HYDROLOGIE STATISTIQUE

Enoncé et Corrigé du Contrôle du 21 Janvier 2000
Promos 3HSEE et DEA-STE 1999/2000
Enseignant : R. Ababou

Exercice II

II. ANALYSE STATISTIQUE BIVARIEE : CORRELATION, REGRESSION, ET ANALYSE EN COMPOSANTES PRINCIPALES OU "A.C.P" *(simple test)*

ENONCE DU II.

Pour tester un programme d'analyse statistique multivariée débouchant sur de l'A.C.P, on étudie la structure de corrélation de 2 vecteurs d'observations (x1,x2) représentant 2 variables distinctes (non précisées). Les résultats de cette analyse sont joints à ce document, et sont décrits ci-dessous.[1]

Premièrement, on donne **ci-joint** les valeurs numériques des Matrices représentant les Covariances et les Composantes Principales, pour N=1000 paires d'observations. On remarque que les moyennes de (x1,x2) ne sont pas nulles, mais quand même relativement proches de (0,0).

Deuxièmement, une visualisation graphique dans le plan (x1,x2), **figure ci-jointe**, illustre les résultats obtenus pour un sous-ensemble des 1000 paires de points. **Noter que les axes (x1,x2) ne sont pas représentés à la même échelle sur ce graphique (important).**

Expliquez, commentez et exploitez brièvement les résultats présentés, comme suit (questions 1 à 8) :

1. Retrouver les écarts-types σ_1,σ_2 des 2 variables, ainsi que leur coefficient de corrélation ρ.
2. Que veut dire ici "variables brutes" et "variables normalisées" ?
3. Que représente la matrice de covariance des CP; pourquoi est-elle diagonale; propriétés?
4. Quelle est la différence entre CP "brutes" et "normalisées"?
5. Ecrire explicitement le système de relations entre CP et variables "brutes"
6. Représenter graphiquement les CP "brutes" dans le plan (x1,x2), sur la figure.
7. Exprimer les régressions linéaires de x2/x1 et de x1/x2, respectivement.
8. Tracer les 2 droites de régression sur la figure. Sont-elles confondues et pourquoi ?

[1] Outre l'aspect "test" évoqué, cet exercice permet aussi de voir, dans le cas multivarié, ce qu'on peut faire comme analyses en prenant les variables 2 à 2.

REPONSES DU II.

1. Valeurs numériques des écarts-types s_1,s_2 et du coefficient de corrélation r des 2 variables?.

Il fallait "intuiter", comme l'énoncé le suggère, que les 2 vecteurs d'observations $(x1^{(i)},x2^{(i)})$ sont générés artificiellement avec des statistiques exactes (σ_1, σ_2, ρ) spécifiées d'avance. Ces statistiques sont sans doute bien reproduites car l'échantillon est assez grand (N=1000).

D'après cette remarque, et vu la matrice de covariance des variables brutes, on a :

$$\sigma_1 = 1.0311 \approx 1,$$

$$\sigma_2 = 2.0234 \approx 2,$$

$$\rho = -0.5072 \approx -1/2.$$

On peut chercher à confirmer l'hypothèse des chiffres ronds par un calcul rapide d'erreur d'échantillonnage (erreur quadratique moyenne ou erreur "*rms*")[2] normalisée par les écarts-types des variables (erreur *rms* relative). On sait *(cf.cours)* que l'erreur relative commise sur l'écart-type (et aussi sur la moyenne) est grossièrement en $1/\sqrt{N}$, avec ici $1/\sqrt{N} = 1/\sqrt{1000} \approx 1/30 \approx 3\%$. Or on constate qu'on a bien $\sigma_1=1$ et $\sigma_2=2$ à quelques % près. L'hypothèse est ainsi confortée.

Noter que les variables sont partiellement anti-corrélées, et que leurs écart-types sont significativement différents.

[2] De l'anglais "*rms* error": root-mean-square error.

2. Signification des variables "brutes" et "réduites"?.

Les variables brutes correspondent aux observations brutes ($x1^{(i)}$,$x2^{(i)}$)

Les variables normalisées ou réduites correspondent aux observations débarrassées de leur moyenne (estimée) et divisées par leur écart-type (estimé) :

$$(y1^{(i)},y2^{(i)}) = \left(\frac{x_1^{(i)} - m_1}{s_1}, \frac{x_2^{(i)} - m_2}{s_2} \right)$$

La normalisation conduit à homogénéiser les données d'une certaine manière : les nouvelles variables sont toutes adimensionnelles, de moyennes nulles et d'écart-types unités.

Il est équivalent de faire de l'analyse multivariée et de l'ACP sur la matrice de covariance des variables réduites ou sur la matrice de corrélation des variables brutes. Par contre, l'ACP en covariance brute n'est pas équivalente à l'ACP en corrélation ou en covariance réduite. L'ACP "réduite" a l'avantage de rendre toutes sortes de données comparables, mais peut avoir comme désavantage d'écraser les variables à fortes fluctuations. D'autres types de normalisations peuvent être envisagés selon la physique du problème (exemple : passer des débits aux débits spécifiques).

3. Signification de la matrice de covariance des CP (diagonalité, autres propriétés)?

Les CP obtenues par diagonalisation de la matrice de covariance des variables (brutes ou réduites) sont toujours orthogonales entre elles, c'est-à-dire non corrélées, et leur matrice de corrélation (ou de covariance) est donc diagonale.

On peut le montrer formellement comme suit (interprétation "algébrique" des CP).

Il n'y a ici que K=2 variables; organisons les K variables en un vecteur colonne, en laissant implicite les observations (i) :

$$\underline{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Soit $\underline{\underline{C}}_{XX} = \begin{bmatrix} Cx_1x_1 & Cx_1x_2 \\ Cx_2x_1 & Cx_2x_2 \end{bmatrix}$,

la matrice de covariance du vecteur $\underline{X}$. Cette matrice est symétrique définie-positive, et non-diagonale en général. La diagonalisation de $\underline{\underline{C}}_{XX}$ conduit aux relations ci-dessous (cf. **Annexe "Diagonalisation" ci-jointe**).

Soit P^T (**P transposée**) la matrice de passage de l'ancienne base (où la covariance est la matrice non-diagonale C_{XX}) vers la nouvelle base (où la covariance est devenue diagonale) :

$$(P^T): \quad \underline{\underline{C}}_{XX} \rightarrow \underline{\underline{C}}_{ZZ} = \underline{\underline{\Lambda}} = \underline{\underline{P}}^T \underline{\underline{C}}_{XX} \underline{\underline{P}}$$

$$(P): \quad \underline{\underline{C}}_{ZZ} = \underline{\underline{\Lambda}} \rightarrow \underline{\underline{C}}_{XX} = \underline{\underline{P}}\underline{\underline{\Lambda}}\underline{\underline{P}}^T$$

On montre que P est une matrice orthogonale satisfaisant :

$$\underline{\underline{P}}^T \underline{\underline{P}} = \underline{\underline{P}}\underline{\underline{P}}^T = \underline{\underline{I}}$$

La matrice P contient, en colonnes, les vecteurs propres normés de C_{XX} (P^T contient les mêmes vecteurs en ligne…), et Λ est la matrice diagonale contenant les valeurs propres λ de C_{XX}.

