# 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).

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. Soient 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.

L’ACC cherche à identifier les **combinaisons linéaires** :

* V1 = a1X1 + … + apXp
* W1 = b1Y1 + … + bqYq

V1 et W1 sont des **variables canoniques**. Ces combinaisons linéaires sont choisies pour que la **corrélation** entre les variables canoniques soit **maximisée**.

L’ACC cherche ensuite le couple V2 et W2, avec V2 une combinaison linéaire des X non corrélée à V1 et W2 une combinaison linéaire de Y non corrélée à W1, telle que V2 et W2 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 (Vs , Ws) 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.

## Approche mathématique