# Cours analyse canonique des corrélations

## 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 connait un essor dans les années 1990 avec le développement de la régression PLS, et plus récemment avec l’apparition des données de biopuces.

## 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 : Y1, Y2, …, Yq
* Acides gras : X1, X2, …, Xp

On désigne Y la matrice de dimension n x q contenant les observations relatives au premier groupe (gènes) et X la matrice de dimension n x 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 peux 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 |  |  |  |  |  |  |

Y

X

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

* X1 = a1Gène1 + … + a10Gène10
* Y1 = b1AG1 + … + 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.

L’AC produit ainsi une suite de p couples de variables (Xs , Ys) avec 1 < s < p. Leurs corrélations successives sont appelées les **coefficients de corrélation canonique** et sont notées ρs. Les valeurs des corrélation canoniques sont décroissantes : 1 ≥ ρ1 ≥ ρ2 ≥ … ≥ 0.

## Démarche

Étapes pour déterminer la corrélation entre les deux groupes :

1. Calcul des matrice variance co-variance

Matrices de covariance intra-groupe + matrices intergroupes

On peut alors calculer les matrices Rv et Rw :

Où :

* RXX​ est la matrice de variance-covariance 10×10 des variables du groupe X (Gène1, …, Gène10),
* RYY est la matrice de variance-covariance 11×11 des variables du groupe Y (AG1, …, AG11​),
* RXY est la matrice de covariance entre les variables de X et Y, c'est une matrice 10×11 (entre X1, X2, …, X10​ et Y1, Y2, …, Y11​),
* RYX​ est simplement la transposée de RXY, c’est une matrice 11×10.

1. Combinaison linéaire

Il est alors possible de calculer les valeurs propres des deux matrices Rx et RY. 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.

1. Variables canoniques
2. Calcul du coefficient de corrélation