Le changement de base (P^T) ci-dessus transforme le vecteur des variables étudiées (X), de covariance C_{XX}, en un vecteur de variables (Z) dites *Composantes Principales* (C.P.) et dont la covariance est diagonale ($C_{ZZ} = \Lambda$).

$$(P^T): \quad \underline{X}^{(i)} \rightarrow \underline{Z}^{(i)} = \underline{\underline{P}}^T \underline{X}^{(i)}$$

$$(P): \quad \underline{Z}^{(i)} \rightarrow X^{(i)} = \underline{\underline{P}}\,\underline{Z}^{(i)}$$

où l'on a utilisé l'exposant " $^{(i)}$ " pour représenter les observations ou réalisations (i) avec i=1,…,N.

Le système $Z=P^TX$ exprime en particulier les C.P. (Z) en fonction des variables de départ (X). Ce système s'écrit, en notations indicielles :

$$Zk^{(i)} = Pjk^{(i)}\, Xj^{(i)} \quad \text{(somme sur j)}$$

avec j,k = 1,…,K pour les variables, et (i)=1,…,N pour les observations (par exemple ici K=2 et N=1000).

Montrons maintenant que l'on a bien $C_{ZZ} = \Lambda$, c'est-à-dire, que la covariance des C.P. est bien la matrice diagonale Λ des valeurs propres de C_{XX}. En supposant les variables centrées (pour simplifier…) :

$$\begin{aligned} C_{ZZ} &= \langle Z Z^T \rangle \\ &= \langle P^T X (P^T X)^T \rangle \\ &= \langle P^T X X^T P \rangle \\ &= P^T \langle X X^T \rangle P \\ &= P^T \, C_{XX} \, P \\ &= P^T (P L P^T) P \\ &= L \qquad \text{car } P^T P = I. \end{aligned}$$

(CQFD).

Enfin, il est clair que la trace de la matrice C_{XX} reste invariante après changement de base. On a donc :

$$Trace(C_{XX}) = Trace(C_{ZZ}) = Trace(L).$$

Selon que les variables de départ (X) sont "brutes" ou "réduites", on obtient ceci :

- Var.Brutes: *Trace* $= S\, l_j = S\, s_j^2$ » *5 ici.*
- Var.Réduites: *Trace* $= S\, l_j = K = 2$ *ici.*

Voir la trace de C_{ZZ} calculée : on voit que les propriétés ci-dessus sont satisfaites, ce qui est évidemment rassurant (test du programme de calcul des C.P).

4. Significations des CP brutes et des CP normalisées (différences?)

Cette question est liée à la question 2 sur les variables brutes/réduites. Brièvement :

i) Les CP "brutes" sont obtenues en diagonalisant la matrice de covariance des données brutes (X)

ii) Les CP "réduites" sont obtenues en diagonalisant la matrice de covariance des données réduites (qui est aussi la matrice de corrélation des données brutes…).

Les deux procédures ne donnent pas les mêmes résultats, comme on peut le constater en examinant la matrice de passage orthogonale P dans chaque cas. Dans la feuille de calcul, la matrice P est notée V en variables brutes et U en variables réduites. On voit bien que V≠ U.

En conclusion, l'ACP réduite n'est pas équivalente à l'ACP brute. Les CP brutes accordent le plus de poids aux variables dont les fluctuations sont les plus grandes dans les unités choisies. Les CP réduites ne prennent en compte que les corrélations, et non pas les intensités de fluctuations, ni les différences induites par le choix des unités; ce dernier point étant un avantage dans le cas de variables hétéroclites (débits, températures,…).

5. Formulation du système reliant les CP aux variables brutes (expression symbolique, numérique).

Voir la réponse à la question 3 (théorie), et voir aussi l'Annexe "Diagonalisation". On en tire une relation sur les C.P. de la forme $Z=P^TX$, ou encore, dans les notations de la feuille de calcul en variables brutes :

$$\underline{Z}^{(i)} = \underline{\underline{V}}^T \underline{X}^{(i)}$$

Ceci pour chaque observation (i). Pour alléger la notation, (i) sera implicite dans ce qui suit. Nous obtenons (cas bivarié):

$$\begin{cases} Z_1 = V_{11} X_1 + V_{21} X_2 \\ Z_2 = V_{12} X_1 + V_{22} X_2 \end{cases}$$

D'où finalement (application numérique) :

$$\begin{cases} Z_1 = +0.9539 X_1 + 0.3001 X_2 \\ Z_2 = -0.3001 X_1 + 0.9539 X_2 \end{cases}$$

La 1ère ligne de ce système définit la 1ère CP, et la 2nde ligne la 2nde CP, du moins avant reclassement. En effet, la matrice des valeurs propres donne (avant reclassement):

CP1: $\lambda_1 = Var(Z_1) = 0.7303$ (15%)

CP2: $\lambda_2 = Var(Z_2) = 4.4271$ (85%)

soit respectivement 15% et 85% de la variance totale. Noter que la 1ère CP n'est pas la plus grande.

En principe, il est conseillé d'achever la mise en forme des calculs en reclassant les CP par ordre décroissant, la plus grande valeur propre en 1er (nous ne le ferons pas dans cet exercice test).

6. Représentation graphique des CP brutes dans le plan (x1,x2).

En faisant varier les observations (i), le système ci-dessus définit un nuage de points dans le plan (X1,X2) et/ou dans le plan transformé des C.P. (Z1,Z2).

Pour obtenir l'équation des axes des CP, ou axes principaux, il suffit d'écrire :

Axe CP1 : $Z_2=0 : X_2 \approx +0.314\, X_1$

Axe CP2 : $Z_1=0 : X_2 \approx -3.180\, X_1$

Une difficulté d'interprétation apparaît dans le cas où les axes (X1,X2) sont représentés avec une distortion graphique (unités papier différentes suivant X1, X2). Les axes (CP1,CP2) n'apparaissent pas orthogonaux entre eux sur le graphique, alors qu'ils le sont en unités réelles.

Il suffit de retracer le nuage de points dans le plan (X1,X2) **sans distorsion** des échelles de longueurs pour se rendre compte en effet que :

i) Les axes CP1 et CP2 définis plus haut sont bien orthogonaux entre eux (comme il se doit);

ii) Les points observations apparaissent non corrélés dans le repère des axes principaux (comme il se doit);

iii) De plus, comme (X1,X2) sont ici conjointement gaussiens, les axes (CP1,CP2) sont aussi les axes principaux (grand et petit) des ellipses de concentration (isovaleurs) de la densité de probabilité gaussienne de (X1,X2).

Pour comprendre et préciser ce dernier point, il faut considérer la forme de la densité bivariée gaussienne $f_{X1,X2}(x_1,x_2)$ de deux variables corrélées. On montre que les isovaleurs d'équations $f_{X1,X2}(x_1,x_2)=f_0$ constante sont bien des ellipses déterminées par la matrice de covariance.[3]

[3] Ce problème est traité par ailleurs dans un Bureau d'Etude. Voir cours/poly R.A (vecteurs gaussiens), et H.Ventsel (Théorie des Proba., Moscou).

