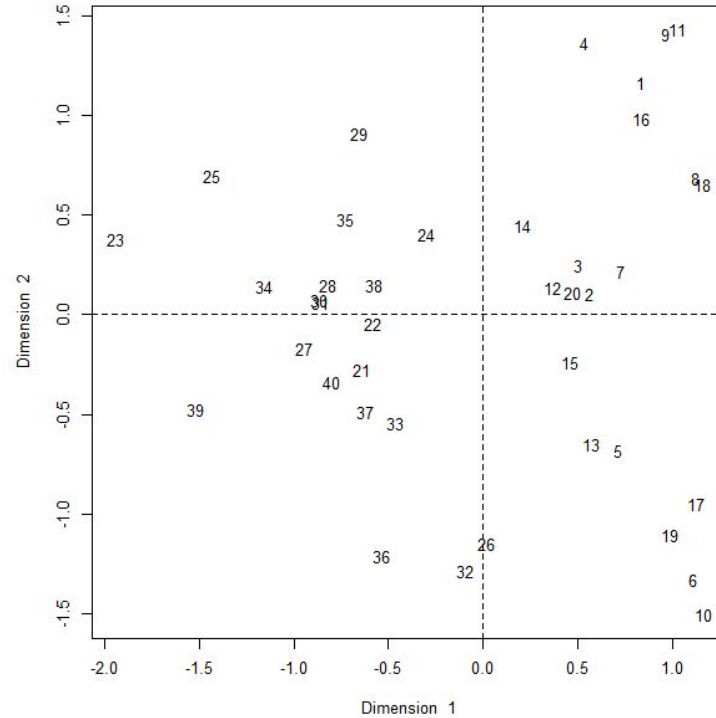
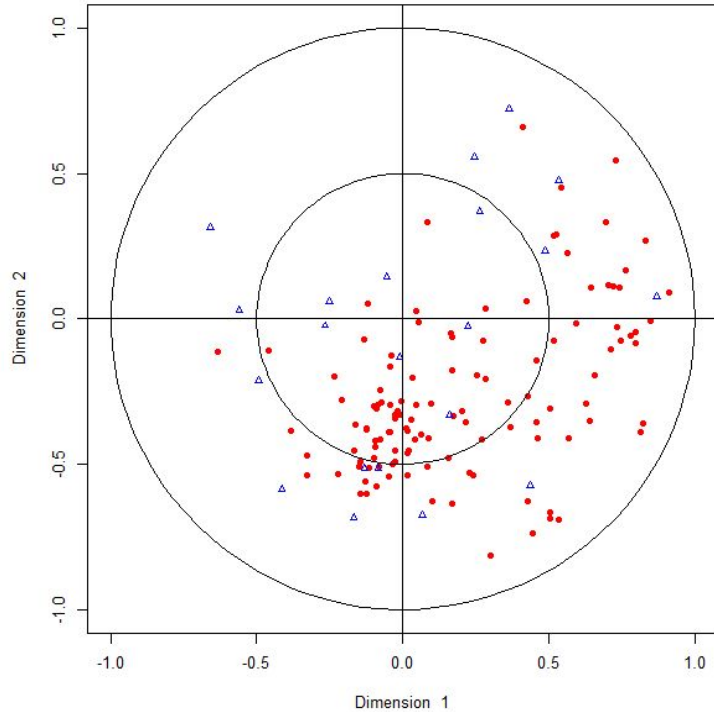
```
# on réalise l'analyse canonique des corrélations
```

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

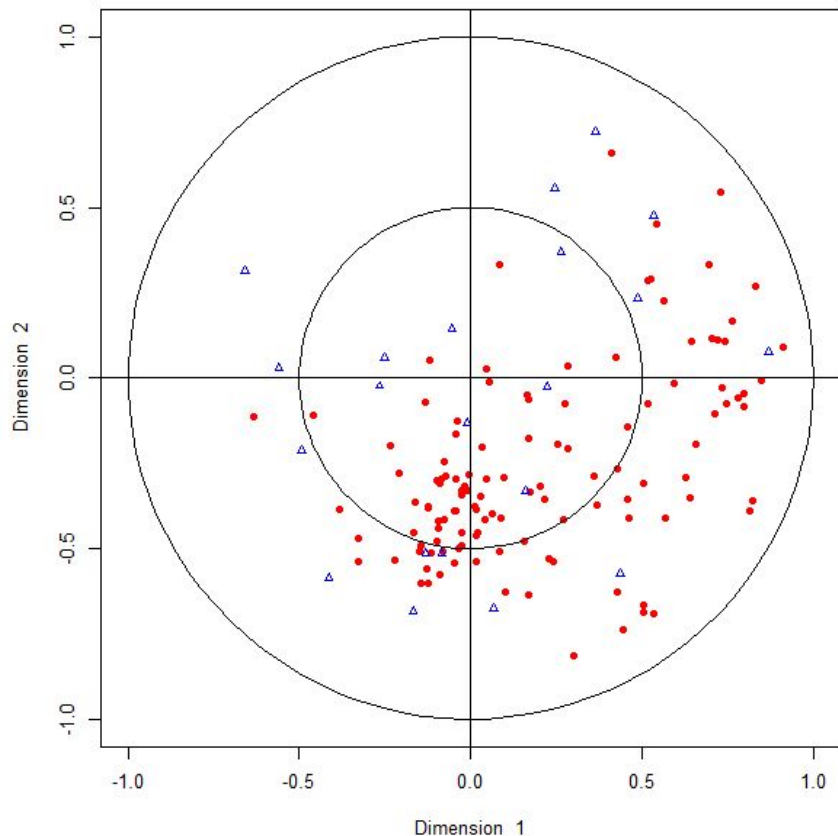
```
# on visualise les résultats sous forme de graphiques
```

