Université de Lorraine - Site de Metz - UFR MIM **Année 2012/2013**
Master de Mathématiques, année M1
Module *Introduction à l'analyse des données*

Exercices - Feuille 2

Analyse élémentaire des données multivariées

1- Matrice des covariances, matrice des corrélations
Soit $X \in \mathbb{M}_{n,p}(\mathbb{R})$, une matrice de données correspondant à un n échantillon de p variables.
1) Rappeler la définition de la moyenne $m \in \mathbb{R}^p$ et de la matrice des covariances $C \in \mathbb{M}at_p(\mathbb{R})$ de X.
2) Montrer que C peut s'écrire matriciellement

$$C = \frac{1}{n-1}\left(X^T X - \frac{1}{n} X^T \mathbb{1}_n \mathbb{1}_n^T X\right) \tag{1}$$

et également

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)(x_i - m)^T \tag{2}$$

3) Rappeler la définition de la matrice des corrélations. On note D la matrice

$$D = \text{diag}\ (s_1, s_2, \cdots, s_p) \tag{3}$$

où s_j^2 est la variance empirique des données $x_{i,j}, 1 \leq i \leq n$. Montrer que la matrice des corrélations est $R \in \mathbb{M}at_p(\mathbb{R})$ donnée par

$$R = D^{-1} C D^{-1} \tag{4}$$

2- Calcul de la matrice des covariances
On considère la matrice $X \in \mathbb{M}at_{5,3}(\mathbb{R})$ des données suivantes, avec $n = 5$ observations et $p = 3$ caractères observés,

$$X = \sqrt{10}\begin{bmatrix} 2 & 2 & 3 \\ 3 & 1 & 2 \\ 1 & 0 & 3 \\ 2 & 1 & 4 \\ 2 & 1 & 3 \end{bmatrix} \tag{5}$$

1) Calculer la moyenne empirique et les variances empiriques de chacun des caractères X^1, X^2, X^3.
2) Calculer les covariances de deux caractères X^k, X^l pour $k \neq l$.
3) Calculer la matrice C des covariances.
4) Calculer la matrice R des corrélations.
5) Donner une interprétation des résultats.

3- Matrice des covariances, matrice des corrélations (3)
On considère les tableaux de données

- `pullover.dat`
  Données des ventes de pullovers bleus.
- `mandible.dat`
  Données de paléontologie sur l'origine du chien préhistorique.
- `frenchfood.dat`
  Données sur l'alimentation en France par catégories socioprofessionnelles.

Pour chacune de ces données, effectuer les calculs suivants:
1) Calcul du vecteur moyen.
2) Calcul de la matrice des covariances.
3) Calcul de la matrice des corrélations.
On pourra utiliser les logiciels `R` ou `matlab`. En `matlab`, le tableau des données se lit à l'aide de la commande `X=tblread('pullover.dat');`. Pour le calcul du vecteur moyen m, des matrices C des covariances et R des corrélations, utiliser les commandes `mean`, `cov` et `corrcoef`.

4- Coefficient de corrélation de Spearman
Il arrive que pour comparer deux variables aléatoires X et Y on utilise le coefficient de corrélation de Spearman basé sur le **classement** des valeurs x_i et y_i. Ceci arrive quand par exemple les valeurs x_i et y_i n'ont pas de sens numérique réel. (Par exemple une note de 12/20 ne signifie pas le double d'une note de 6/20). On classe les valeurs x_i et y_i par ordre croissant et on note $m_i \in \{1, \cdots, n\}$, le rang de la valeur x_i et $p_i \in \{1, \cdots, n\}$, le rang de la valeur y_i. Les entiers m_i d'une part et p_i d'autre part définissent donc une permutation de l'ensemble $\{1, \cdots, n\}$.

objet	1	2	$\cdots$	n
rang no 1	m_1	m_2	$\cdots$	m_n
rang no 2	p_1	p_2	$\cdots$	p_n

Le coefficient de corrélation de Spearman est défini par

$$r_s = \frac{\mathrm{cov}(m,p)}{s_m s_p} \tag{6}$$

où s_m, s_p désignent les covariances empiriques des deux vecteurs m et p.
1) On suppose qu'il n'y a pas d'ex-æquos. Montrer que les moyennes empiriques $\bar{m}, \bar{p}$ sont

$$\bar{m} = \bar{p} = \frac{n+1}{2} \tag{7}$$

2) Montrer que

$$\sum_{i=1}^{n} m_i^2 = \sum_{i=1}^{n} p_i^2 = \frac{n(n+1)(2n+1)}{6} \tag{8}$$

En déduire que les variances empiriques sont s_m^2, s_p^2 sont

$$s^2(m) = s^2(p) = \frac{n(n+1)}{12} \tag{9}$$

3) On note $d_i = m_i - p_i$ l'écart entre les rangs m_i et p_i. Verifier que $\mathrm{cov}(m,p)$ est

$$\mathrm{cov}(m,p) = \frac{1}{n-1}\left(-\frac{1}{2}\sum_{i=1}^{n} d_i^2 + \frac{n(n+1)(2n+1)}{6} - n\left(\frac{n+1}{2}\right)^2\right) \tag{10}$$

4) En déduire que le coefficient r_s peut se calculer par

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{11}$$

5) On considère le tableau de données suivant (comparaison de 10 vins par deux œnologues)

x_i	1	2	3	4	5	6	7	8	9	10
y_i	3	1	4	2	6	5	9	8	10	7

En utilisant la formule (11), calculer le coefficient de Spearman de ces données.

5- Comparaison chiffre d'affaire/employés

On souhaite comparer le chiffre d'affaire et le nombre d'employés dans dix grandes entreprises, numerotées de 1 a 10. On obtient le tableau suivant (en milliards d'euros et en milliers d'employés).

entreprise	CA (y)	Employés (x)
1	103.54	311.0
2	88.76	373.0
3	88.12	242.4
4	72.37	125.2
5	65.50	135.1
6	52.17	161.6
7	49.40	106.6
8	46.14	115.8
9	44.58	142.9
10	41.93	83.8

1) Représenter le scatterplot x/y de ces données et en donner l'interprétation.
2) Calculer les coefficients de Pearson et de Spearman.
3) Comment sont modifiés les coefficients quand on effectue le calcul avec le CA en euros et avec le nombre d'employes ?