7. Expressions des régressions linéaires de x2/x1 et de x1/x2 (symboliquement, numériquement).

Formule classique de régression linéaire:

$$y = \boxed{a\,x + b} + \varepsilon = \boxed{y^*} + \varepsilon$$

où $y^* = a\,x + b$ représente l'estimation de y (modèle linéaire), et ε l'erreur d'estimation. Le critère y* sans biais impose :

$$b = m_Y - a\, m_X$$

et le critère de minimisation de variance d'erreur, Min Var(ε), donne :

$$a = \rho_{XY}\, \sigma_Y / \sigma_X$$

Au total, la régression s'écrit :

$$y^* = m_Y + a\,(x-m_X)$$

et la variance d'erreur (à l'optimum) est :

$$\sigma_\varepsilon^2 = (1-\rho^2)\, \sigma_Y^2$$

Appliquons ces formules aux données:

i) <u>Régression de X2/X1</u>:

$x_2 \approx \rho\, \sigma_2/\sigma_1\, x_1 + \varepsilon_2$
$x_2 \approx -0.995\, x_1 - 0.10 + \varepsilon_2$
$\approx - x_1 + \varepsilon_2$
avec $\sigma_{\varepsilon 2} \approx \sqrt{3} \approx 1.7$, d'où:
$I_{80\%} \approx \pm 1.28\, \sigma_{\varepsilon 2} \approx \pm 2.2$,
l'intervalle de confiance à 80% autour de $y=x_2$ ("verticalement").

ii) <u>Régression de X1/X2</u>:

$x_1 \approx \rho\, \sigma_1/\sigma_2\, x_2 + \varepsilon_1$
$x_1 \approx -0.258\, x_2 - 0.012 + \varepsilon_1$
$\approx - 0.25\, x_2 + \varepsilon_1$
avec $\sigma_{\varepsilon 1} \approx \sqrt{3}/2 \approx 0.9$, d'où:
$I_{80\%} \approx \pm 1.28\, \sigma_{\varepsilon 1} \approx \pm 1.1$,
l'intervalle de confiance à 80% autour de $y=x_1$ ("horizontalement").

8. Représentation graphique des régressions x2/x1 et x1/x2; sont-elles confondues (pourquoi)?

Voir figure ci-jointe. Les droites de régression ne sont pas confondues entre elles, et ne coïncident pas le grand axe de l'ellipse (qu'elles encadrent). La régression X2/X1 donne une estimation optimale de X2 conditionnée par les observations de X1 (gelées). Idem et vice-versa pour X1/X2…

ANNEXE DIAGONALISATION

Cette annexe traite de la diagonalisation d'une matrice (voir entre autres le Chap.1 du livre de *Y. Saâd, 1996*)[4] et en particulier de la diagonalisation de matrices de covariances ou de corrélations (qui sont par construction symétriques définies-positives).

Matrices $\underline{\underline{A}}$ et $\underline{\underline{B}}$ similaires (*similar matrices*) et similarités (*similarity transform*) :

Les matrices $\underline{\underline{A}}$ et $\underline{\underline{B}}$ sont similaires si elles sont liées par une transformation:

$$\underline{\underline{B}} \rightarrow \underline{\underline{A}} = \underline{\underline{P}}\ \underline{\underline{B}}\ \underline{\underline{P}}^{-1}$$

Cette transformation représente la matrice $\underline{\underline{B}}$ dans une autre base.

Elle préserve les valeurs propres λ :

$$\lambda_B \rightarrow \lambda_A = \lambda_B$$

Elle transforme les vecteurs propres $\underline{V}$ comme suit :

$$\underline{V}_B \rightarrow \underline{V}_A = \underline{\underline{P}}\ \underline{V}_B$$

Le changement de base correspondant s'écrit, pour n'importe quel vecteur :

$$\underline{U}_B \rightarrow \underline{U}_A = \underline{\underline{P}}\ \underline{U}_B$$

Diagonalisation d'une matrice $\underline{\underline{A}}$:

Définition : Une matrice carrée A est diagonalisable ssi $\exists\ \underline{\underline{D}}$ diagonale : $\underline{\underline{A}} = \underline{\underline{P}}\ \underline{\underline{D}}\ \underline{\underline{P}}^{-1}$

Théorème : Une conséquence de cette définition est que la matrice $\underline{\underline{A}}$ de taille (KxK) est diagonalisable ssi elle possède K vecteurs propres linéairement indépendants :

$$\underline{\underline{A}} = \underline{\underline{P}}\ \underline{\underline{D}}\ \underline{\underline{P}}^{-1} \Leftrightarrow \underline{\underline{A}}\ \underline{\underline{P}} = \underline{\underline{P}}\ \underline{\underline{D}} \Leftrightarrow \underline{\underline{A}}\ \underline{P}_j = \lambda_j\ \underline{P}_j \quad \text{(sans sommation sur j)}$$

Conséquence (corollaire) : On en déduit que les termes diagonaux λ_j de la matrice diagonale $\underline{\underline{D}}$ représentent les valeurs propres de $\underline{\underline{A}}$, mais aussi que les vecteurs colonnes $\underline{P}_j$ de la matrice $\underline{\underline{P}}$ représentent les vecteurs propres de $\underline{\underline{A}}$ (j=1,…,K).

Diagonalisation d'une matrice symétrique définie-positive $\underline{\underline{C}}$:

Une matrice symétrique définie-positive $\underline{\underline{C}}$ est toujours diagonalisable, et la matrice $\underline{\underline{P}}$ est dans ce cas une matrice orthogonale, véritablement "orthonormale" en fait (*unitary matrix*), telle que :

$$\underline{\underline{P}}^T\ \underline{\underline{P}} = \underline{\underline{P}}\ \underline{\underline{P}}^T = \underline{\underline{I}}$$

On a donc les relations suivantes pour la matrice $\underline{\underline{C}}$:

$$\underline{\underline{C}} = \underline{\underline{P}}\ \underline{\underline{D}}\ \underline{\underline{P}}^T \Leftrightarrow \underline{\underline{D}} = \underline{\underline{P}}^T\ \underline{\underline{C}}\ \underline{\underline{P}} \Leftrightarrow$$

$$\underline{\underline{C}}\ \underline{\underline{P}} = \underline{\underline{P}}\ \underline{\underline{D}} \Leftrightarrow \underline{\underline{C}}\ \underline{P}_j = \lambda_j\ \underline{P}_j \quad \text{(sans sommation sur j)}$$

On voit à nouveau que les vecteurs colonnes $\underline{P}_j$ (j=1,…,K) de la matrice $\underline{\underline{P}}$ sont les vecteurs propres $\underline{V}_C$ de la matrice $\underline{\underline{C}}$. Les vecteurs propres $\underline{V}_D$ de $\underline{\underline{D}}$, par construction, sont les vecteurs de la base dans laquelle $\underline{\underline{C}}$ est diagonalisée, et ils sont appelés "Composantes Principales" (C.P) de la matrice $\underline{\underline{C}}$. En résumé :

$$\lambda_C = \text{Diag}(\underline{\underline{D}}) ; \qquad \underline{V}_C = \text{Col}(\underline{\underline{P}}) = \underline{\underline{P}}\ \underline{V}_D ; \qquad \underline{V}_D = \underline{\underline{P}}^T\ \underline{V}_C$$

Conséquences en analyse statistique multivariée (matrice de covariance $\underline{\underline{C}}$) :

K variables: $\underline{X} \Rightarrow$ Covar(KxK): $\underline{\underline{C}}_{XX} \Rightarrow$ Diag: $\underline{\underline{D}} = \underline{\underline{P}}^T\ \underline{\underline{C}}_{XX}\ \underline{\underline{P}} \Rightarrow$ Comp.Princ.: $\underline{Z} = \underline{\underline{P}}^T \underline{X}$

[4] Y. Saâd, 1996*: Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston.

