TD 3 : Analyse multivariée

Objectifs :
- représentation et interprétation d'un volume de données important dans un espace de dimension réduite,
- détermination des relations entre variables et individus,
- classification des données.

1 Analyse de données : réseau métabolique du foie

On considère les données contenues dans le fichier `mitochondrie.csv` récoltées expérimentalement puis organisées sous forme de matrice :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{1}$$

Ainsi chaque colonne numérotée de 1 à p correspond à une variable v_i. Chaque variable correspond à une réaction donnée modélisée par un sommet dans un graphe. Chaque ligne allant de 1 à n est une mesure et correspond au chemin optimal (minimal et unique) dans le graphe correspondant.

Exercice 1 : Récupérez et chargez le fichier fichier `mitochondrie.csv` contenu dans l'archive `http://www.labri.fr/~fourer/Ens/1112/stats/data2.tar.gz` (c.f. TD 2). Vous pourrez utiliser la fonction `read.table()` en vous assurant que vous récupérez une variable X de type `matrix`.

1.1 Distance entre variables

Soit, a et b les échantillons de taille N correspondant à 2 variables.

distance Euclidienne $[0, +\infty[$	$\sqrt{\sum_{i=1}^{N}(b_i - a_i)^2}$
covariance $]-\infty, +\infty[$	$d(a,b) = \frac{1}{n-1}\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})$
corrélation $[-1, +1]$	$d(a,b) = \frac{\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\|\|a\|\| \cdot \|\|b\|\|}$

Remarque : la corrélation correspond à la covariance centrée et réduite.

La matrice de distance entre p variables est une matrice symétrique (car $d(a,b) = d(b,a)$) définie par :

$$D = \begin{pmatrix} d(v_1, v_1) & \cdots & d(v_1, v_p) \\ \vdots & \ddots & \vdots \\ d(v_p, v_1) & \cdots & d(v_p, v_p) \end{pmatrix} \tag{2}$$

Exercice 2 : Choisissez une fonction de distance parmi celles proposées ci-dessus puis construisez la matrice D correspondante entre les variables (colonnes de la matrice X). Pour la covariance (resp. corrélation), vous prendrez la valeur absolue des valeurs calculées. Vous pourrez utiliser au choix les fonctions `dist()`, `cov()` ou `cor()`.

Exercice 3 : A l'aide de la fonction `hclust()`, appliquez une classification hiérarchique des variables sur la matrice de distance calculée précédemment. Vous pourrez utiliser la fonction `as.dist()` pour créer une matrice de distance compatible avec `hclust`. Les matrices de distance peuvent être converties en matrice numérique classique en utilisant la fonction `as.matrix()`.

Exercice 4 : Visualisez le résultat de la classification effectuée précédemment à l'aide de la fonction `plot()`. Déduisez quelles sont les variables les plus proches pour la distance que vous aurez choisi. Attention, pour la covariance (resp. corrélation), la fonction de distance mesure l'indépendance statistique entre 2 variables. Ainsi, une covariance de valeur absolue plus importance signifie que les variables sont dépendantes (et donc proches) d'un point de vue statistique.

Exercice 5 : A votre avis, quelle est la distance la mieux adaptée pour mettre en évidence une synchronicité entre les variables ? une relation linéaire ?