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

V. Interprétation des résultats



V. Interprétation des résultats



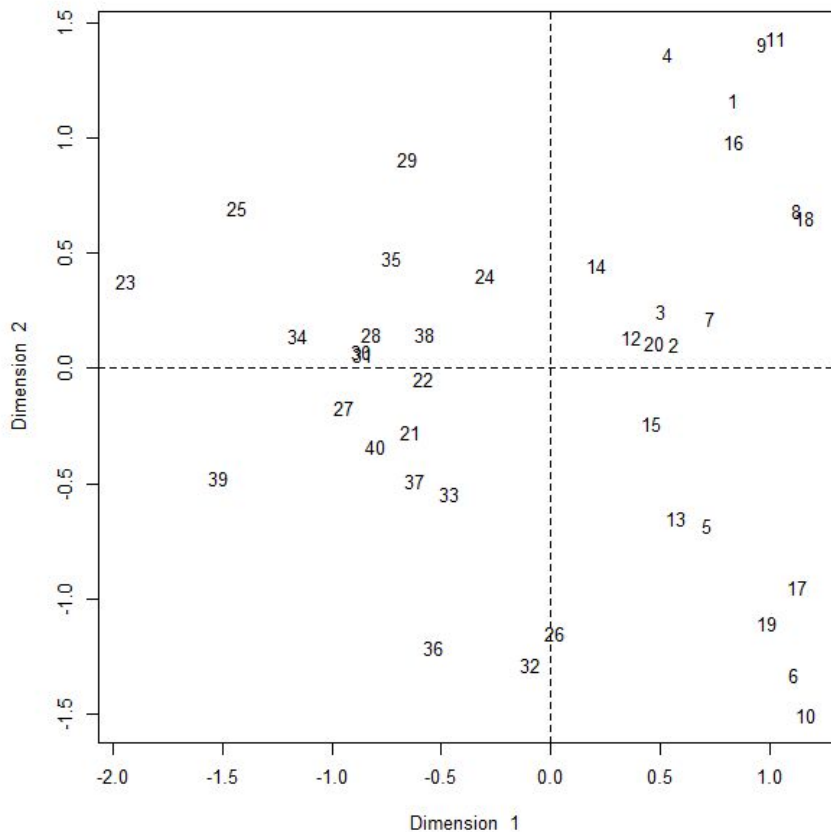
Axes : 2 premières variables canoniques

Triangles bleus : variables du groupe X (gènes)

Ronds rouges : variables du groupe Y (lipides)

Cercles : niveau de corrélation entre les variables canonique et les variables d'origines

V. Interprétation des résultats



Axes : 2 premières variables canoniques

Points numérotés : individus (souris)

VI. 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

VII. Limites de l'ACC

- Hypothèse de linéarité
- Hypothèse de normalité
- Sensibilité aux valeurs extrêmes
- Difficultés d'interprétation



Pour pallier ces limites, utiliser des méthodes alternatives ou complémentaires (ACP, techniques de rééchantillonnage,...)

Conclusion

- Relations linéaires maximales entre deux ensembles de variables quantitatives
- Utilisée dans de nombreux domaines (biologie, psychologie, économie, ...)
- Limites : hypothèse de linéarité, de normalité, sensible aux valeurs extrêmes et interprétation difficile

Bibliographie

Besse, P.-A. (n.d.). *Statistiques multivariées - Exploitation des données : Analyse Canonique des Corrélations (ACC)*. Université de Toulouse. <https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-acc.pdf>

SysBio_SU. (2023). *Introduction à l'Analyse Canonique des Correspondance* [Video]. YouTube. <https://www.youtube.com/watch?v=GolRbqBq04c>

TileStats. (2022). *Canonical correlation analysis - explained* [Vidéo]. YouTube. <https://www.youtube.com/watch?v=2tUuyWTtPqM>