FEUILLE DE RESULTATS DE CALCULS DE L'EXERCICE (II): RESULTATS DE CORRELATION & DIAGONALISATION OU "A.C.P" (BIVARIEE)

Taille totale des vecteurs de données [x1],[x2]: **N = 1000**

Moyennes des vecteurs de données: **mu1, mu2 = -0.0138;+0.0037**

Matrice de covariance **CX** des données brutes [x1 x2] : **CX** =

1.0632	**-1.0581**
-1.0581	**4.0943**

Matrice de covariance **CY** des données réduites [y1 y2]: **CY=**

1.0000	**-0.5072**
-0.5072	**1.0000**

Matrice de rotation / diagonalisation des données brutes = matrice des vecteurs propres [v1 v2] = **VX =**

0.9539	**-0.3001**
+0.3001	**0.9539**

Matrice de rotation / diagonalisation des données réduites ou matrice des vecteurs propres [u1 u2] = **UY** =

-0.7071	**-0.7071**
+0.7071	**-0.7071**

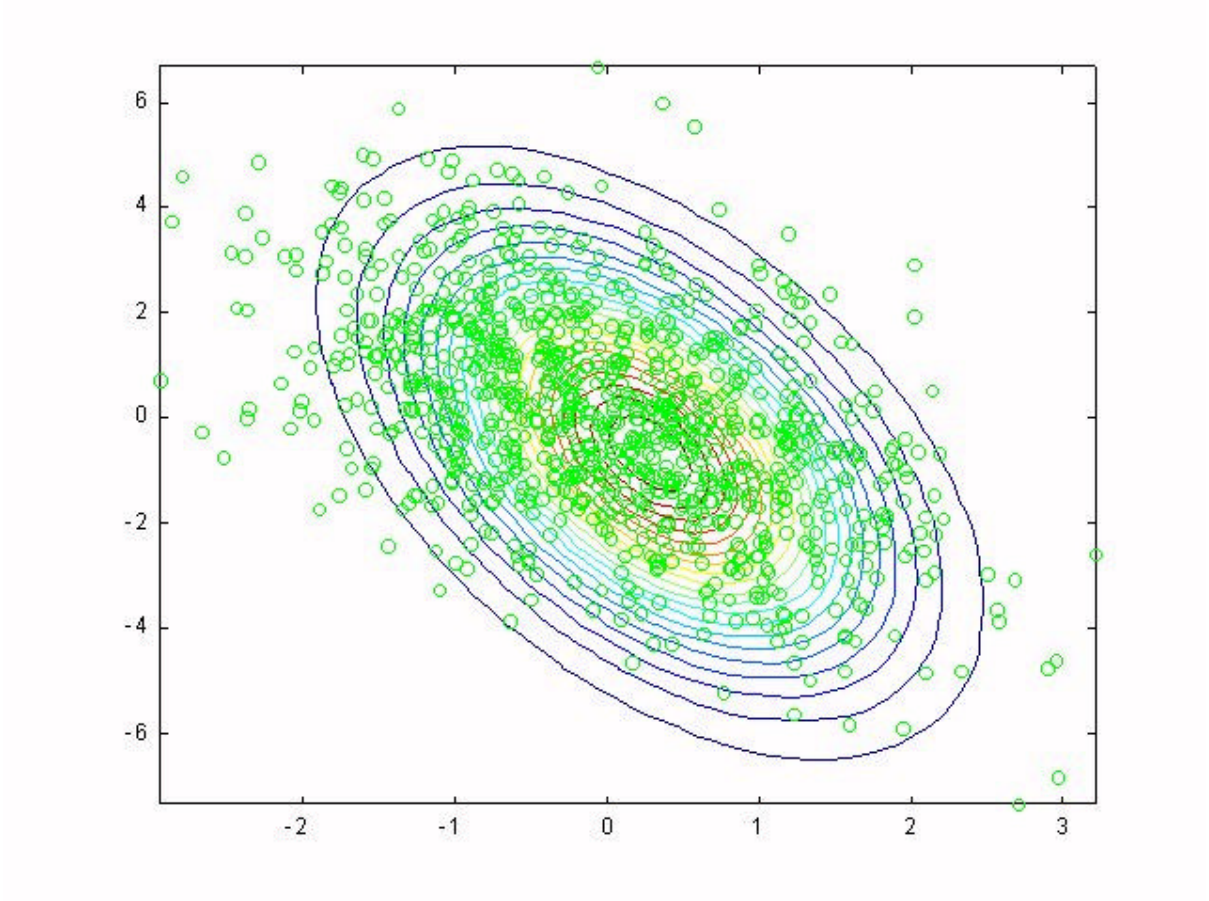
Matrice de covariance **CZ** des CP "brutes" [z1 z2]: **CZ =**

0.7303	**0.0000**
0.0000	**4.4271**

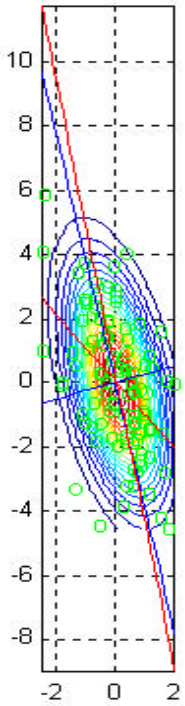
Matrice de covariance **CW** des C.P. "réduites" [w1 w2]: **CW =**

