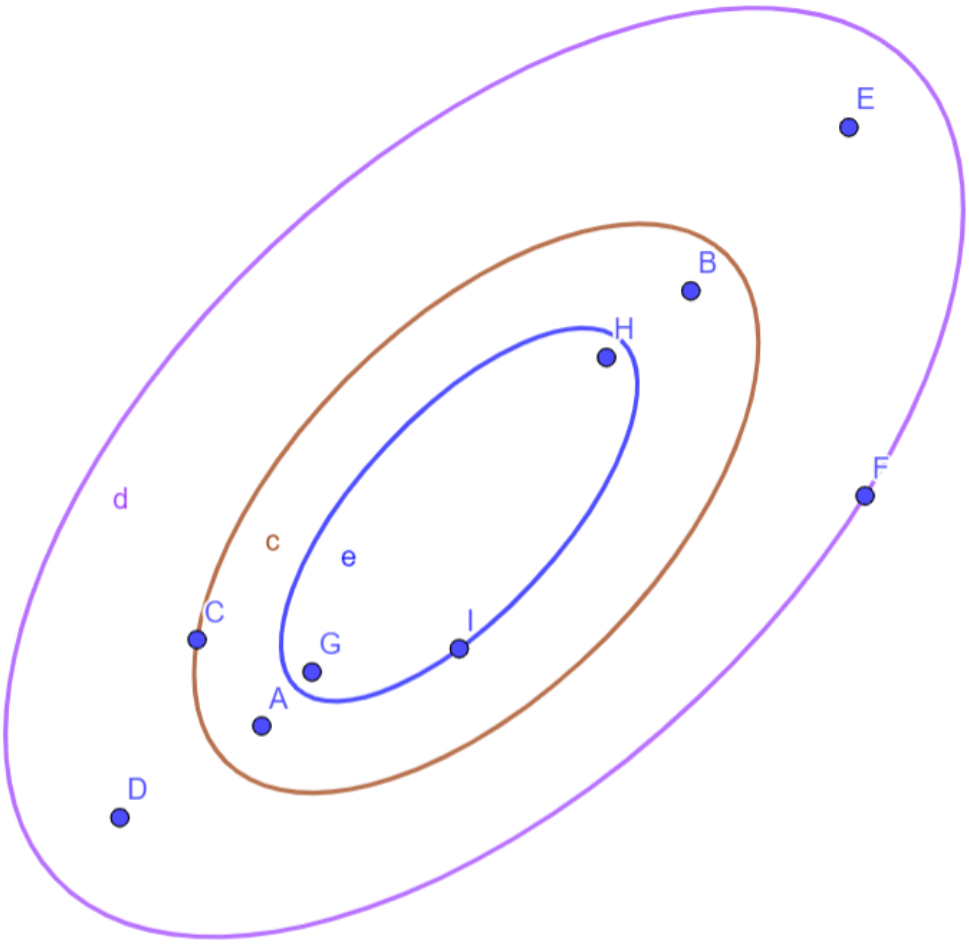
Projet Mathématiques

Modélisation d'une loi Normale Bidimensionnelle



COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Sommaire

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Introduction

L'objectifs de ce rapport est d'étudier et de modéliser une variable aléatoire Z suivant une loi normale bidimensionnelle caractérisée par ses paramètres μ et Σ sa matrice de covariance.

Dans cette étude, nous considérons d'abord Σ comme définie positive pour éviter les cas dégénérés.

La première partie consiste en une étude mathématique de la densité de probabilité f_Z associée à Z d'abord en caractérisant ses contours d'isodensité, puis en trouvant des estimateurs μ et Σ dont on justifiera la pertinence.
S'en suit une étude des cas dégénérés où l'on va considérer d'abord une, puis deux valeurs propres nulles de la matrice de covariance Σ. Cette étude sera notamment utile dans la compréhension de ce que la matrice de covariance apporte à une distribution de données.

La seconde partie propose des simulations numériques, visuelles, de ce qui a été montré dans l'étude mathématique ainsi qu'un calcul et une analyse des estimateurs.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Etude Mathématiques

I. Démontrer que les lignes d'isodensité de la densité de probabilité de la loi normale bidimensionnelle sont des ellipses

Soit Z une variable aléatoire suivant une loi normale bidimensionnelle d'espérance μ ∈ $\mathbb{R}^2$ et de matrice de covariance ∑ ∈ $M_2(\mathbb{R})$.∑ est symétrique et semi-définie positive. Soit z un couple de point de $\mathbb{R}^2$,la densité de probabilité de Z s'exprime :

$$\text{(E)} \qquad f_Z(z) = \frac{1}{\sqrt{(2\pi)^2 \det \sum}} \exp\left(-\frac{1}{2}{}^t(z-\mu)\sum{}^{-1}(z-\mu)\right) \quad \forall\, m \in \mathbb{R}^2$$

Afin d'éviter les cas dégénérés, on considérera dans la suite la matrice Σ définie positive.

Le but de cette première partie est d'étudier les lignes d'isodensité de f_Z, puis de calculer la probabilité qu'un point tiré selon la loi de Z appartienne à la surface interne de cette courbe, qui est une ellipse dont les caractéristiques dépendent de μ et Σ. Plus précisément, pour tout p ∈ [0 ; 1], il s'agira de déterminer la ligne d'isodensité K délimitant une surface intérieure S_K telle que P(Z ∈ S_K) = p.