2 Régression linéaire d'ordre 1

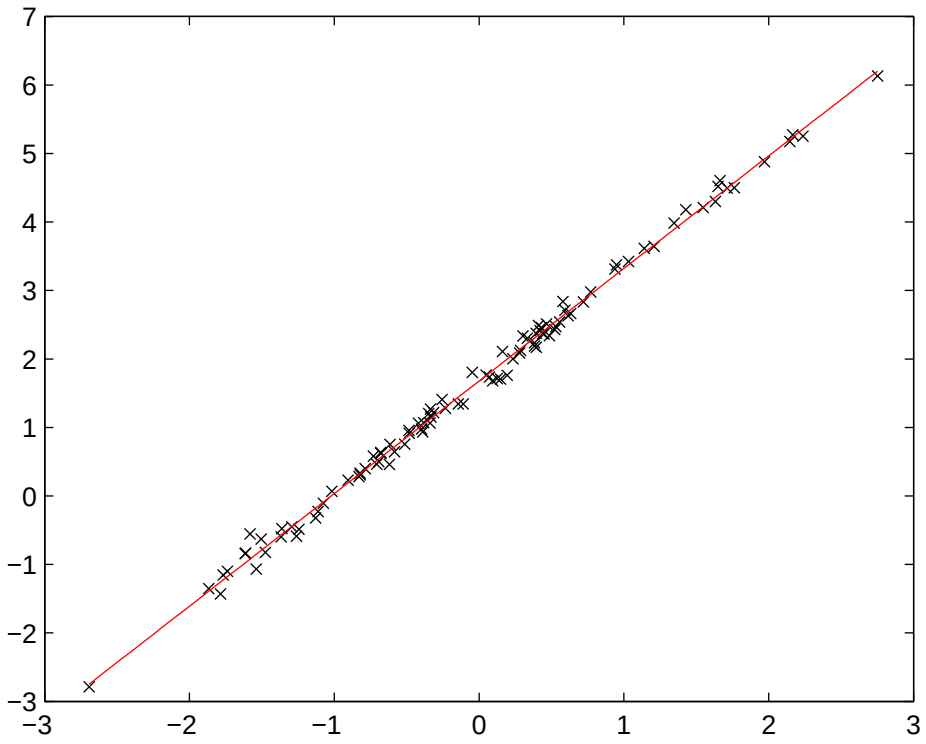

FIG. 1 – Regression linéaire

Une régression linéaire d'ordre 1 modélise une fonction affine entre 2 variables x et y de taille N tel que :

$$y_i = \alpha \cdot x_i + \beta + \epsilon_i \text{ pour } i \in [1, N] \tag{3}$$

Ainsi, on recherche les paramètres α et β de la droite qui minimise $||\epsilon||^2$ pour tous les couples (x_i, y_i). On donne sans démonstration, un estimateur :

$$\hat{\alpha} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4}$$

où $\bar{x}$ est la moyenne du vecteur x. L'ordonnée à l'origine est donnée par

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x} \tag{5}$$

Exercice 6 : Appliquez une régression linéaire entre les colonnes de la matrice X les plus corrélées . Pour cela vous pourrez déclarer une fonction `regression1d(x,y)`. Pour chaque droite estimée, calculez $\bar{\epsilon}$, c'est à dire l'erreur moyenne des différences entre y_i et l'ordonnées de la droite correspondant à x_i.

Exercice 7 : Affichez avec `plot` vos données ainsi que la droite de régression que vous aurez calculé de manière à obtenir une figure similaire à la figure 1. Conclusions ?

3 Analyse en composantes principales

A présent on souhaite trouver une représentation adéquate pour les variables de la matrice X. L'analyse en composante principale (ACP) propose une solution qui consiste à trouver un sous-espace mieux adapté qui préserve les distances.

L'ACP peut se résumer de la manière suivante :

1. Calcul de la matrice de distance D :
2. diagonalisation de la matrice $D = U\Lambda U^T$ telle que Λ est la matrice diagonale des valeurs propres. (on pourra utiliser la fonction `eigen()`)
3. tri des valeurs propres par ordre décroissant
4. projection des unités (lignes de X) sur la base des vecteurs propres. (on choisit en général 2 ou 3 vecteurs propres correspondant aux valeurs propres les plus importantes)

Exercice 8 : Appliquez une ACP à la matrice X en utilisant une matrice D de covariance. Affichez les unités calculées pour le premier plan factoriel (défini par les 2 premiers vecteurs propres).

Exercice 9 : Calculez l'indice de qualité global définit par :

$$\text{IQG} = \frac{\sum_i^k \lambda_i}{\sum_j \lambda_j} \tag{6}$$

qui correspond la pertinence (comprise entre 0 et 1) de la représentation des données sur les k premiers vecteurs propres.
Vous pourrez visualiser graphiquement les valeurs propres avec la fonction `barplot()`.

Exercice 10 : Tracez le cercle de corrélation du premier plan factoriel. Chaque coordonnée est définie par la corrélation entre chaque variable et les variables projetées sur les axes factoriels.