1.5072	**0.0000**
0.0000	**0.4928**

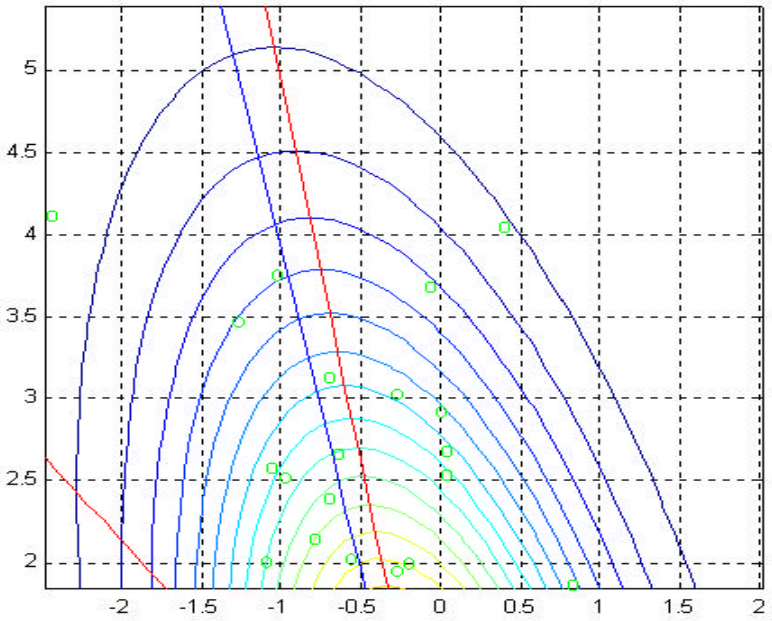
"ANALYSE BIVARIEE" : FIGURE INDICATIVE

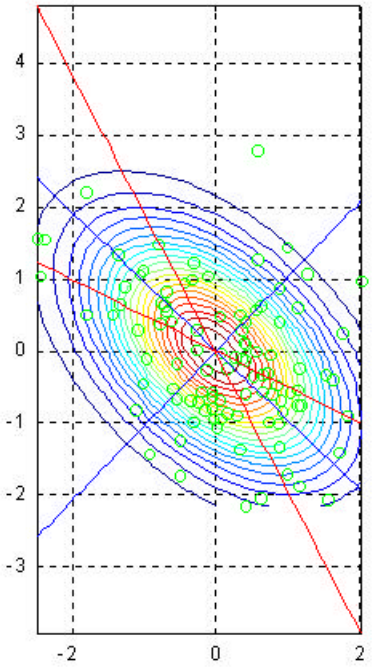


ANALYSE CORRELATOIRE BIVARIEE :
VISUALISATION D'UNE PARTIE DES 1000 PAIRES DE POINTS-OBSERVATIONS BRUTES [X1,X2].
Remarquer que les échelles des coordonnées sont très distordues (rapport d'affinité 1/3 environ).

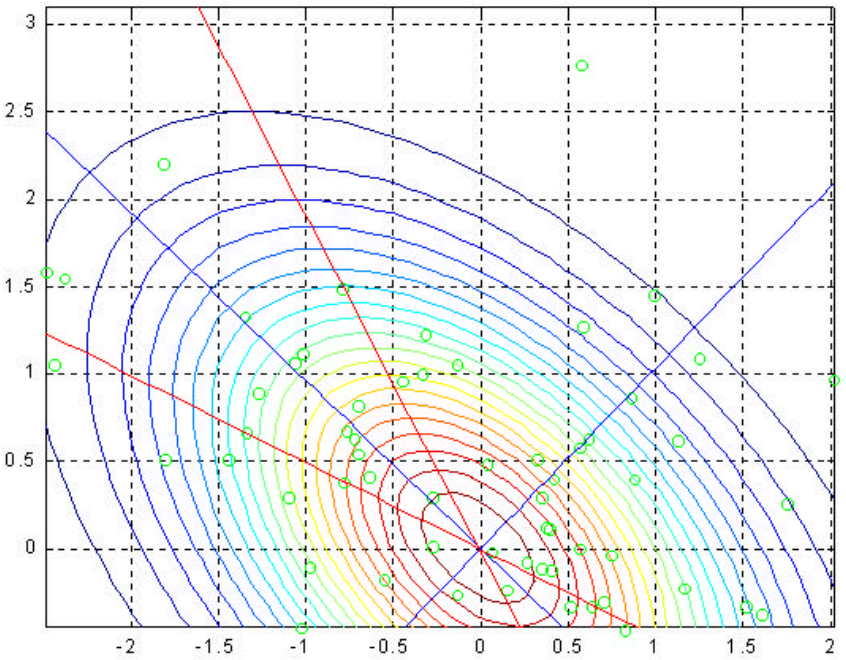


VISUALISATION DE LA CORRÉLATION BIVARIÉE (X_1,X_2) EN ÉCHELLES "PEU DISTORDUES" : AXES PRINCIPAUX, ELLIPSES DE CONCENTRATION, ET DROITES DE RÉGRESSION.
MOMENTS THÉORIQUES r = -0.5, s_1 = 1, s_2 = 2, m_1 = 0, m_2 = 0 (N=100 POINTS).
EN HAUT : VISUALISATION COMPLÈTE. EN BAS : DÉTAIL (ZOOM).





VISUALISATION DE LA CORRÉLATION BIVARIÉE (X_1,X_2) EN ÉCHELLES "PEU DISTORDUES" : AXES PRINCIPAUX, ELLIPSES DE CONCENTRATION, ET DROITES DE RÉGRESSION.
MOMENTS THÉORIQUES r = -0.5, s_1 = 1, s_2 = 1, m_1 = 0, m_2 = 0
N = 1000 POINTS (SEULEMENT 100 POINTS REPRÉSENTÉS GRAPHIQUEMENT).
EN HAUT : VISUALISATION COMPLÈTE. EN BAS : DÉTAIL (ZOOM).



RÉSULTATS STATISTIQUES :

VALEURS NUMÉRIQUES OBTENUES POUR LES MOMENTS THÉORIQUES

r = -0.5, s_1 = 1, s_2 = 1, m_1 = 0, m_2 = 0 AVEC N = 1000 POINTS

(données gaussiennes générées selon ces spécifications par le programme Matlab)

```
Total size of gaussian data vectors [x1],[x2]: ... N =
        1000

 Input correl. coeff. of gaussian vectors [x1],[x2]: Rho =
   -0.5000

 Computed correl coefficient of gaussian vectors:  rho =
   -0.5072

 Input means of gaussian vectors [x1],[x2]: ..Mu1, Mu2 =
     0
     0

 Computed means of gaussian vectors :  ....mu1=...,mu2 =
   -0.0138
    0.0019

 Input std. dev. of gaussian vectors: ...Sigma1, Sigma2 =
     1
     1

 Computed std. dev. of gaussian vectors: sigma1=..,sigma2=
    1.0311
    1.0117

 Covariance matrix of raw data [x1 x2] : .......... CX =
    1.0632   -0.5291
   -0.5291    1.0236

 Covariance matrix of normalized data [x1 x2] : ... CY =
    1.0000   -0.5072
   -0.5072    1.0000

 Raw data : Rotation matrix=eigenvectors [v1 v2]: ..VX =
   -0.7202   -0.6938
    0.6938   -0.7202

 Norm. data: Rotation matrix=eigenvectors [u1 u2]: ..UY =
   -0.7071   -0.7071
    0.7071   -0.7071

 Raw data: Covar matrix of principal compon.[z1 z2]:CZ =
    1.5728    0.0000
    0.0000    0.5139

 Norm. data: Covar matrix of principal compon.[w1 w2]:CW =
    1.5072    0.0000
    0.0000    0.4928

 Pentes des régressions linéaires de x2/x1 (a21) et de x1/x2 (aa21=1/a12):
   -0.4976   -1.9347
```

ANNEXE "MATLAB" :

