Objectifs de la statistique multivariée

- Observer simultanément des individus d'une population sur p caractères

$$\begin{array}{lll} \boldsymbol{P} & \rightarrow & \boldsymbol{M} = \boldsymbol{M}_1 \times \boldsymbol{M}_2 \times \ldots \times \boldsymbol{M}_p \\ \omega & & \boldsymbol{c}(\omega) = (\boldsymbol{c}_1(\omega), \boldsymbol{c}_2(\omega), \ldots, \boldsymbol{c}_p(\omega)) \end{array} \quad \text{où } \boldsymbol{M}_j = \boldsymbol{R} \text{ ou } \boldsymbol{M}_j = \boldsymbol{N} \text{ pour } 1 \leq j \leq p$$

- Analyse par individus
  - Représenter ces individus dans des espaces de faibles dimensions représentant au mieux leurs différences.
  - Qualifier les spécificités de chaque individu en fonction des caractères.

- Analyser par caractères
  - Calculer les liens par couple de caractères
  - Chercher des caractères synthétiques signifiants comme combinaison des caractères initiaux

- Ces deux analyses sont duales l'une de l'autre

Analyse en composantes principales

- Position du problème : On désire révéler les différences les plus significatives entre les individus observés à travers p caractères quantitatifs.

$$\begin{array}{ccc} P & \rightarrow & R^p \\ \omega & & c(\omega) = (c_1(\omega), c_2(\omega), \ldots, c_p(\omega)) \end{array}$$

- On supposera que tous les individus ont le même poids.
- Centrage des données : On calcule le vecteur $i_m = (\bar{c}_1, \ldots, \bar{c}_p)$ qui représente l'individu moyen. On fait une translation de ces données du vecteur i_m. Cette translation ne change pas la nature du problème car une translation conserve les distances.
- Réduction des données : Les caractères ne sont pas exprimés dans la même unité. Calculer les différences entre les individus en utilisant la distance canonique n'a donc pas de sens. Pour chaque caractère c_i, on divisera les données par $\sigma(c_i)$.
- En résumé les données traitées seront :

$$\begin{array}{ccc} P & \rightarrow & R^p \\ \omega & & c(\omega) = ((c_1(\omega) - \bar{c}_1)/\sigma(c_1), (c_2(\omega) - \bar{c}_2)/\sigma(c_2), \ldots, (c_p(\omega) - \bar{c}_p)/\sigma(c_p)) \end{array}$$

ACP : Inerties

- On suppose que R^p est muni du produit scalaire définie par la matrice Id et de la distance d associée. Pour simplifier, on notera x_i les vecteurs $c(\omega_i)$

- On définit l'inertie du nuage de points par rapport à l'origine.

$$I_N(0) = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, 0)$$

  Ce nombre représente la dispersion du nuage par rapport à l'origine. C'est une variance généralisée dans R^p.

- On cherche des droites vectorielles Δ_u dans R^p telles que la dispersion du nuage projeté sur ces droites soit maximum.

$$I_N(\Delta_u^{\perp}) = \frac{1}{n}\sum_{i=1}^{n} d^2(\mathrm{Pr}_{\Delta_u}(x_i), 0)$$

- Remarque : $I_N(0) = I_N(\Delta_u^{\perp}) + I_N(\Delta_u)$

- On note M la matrice (p,p) telle que $m_{j_1, j_2} = r(c_{j_1}, c_{j_2})$
- On peut montrer que si u est un vecteur unitaire alors

$$I_N(\Delta_u^{\perp}) = ({}^t U.M.U)$$

ACP : Axes principaux (1)

- On doit résoudre le problème suivant :

$$\text{Max } {}^{t}\boldsymbol{U}.\boldsymbol{M}.\boldsymbol{U} \text{ sous la contrainte } \|\boldsymbol{U}\| = 1$$

- On peut démontrer en utilisant la méthode des multiplicateurs de Lagrange et le fait que M est une matrice définie positive qu'une solution du problème est u_1 où u_1 est un vecteur propre unitaire de M associée à la plus grande valeur λ_1.
- Remarque : $\boldsymbol{I}_N(\Delta_U^{\perp}) = {}^{t}\boldsymbol{U}_1.\boldsymbol{M}.\boldsymbol{U}_1 = {}^{t}\boldsymbol{U}_1.(\lambda_1\boldsymbol{U}_1) = \lambda_1.{}^{t}\boldsymbol{U}_1.\boldsymbol{U}_1 = \lambda_1$
- Le vecteur u_1 est le premier axe principal.

ACP : Axes principaux (2)

- Pour chercher le deuxième axe principal, nous devons travailler dans l'espace $\Delta_u^{\perp}$. On doit donc résoudre le problème suivant :

$$\text{Max } {}^t\boldsymbol{U}.\boldsymbol{M}.\boldsymbol{U} \text{ sous les contraintes } \|\boldsymbol{U}\| = 1, {}^t\boldsymbol{U}.\boldsymbol{U}_1 = 0$$

- On peut démontrer en utilisant encore la méthode des multiplicateurs de Lagrange et le fait que M est une matrice définie positive qu'une solution du problème est u_2 où u_2 est un vecteur propre unitaire de M associée à la deuxième plus grande valeur λ_2.
- Attention : Si λ_1 a un ordre de multiplicité supérieur ou égal à 2 alors $\lambda_2 = \lambda_1$
- D'un point de vue général, le $j^{ème}$ axe principal est un vecteur propre unitaire de M associée à la $j^{ème}$ grande valeur propre de M.
- On reconstitue la dispersion avec les différents axes : $\boldsymbol{I}_N(0) = \sum_{j=1}^{p} \boldsymbol{I}_N(\Delta_{u_j}^{\perp})$

ACP : Composantes principales

- On visualise les individus sur les axes principaux . La projection z_i^l du i^{ème} individu sur le le l^{ème} axe vaut $z_i^l = {}^t U_l . x_i$
- On définit dans R^n, les vecteurs ζ^l dont la i^{ème} coordonnée est . Ces vecteurs représentent de nouveaux caractères qui sont en quelques sortes des synthèses des caractères initiaux c_i. Ces vecteurs ζ^l sont appelés composantes principales.
- les vecteurs ζ^l en tant que caractères statistiques vérifient $var(\zeta^l) = \lambda_l$.
- Pour interpréter ces caractères synthétiques, il faut calculer les coefficients de corrélation $r(c_j, \zeta^l)$. On peut démontrer l'égalité suivante :

$$r(c_j, \zeta^l) = \sqrt{\lambda_l} \frac{u_l^j}{\sqrt{m_{j,j}}}$$

- Remarque : comme on a réduit les données, $m_{j,j} = 1$.

Algorithme de calcul des u_j et λ_j

- Pour diagonaliser une matrice M symétrique définie positive, on peut utiliser l'algorithme ci-dessous :

$$\boldsymbol{v}_0 \text{ un vecteur unitaire}$$

$$v_n = \boldsymbol{M}.v_{n-1} \Big/ \left\| M.v_{n-1} \right\|$$

- La suite v_n converge vers un vecteur propre unitaire associée à la plus grande valeur propre de M. On note u_1 la limite de cette suite et λ_1 la valeur propre associée.

- On note $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ les différentes valeurs propres de M.

- La matrice M' définie par $M' = M - \lambda_1 u_1 .{}^t u_1$ est une matrice symétrique définie positive qui a les mêmes vecteurs propres que M et dont les valeurs propres sont $(\lambda_2, \ldots, \lambda_p, 0)$.