INSA - Méthodes Statistiques M8 - Examen Médian 2010 - Formulaire autorisé - **1 h 30**

Exercice 1 — pour faire court — 5 points

1. Parmi ces résultats vus en cours lequel vous semble le plus important et pourquoi ?
   1. $\mathbf{a} = (X^\top X)^{-1} X^\top \mathbf{y}$
   2. $\widehat{\sigma}_x^2 = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$
   3. quand deux variable aléatoires x et y sont indépendantes on a $\mathbb{P}(x, y) = \mathbb{P}(x)\mathbb{P}(y)$
   4. $\min_{\mathbf{u},\mathbf{v}} \|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \quad \Leftrightarrow \quad X^\top X \mathbf{v} = \lambda \mathbf{v}$
   5. $c_i = \dfrac{H_{ii}}{p(1 - H_{ii})^2} \dfrac{\widehat{\varepsilon}_i^2}{s^2}$

   Attention : il ne faut donner qu'une et une seule réponse

Exercice 2 — Croissance et Régression — 5 points

D'après Monsieur Marcotte, professeur au laboratoire de Géophysique et Géostatistique de l'École Polytechnique de Montréal, il y a dans Excel une fonction « Growth » (ou croissance en français) qui permet d'estimer les paramètres c et d du modèle :

$$z = c\, d^x$$

1. indiquez comment « linéariser » ce modèle et ainsi obtenir des estimés pour c et d avec un programme de régression linéaire. Indiquez clairement le vecteur y et la matrice X qui seront soumis au programme de régression, les coefficients obtenus par la régression et le lien avec les coefficients recherchés.
2. les prédictions obtenues avec ce modèle minimiseront-elles la somme des carrés des erreurs $\sum_{i=1}^{n} (y_i - z_i)^2$. Justifiez votre réponse.

Exercice 3 — Lasse épée — 10 points

Au verso vous trouverez la copie d'une session Matlab. On y trouve des lignes de code qui commencent par le chargement d'un tableau de données X comportant $p = 3$ variables et $n = 15$ observations. Rappel : la fonction matlab `[U,d] = eig(M)` calcule les vecteurs propres (U) et les valeurs propres associées (sur la diagonale de d) de la matrice M.

1. dessinez la fonction de répartition empirique de la première variable,
2. donnez la boite à moustache correspondant à la première variable,
3. représentez graphiquement la médiane de Tukey et le sac médian, des deux premières variables
4. y a-t'il un point aberrant ?
5. quelle est la moyenne et la variance de Xn ?
6. quelle est la première composante principale et quelle est sa relation avec les trois variables originales ?
7. représentez le nuage de points dans les deux premiers axes de l'analyse en composantes principales (ACP).
8. comment caractériser la qualité de la représentation du nuage de points dans les deux premiers axes de l'ACP ?

```
load X
X = 8.5635   -0.4249   -1.1302
   10.6189   -0.6752    0.5311
   10.3912   -0.5847    0.2538
   10.4251   -0.6200    0.7893
   11.1107   -0.5394    0.6410
    9.1614   -0.5497    0.1627
   10.5240   -0.4315    0.2750
   10.8754   -0.4242    0.4208
    8.4377   -0.3782   -1.3738
    9.0995   -0.3888   -1.2308
   10.6006   -0.5199    0.9124
   11.1059   -0.5205    0.8106
    8.7338   -0.3190   -1.1281
    9.8828   -0.5151    0.2720
    8.3042   -0.3916   -1.5087

mean(X) = 9.8557   -0.4855   -0.0868

cov(X) = 1.0553   -0.0647    0.8529
        -0.0647    0.0101   -0.0692
         0.8529   -0.0692    0.8099

Xn = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X));
cov(Xn) =  1.0000   -0.6251    0.9226
          -0.6251    1.0000   -0.7635
           0.9226   -0.7635    1.0000

[U,d]=eig(X'*X)
U =    0.0493   -0.0010   -0.9988
       0.9982   -0.0334    0.0493
       0.0334    0.9994    0.0006
diag(d) =   0.1      11.5     1475.4

[Un,dn]=eig(Xn'*Xn)
Un = -0.6025    0.5467    0.5814
      0.2211    0.8143   -0.5366
      0.7669    0.1948    0.6115
diag(dn) =   0.7704     5.5677     35.6618

V = X*U
V = -0.0395   -1.1243   -8.5747
    -0.1326    0.5423  -10.6390
    -0.0627    0.2623  -10.4073
    -0.0784    0.7987  -10.4425
     0.0309    0.6470  -11.1234
    -0.0914    0.1713   -9.1772
     0.0974    0.2782  -10.5323
     0.1270    0.4234  -10.8828
    -0.0072   -1.3692   -8.4469
     0.0195   -1.2267   -9.1084
     0.0343    0.9182  -10.6128
     0.0552    0.8159  -11.1176
     0.0746   -1.1259   -8.7396
    -0.0177    0.2787   -9.8961
    -0.0318   -1.5034   -8.3143

Vn = Xn*Un
Vn =
     0.0019   -0.4233   -1.7634
    -0.3375   -0.9941    1.8629
    -0.2415   -0.4430    1.0630
     0.1174   -0.5950    1.6345
    -0.2342    0.3898    1.4921
     0.4790   -0.8342    0.1184
     0.0349    0.8706    0.3364
    -0.0308    1.1484    0.5954
    -0.0293   -0.1650   -2.2492
    -0.3191    0.1317   -1.7205
     0.3391    0.3347    1.2839
    -0.0453    0.5768    1.5038
     0.1362    0.5239   -2.2297
     0.2248   -0.1473    0.4171
    -0.0955   -0.3742   -2.3447
```

Formulaire

- moyenne : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$
- variance : $\widehat{\sigma}_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- médiane : $\mathbb{P}(X < M) = 0,5$
- fonction de répartition empirique : $\widehat{F}(x) = \dfrac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x<x_i\}}$
- covariance : $c_{xy} = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
- correlation : $\mathrm{cor}(x,y) = \dfrac{c_{xy}}{\sqrt{\widehat{\sigma}_x^2\widehat{\sigma}_y^2}}$
- probabilité conditionnelle : $\mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(X=x_i, Y=y_j)}{\mathbb{P}(Y=y_j)}$
- moyenne conditionnelle : $\mathbb{E}[Y|X = a] = \sum_{i=1}^{n} y_i \mathbb{P}(Y = y_i | X = a)$
- estimateur des moindres carrés : $\widehat{\alpha} = (X^\top X)^{-1} X^\top y$
- $\log ab = \log a + \log b$