PROGRAMME TEST D'ANALYSE STATISTIQUE MULTIVARIEE & A.C.P EN LANGAGE MATLAB

```
% ========================================================================================
% STAT*ACP.M :        Programme d'Analyse Statistique Multivariée et A.C.P.
% Ver.4 (20Mars00):  Test bivarié de DEMO avec "données" générées en interne
% ========================================================================================
% Objectifs: CORRÉLATION ET RÉGRESSION MULTIPLES + ANALYSE EN COMPOSANTES PRINCIPALES
% ========================================================================================
% Auteur:    R.ABABOU, depuis 1996+ (Cours d'Hydrologie Statistique, N7)
% ========================================================================================
% Dates:     Mars 1996 :  Version # 0.0 (test préliminaire)
%            22 Fev.97 + 20 Jan.2000: Version # 0.1 (test préliminaire retouché)
%            24 Jan.2000:  Version # 0.2 (correction de plots (!) + axes ACP)
% ========================================================================================
% Features:  -Commentaires:      bilingual comments (french/anglais)
%                  -Matlab algebra/statistics:      cov(); eig();
%                         -Matlab I/O modules:              disp();
%                         -Matlab string operators : num2str(); str2mat(S1,S2,...);
% ========================================================================================

  clear all;

% ----------------------------------------------------------------------------------------
% DONNEES D'ENTREE POUR LA VERSION "DEMO" BIVARIEE DE STAT*ACP (voir ci-dessous)
% ----------------------------------------------------------------------------------------
%  N_tot=1000;      % Total number of points in stat.analyzis (usually < 10000)
%  N_plot_max=100;        % Max number of points displayed in plots (should be << 1000)
%  Rho=-0.50;             % Theoretical correlation coefficient, must be in [-1,+1]
%  Sigma1=1; Mu1=0;       % Theoretical mean (mu1) and std.dev.(sigma1) of X1
%  Sigma2=1; Mu2=0;   %  Theoretical mean (mu2) and std.dev.(sigma2) of X2
% ----------------------------------------------------------------------------------------
% NOTE ON SYNTAX OF INPUT('string'): USE (`) INSTEAD OF (') WITHIN CHARACTER STRINGS
% ----------------------------------------------------------------------------------------
input(' Matlab program ACP*STAT.M by R.Ababou (1996-2000) - TO CONTINUE, PRESS RETURN...');
input(' Analyse statistique multivariée et A.C.P : DEMO pour le cas bivarié - RETURN...');
input(' On donnera ci-dessous les moments bivariés des données gaussiennes - RETURN...');
% ----------------------------------------------------------------------------------------
  N_tot=input(' ENTER total number of data points (X1,X2) to be analyzed (usually < 10000):');
  N_plot_max=input(' ENTER max number of points to be shown in plots (should be << 1000): ');
  Rho=input(' ENTER theoretical correlation coefficient Rho in [-1,+1], for example -0.5: ');
  Sigma1=input(' ENTER theoretical standard deviation (Sigma1) of X1, for example 1.0 : ');
  Sigma2=input(' ENTER theoretical standard deviation (Sigma2) of X2, for example 2.0 : ');
  Mu1=input(' ENTER the theoretical mean (Mu1) of X1, for example 0.0 : ');
  Mu2=input(' ENTER the theoretical mean (Mu2) of X2, for example 0.0 : ');
% ----------------------------------------------------------------------------------------
  PDF2PLOT=1;   % Option traçage ellipses (courbes iso-probabilités bivariées)
  ncontours=20; % Nombre de contours/ellipses pour plots DdP et Proba (FdR)
  vcontours=[0.05 0.10 0.20 0.50 0.80 0.90 0.95]; % Contours : Isovaleurs de Proba (FdR)
% ----------------------------------------------------------------------------------------

  N = N_tot;         % Noter l'alias utilisé : équivalence N <====> N_tot
  Rho2=Rho*Rho;           % Notation: Rho2 is the Squared Rho

% ----------------------------------------------------------------------------------------
% Premièrement:  CONSTRUCTION DE 2 VECTEURS D'OBSERVATIONS GAUSSIENS ET CORRÉLÉS:
%      ON UTILISE POUR CELA UN GÉNÉRATEUR DE NOMBRES ALÉATOIRES INTERNE À MATLAB,
%     "RANDN()", QUI GÉNÈRE DES NOMBRES RÉELS ALÉATOIRES GAUSSIENS NORMALISÉS N(0,1).
% ----------------------------------------------------------------------------------------

  randn('seed',0);   % Reset the seed at its startup value
  xr=randn(N,1);     % Generate two normalized gaussian vectors
  yr=randn(N,1);     % (xr) and (yr) for testing (or simulation)

% CONSTRUCTION OF TWO CORRELATED VARIABLES:
```

```
  x1=Mu1+Sigma1*xr;               % note that xr and yr are independent,
  x2=Rho.*xr + sqrt(1-Rho2).*yr;  % while x1 and x2 are now correlated;
  x2=Mu2+Sigma2*x2;               % note that x1 and x2 are not normalized.

% ------------------------------------------------------------------------------------------
% Deuxièmement :  ANALYSE DE VARIANCES-COVARIANCES DES 2 VECTEURS D'OBSERVATIONS
% ------------------------------------------------------------------------------------------
% FORMATS: NOTER QUE X EST LE VECTEUR LIGNE DES 2 VARIABLES (OU MATRICE RECTANGLE AVEC
OBSERVATIONS EN COLONNES)
% ------------------------------------------------------------------------------------------

  X=[x1 x2];
  mu1=mean(x1); sigma1=std(x1); % Estimated means and standard deviations
  mu2=mean(x2); sigma2=std(x2); % of each raw dataset, x1 and x2.

  CX=cov(X);                   % 2x2 covariance matrix of raw data

  rho=(CX(1,2)/sigma1)/sigma2;  % Estimated correlation coefficient
  rho2=rho*rho;                 % Square of estimated correlation coeff.
  var1=sigma1*sigma1;           % Estimated variance of x1
  var2=sigma2*sigma2;           % Estimated variance of x2

  CY=[ 1,rho ; rho,1] ;         % 2x2 correlation matrix of raw data
                               %(2x2 covar matrix of normalized data)
  [VX,DX]=eig(CX);            % Eigenvectors and eigenvalues of CX
  [UY,DY]=eig(CY);            % Eigenvectors and eigenvalues of CY

  v1=VX(:,1);              v2=VX(:,2);
  v1=v1./norm(v1);   v2=v2./norm(v2);
  VX=[v1 v2];
  u1=UY(:,1);              u2=UY(:,2);
  u1=u1./norm(u1);   u2=u2./norm(u2);
  UY=[u1 u2];

%%z1=VX(1,1)*x1+VX(2,1)*x2;  % OK à un détail près : remplacer xj=>(xj-muj) (j=1,2)
%%z2=VX(1,2)*x1+VX(2,2)*x2;  % ceci afin d'obtenir z1=z2=0 au point x1=mu1 et x2=mu2
  z1=VX(1,1)*(x1-mu1)+VX(2,1)*(x2-mu2);
  z2=VX(1,2)*(x1-mu1)+VX(2,2)*(x2-mu2);
  Z=[z1 z2];                % Composantes Principales Brutes 1 et 2 (non classées ici)
  CZ=cov(Z);                % Covariance matrix of raw principal components

  w1=UY(1,1)*(x1-mu1)/sigma1+UY(2,1)*(x2-mu2)/sigma2;
  w2=UY(1,2)*(x1-mu1)/sigma1+UY(2,2)*(x2-mu2)/sigma2;
  W=[w1 w2]; % C.Principales Normalisées 1 et 2 (non classées pour l'instant)
  CW=cov(W); % Covariance matrix of normalized principal components

% ------------------------------------------------------------------------------------------
% Troisièmement :  RÉGRESSION LINÉAIRE SIMPLE DES VARIABLES PRISES 2 A 2 :
%                  2 VECTEURS D'OBSERVATIONS X1,X2 ==> REG.LIN. X2/X1 ET REG.LIN. X1/X2
%                  (À COMPLÉTER PLUS TARD PAR DE LA RÉGRESSION MULTIPLE)
% ------------------------------------------------------------------------------------------

% Formule de régression simple :              Y -My = a * (X - Mx) + E
% Pente :                                     a = Rho * sigmaY / sigmaX
% Ordonnée à l'origine :                             b = My - a * Mx
% Variance d'erreur :                                      sigmaE**2 = (1-Rho**2) * sigmaY**2

% Régression de x2/x1 : x2 = a21*x1 + b2 + e2 => Estim(x2) = a21*x1 + b2
  a21=rho*sigma2/sigma1; b2=mu2-a21*mu1; sigma2E=sqrt((1-rho2))*sigma2;

% Régression de x1/x2 : x1 = a12*x2 + b1 + e1 => Estim(x1) = a12*x2 + b1
  a12=rho*sigma1/sigma2; b1=mu1-a12*mu2; sigma1E=sqrt((1-rho2))*sigma1;

% Régression de x1/x2 exprimée en termes de x2/x1  =>  x2 = aa21*Estim(x1) + bb2
  aa21 = 1/a12; bb2 = -b1/a12;

% ------------------------------------------------------------------
% PRÉPARATION DES OUPUTS (PROVISOIRES -- PROGRAMME TEST PRÉLIMINAIRE):
% ------------------------------------------------------------------
```

```
  N_tot  = N;   N_plot = min(N_tot,N_plot_max);

% !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
if PDF2PLOT==1, % Begin plots (pre-processing and graphics)
% !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

  S_plot = num2str(N_plot);
  S_Plot = str2mat(' # of points Nplot  = ',S_plot);
  S_Plot = [S_Plot(1,:) S_Plot(2,:)];
  S_tot  = num2str(N_tot);
  S_Tot  = str2mat(' # of points Ntotal = ',S_tot);
  S_Tot  = [S_Tot(1,:) S_Tot(2,:)];
  T_Title= str2mat(S_Tot,S_Plot);
  T_Title= [T_Title(1,:) T_Title(2,:)]; % Produces a one line title
  x1plot=x1(1:N_plot);
  x2plot=x2(1:N_plot);                  % Data points to be displayed
  x1min=min(x1plot); x1max=max(x1plot);
  x2min=min(x2plot); x2max=max(x2plot);

% ------------------------------------------------------------------------
%     COLOR RASTER MAP AND LINE CONTOUR PLOTS OF JOINT GAUSSIAN P.D.F AND C.D.F
%     CARTE COULEUR RASTER ET ISOVALEURS DES D.D.P ET F.D.R GAUSSIENNES BIVARIÉES
% ------------------------------------------------------------------------

% Dimensions grille 2D pour images raster
  M1=51;   M2=51;
% M1M2max=max(M1,M2);M1=M1M2max;M2=M1M2max; % Pour évacuer pbs tableaux transposés(24Jan.2000)
  L1=x1max-x1min; L2=x2max-x2min;
% L1L2max=max(L1,L2);L1=L1L2max;L2=L1L2max; % Pour évacuer pbs tableaux transposés(Jan.2000)
  Dx1=L1/(M1-1); Dx2=L2/(M2-1);  % Avec ces ajouts, on aura(it) Dx1=Dx2 par construction

% Correction erreur programmation/plot, à partir d'ici et plus bas: x1==>x1grid etc(24Jan2000)
  x1grid=[0:1:M1-1]; x1grid=x1min+(Dx1.*x1grid);
  x2grid=[0:1:M2-1]; x2grid=x2min+(Dx2.*x2grid);   % xjgrid: vecteur de coord./axe 1D
  [X1grid,X2grid]=meshgrid(0:1:M1-1,0:1:M2-1);       % Xjgrid: matrice de coord./grille 2D
  X1grid=x1min+(Dx1.*X1grid);
  X2grid=x2min+(Dx2.*X2grid);

% ------------------------------------------------------------------------
% X1grid=X1grid';X2grid=X2grid'; % Transposées ou non, selon version Matlab:ATTENTION!
% ------------------------------------------------------------------------

  XX1grid= (X1grid-mu1)/sigma1; % Use "./" or "/" (depending on Matlab version)
  XX2grid= (X2grid-mu2)/sigma2;
  X1X2grid=XX1grid.*XX2grid;
  XX1grid= XX1grid.*XX1grid;
  XX2grid= XX2grid.*XX2grid;

% CALCUL DE LA DDP (PDF) BIVARIÉE SUR LA GRILLE BIDIMENSIONNELLE
% ...À COMPLÉTER PAR LE CALCUL DES ELLIPSES DE PROBABILITÉ (...)
% Voir inputs+haut: ncontours=20; % Nombre de contours/ellipses: DdP + Proba(FdR)
% Voir inputs+haut: vcontours=[0.05 0.10 0.20 0.50 0.80 0.90 0.95]; % Contours:Proba
  cstant= 2*pi*sigma1*sigma2*sqrt(1-rho2);
  pdf2Dgrid= -0.5 * (XX1grid - 2*rho*X1X2grid + XX2grid) / (1-rho2);
  pdf2Dgrid= exp(pdf2Dgrid)/cstant;
% Pour référence, rappelons que la F.d.R (C.D.F) gaussienne univariée est:
% s2=1/sqrt(2); cdf1D=0.5*(1+erf(s2.*(X1grid-mu1)./sigma1));

% Equations des axes principaux "bruts" (axes des CP en variables brutes: voir Z + haut):
%    Eq. de l'axe principal "CP1" (z1) dans le repère (x1,x2) ( Droite z2=0 )
     yCP1=Mu2-(VX(1,2).*(x1grid-Mu1))./VX(2,2); % Correction par Mu1 et Mu2 (OK c'est fait)
%    Eq. de l'axe principal "CP2" (z2) dans le repère (x1,x2) ( Droite z1=0 )
     yCP2=Mu2-(VX(1,1).*(x1grid-Mu1))./VX(2,1); % Correction par Mu1 et Mu2 (OK c'est fait)

% EQUATIONS DES DROITES DE RÉGRESSION LINÉAIRES (CALCULÉES PLUS HAUT, EN VARIABLES BRUTES):
  x2REGx2x1 = ( a21*x1grid) +  b2 ; % Droite de régression de x2/x1.
  x2REGx1x2 = (aa21*x1grid) + bb2 ; % Droite de régression de x1/x2 (exprimée en x2/x1).

% ============================================================================
% START PLOTTING :
% ============================================================================
```

```
% We use Matlab's new plot options >> plot(x,y,'LineSTyle','.','Marker','o','Color','r');

% FIGURE(1): OPENING A NEW FIGURE FOR RASTER COLOR MAP (WITH LINE CONTOURS)
% ----------
  figure;
  colormap('hot'); % Choosing a colormap
  pcolor(x1grid,x2grid,pdf2Dgrid),  % Pixel color plot (works like SURF)
  hold on,       % Interversion of x1grid,x2grid may be needed (CAUTION!)
%%contour(x2grid,x1grid,pdf2Dgrid,vcontours,'k-'),
% This adds selected contours to pix map
%%hold on,
% contour(x2grid,x1grid,pdf2Dgrid,ncontours,'k--'),
% This adds many contours (automatic)
% hold on,  % Some available colors are: k=black,y=yellow,g=green,c=cyan
  plot(x1plot,x2plot,'LineStyle','.','Marker','o', 'Color','m');
% Plot a cloud of points '.' or 'o'
  plot(x1grid,yCP1,'LineStyle','-','Color','b');
  plot(x1grid,yCP2,'LineStyle','-','Color','b');
  plot(x1grid,x2REGx2x1,'LineStyle','-','Color','g');
  plot(x1grid,x2REGx1x2,'LineStyle','-','Color','g');
  hold off;            % Available LineStyles: .=dots/null,--=dashed,-=solid
  title(T_Title); xlabel('X');ylabel('Y'); % Older command: plot(...,'LineStyle','o');
  shading('faceted');       % Shading options (faceted, flat, interp)
  colorbar;                 % Display the colorbar inside the figure
  colormenu;                % Display colormenu on top of figure window
  axis equal;               % Plot with equal tick marks (true x,y scales)
% pause(5);

% FIGURE(2) SUPPRIMÉE: OPENING A SECOND NEW FIGURE FOR A 2ND RASTER COLOR MAP
% ------------------- (WITH LINE CONTOURS+TRUE SCALE)
% figure;
% colormap('hot'); % Choosing a colormap
% pcolor(x1grid,x2grid,pdf2Dgrid),  % Pixel color plot (works like SURF)
% hold on,         % Interchange of x1grid,x2grid may be needed (CAUTION!)
%%contour(x2grid,x1grid,pdf2Dgrid,vcontours,'k-'), % Add selected contours to pixel plot
%%hold on,
% contour(x2grid,x1grid,pdf2Dgrid,ncontours,'k--'),% Add a number of contours (automatic)
% hold on,  % Some available colors are : k=black,y=yellow,g=green,c=cyan
% plot(x1plot,x2plot,'LineStyle','.','Marker','o', 'Color','m');
%%Plot a cloud of points '.' or 'o'
% plot(x1grid,yCP1,'LineStyle','-','Color','b');
% plot(x1grid,yCP2,'LineStyle','-','Color','b');
% plot(x1grid,x2REGx2x1,'LineStyle','-','Color','g');
% plot(x1grid,x2REGx1x2,'LineStyle','-','Color','g');
% hold off;            % Available LineStyles: .=dots/null,--=dashed,-=solid
% title(T_Title); xlabel('X');ylabel('Y'); % Older command: plot(...,'LineStyle','o');
% shading('faceted');       % Shading options (faceted, flat, interp)
% colorbar;                 % Display the colorbar inside the figure
% colormenu;                % Display colormenu on top of figure window
% axis image; % Plot within a box containing ALL data points (AND with true equal x,y scales)
%%pause(5);

% FIGURE(3): OPENING A NEW FIGURE FOR LINE CONTOUR PLOT (WITHOUT COLOR MAP)
% ----------
  figure;
%%contour(x2grid,x1grid,pdf2Dgrid,vcontours,'k-'),
%%hold on,
  contour(x1grid,x2grid,pdf2Dgrid,ncontours),
%Interchange of x1,x2 may be needed:CAUTION!
  hold on,
% This plots a cloud of points '.' or 'o'
  plot(x1plot,x2plot,'LineStyle','.','Marker','o','Color','g');
  plot(x1grid,yCP1,'LineStyle','-','Color','b');
  plot(x1grid,yCP2,'LineStyle','-','Color','b');
  plot(x1grid,x2REGx2x1,'LineStyle','-','Color','r');
  plot(x1grid,x2REGx1x2,'LineStyle','-','Color','r');
  hold off;
  title(T_Title);
  xlabel('X');ylabel('Y');
  grid on;
  axis equal;           % Plot with equal tick marks (true x,y scales)
```

```
% pause(5);

% FIGURE(4): OPENING A SECOND NEW FIGURE FOR LINE CONTOUR PLOT
% --------- (WITHOUT COLOR MAP AND WITH TRUE SCALES)
  figure;
%%contour(x2grid,x1grid,pdf2Dgrid,vcontours,'k-'),
%%hold on,
  contour(x1grid,x2grid,pdf2Dgrid,ncontours),
% Interchange of x1,x2 may be needed:CAUTION!
  hold on,
  plot(x1plot,x2plot,'LineStyle','.','Marker','o','Color','g');
% This plots a cloud of points '.' or 'o'
  plot(x1grid,yCP1,'LineStyle','-','Color','b');
  plot(x1grid,yCP2,'LineStyle','-','Color','b');
  plot(x1grid,x2REGx2x1,'LineStyle','-','Color','r');
  plot(x1grid,x2REGx1x2,'LineStyle','-','Color','r');
  hold off;
  title(T_Title);
  xlabel('X');ylabel('Y');
  grid on;
  axis image; % Plot within a box containing ALL data points (AND with true equal x,y scales)
% pause(5);

% !!!!!!!!!!!!!!!!!!!!
end;   % End of Plots
% !!!!!!!!!!!!!!!!!!!!

% ------------------------------------------------------------------------------
% OUTPUTS STATISTIQUES NON-GRAPHIQUES : MOMENTS UNI+BIVARIÉS, COMP.PRINCIPALES, ETC
%-------------------------------------------------------------------------------
  disp(' Total size of gaussian data vectors [x1],[x2]: ... N = '); disp(N);
  disp(' Input correl coeff of gaussian vectors [x1],[x2]: Rho = '); disp(Rho);
  disp(' Computed correl coefficient of gaussian vectors:  rho = '); disp(rho);
  disp(' Input means of gaussian vectors [x1],[x2]: ..Mu1, Mu2 = ');
  disp(Mu1);disp(Mu2);
  disp(' Computed means of gaussian vectors :  ....mu1=...,mu2 = ');
  disp(mu1);disp(mu2);
  disp(' Input std.dev. of gaussian vectors: ...Sigma1, Sigma2 = ');
  disp(Sigma1);disp(Sigma2);
  disp(' Computed std.dev. of gaussian vectors: sigma1=..,sigma2= ');
  disp(sigma1);disp(sigma2);
  disp(' Covariance matrix of raw data [x1 x2] : .......... CX = '); disp(CX);
  disp(' Covariance matrix of normalized data [x1 x2] : ... CY = '); disp(CY);
  disp(' Raw data : Rotation matrix=eigenvectors [v1 v2]: ..VX = '); disp(VX);
  disp(' Norm.data: Rotation matrix=eigenvectors [u1 u2]: ..UY = '); disp(UY);
  disp(' Raw data: Covar matrix of principal compon.[z1 z2]:CZ = '); disp(CZ);
  disp(' NormData: Covar matrix of principal compon.[w1 w2]:CW = '); disp(CW);
  disp(' Pentes des régressions linéaires x2/x1 (a21) et x1/x2 (aa21=1/a12): ');...
                                                          disp([a21 aa21]);

% ============================================================================
%       REMARQUES FINALES :
% ----------------------------------------------------------------------------
% 1) Z est la matrice des Composantes Principales z1,z2 (vecteurs),
%       combinaisons linéaires des vecteurs d'observations x1 et x2.
% 2) Ayant invoqué cov(Z), on peut vérifier que les C.P. ne sont pas
%       corrélées entre elles : vérifier donc que cov(Z) est bien la
%       matrice diagonale DX contenant, comme termes diagonaux, les
%       valeurs propres de la matrice CX.
% 3) Avant de continuer l'analyse en terme de C.P., reclasser les C.P. par
%       ordre d'importance décroissante, en reclassant les valeurs propres
%       correspondantes de la plus grande à la plus petite en valeur absolue.
% ============================================================================

% ============================================================================
% END: this ends the matlab test program STAT*ACP.M  (R.Ababou 1996-2000).
% ============================================================================
```