Pour commencer nous proposons de décrire les courbes d'isodensité K c'est-à-dire les courbes C_K de surface interne S_K tel que f_Z(m= (X, Y)) =K.

Que l'on peut réécrire en vue de (E) :

$$\text{(E-0)} \quad K = \frac{1}{\sqrt{(2\pi)^2 \det \sum}} \exp\left(-\frac{1}{2}{}^t(m-\mu)\sum{}^{-1}(m-\mu)\right)$$

Nous remarquons que vu la définition de Σ c'est-à-dire semi-définie positive et symétrique, Σ est diagonalisable en base orthonormée et d'après la décomposition spectrale :

$\Sigma = \mathrm{U}\,\Lambda\,\mathrm{U}^{-1} = \mathrm{U}\,\Lambda\,\mathrm{U}^t$.
Avec $\mathrm{U}^{-1} = \mathrm{U}^t$ une matrice de passage. De plus remarquons que
$\Sigma^{-1} = (\mathrm{U}\ \Lambda\ \mathrm{U}^t)^{-1} = (\Lambda U^t)^{-1}U^{-1} = (U^t)^{-1}\,\Lambda^{-1}\,U^t = U\,\Lambda^{-1}\,U^t$.

En supposant que U est une matrice de rotation d'angle θ tel que :

$$U = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \qquad U^{-1} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$$

On remarque alors que :

$$det\Sigma = \det(U\,\Lambda\,U^{-1}) = \det(\mathrm{U})\det(\Lambda)\det(\mathrm{U}^{-1}) = \det(\Lambda) = \lambda_1\lambda_2$$

Et en posant X=x-μ_x et Y=y-μ_y, en notant λ_1 , λ_2 les valeurs propres de Σ qui existent en vue du théorème évoqué ci-dessus, on peut réécrire (E-O) :

$$K = \frac{1}{\sqrt{(2\pi)^2\lambda_1\lambda_2}} \exp\left(-\frac{1}{2}\begin{pmatrix} X & Y \end{pmatrix} U\,\Lambda^{-1}\,U^t \begin{pmatrix} X \\ Y \end{pmatrix}\right)$$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

$$\Leftrightarrow K=\frac{1}{\sqrt{(2\pi)^2\lambda_1\lambda_2}}\exp\left(-\frac{1}{2}\begin{pmatrix}X & Y\end{pmatrix}\begin{pmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{pmatrix}\begin{pmatrix}\frac{1}{\lambda_1} & 0\\ 0 & \frac{1}{\lambda_2}\end{pmatrix}\begin{pmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}\right)$$

$$\Leftrightarrow \ln(2\pi K\sqrt{\lambda_1\lambda_2}) = -\frac{1}{2}\begin{pmatrix}X & Y\end{pmatrix}\begin{pmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{pmatrix}\begin{pmatrix}\frac{1}{\lambda_1} & 0\\ 0 & \frac{1}{\lambda_2}\end{pmatrix}\begin{pmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}$$

$$\Leftrightarrow \ln\left(2\pi K\sqrt{\lambda_1\lambda_2}\right) = -\frac{1}{2}\left(X^2\left(\frac{\sin^2(\theta)}{\lambda_2}+\frac{\cos^2(\theta)}{\lambda_1}\right) + 2\,X\,Y\left(\frac{\cos(\theta)\sin(\theta)}{\lambda_1}-\frac{\cos(\theta)\sin(\theta)}{\lambda_2}\right)\right.$$
$$\left. + Y^2\left(\frac{\sin^2(\theta)}{\lambda_1}+\frac{\cos^2(\theta)}{\lambda_2}\right)\right)$$

On distingue deux identités remarquables de la forme $(a+b)^2$ et $(a-b)^2$,donc en factorisant par $\frac{1}{\lambda_1}\ et\ \frac{1}{\lambda_2}$, la formule précédente peut être écrite :

$$\Leftrightarrow \ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right) = \frac{1}{2}\left(\frac{1}{\lambda_1}\ (X\cos(\theta)+Y\sin(\theta))^2 + \frac{1}{\lambda_2}\ (X\sin(\theta)-Y\cos(\theta))^2\right.$$

$$\Leftrightarrow \frac{(X\cos(\theta)+Y\sin(\theta))^2}{2\lambda_1\ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)} + \frac{(X\sin(\theta)-Y\cos(\theta))^2}{2\lambda_2\ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)} = 1$$

$$\Leftrightarrow \frac{((x-\mu_x)\cos(\theta)+(y-\mu_y)\sin(\theta))^2}{2\lambda_1\ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)} + \frac{((y-\mu_y)\cos(\theta)-(x-\mu_x)\sin(\theta))^2}{2\lambda_2\ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)} = 1$$

On tombe sur une équation que l'on va appeler (E-1) qui est une équation d'ellipse d'après source (1) de centre (μ_x, μ_y), de demi grand axe a = $2\lambda_1\ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)$ et de demi petit axe b = $2\lambda_2\ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)$ et d'angle de rotation θ qui représente l'angle entre les axes de l'ellipse et les axes de l'ellipse s'ils étaient parallèle aux axes de plan de base orthonormée.

On obtient finalement l'équation d'ellipse :

$$\Leftrightarrow \frac{((x-\mu_x)\cos(\theta)+(y-\mu_y)\sin(\theta))^2}{a^2} + \frac{((y-\mu_y)\cos(\theta)-(x-\mu_x)\sin(\theta))^2}{b^2} = 1$$

Nous venons de montrer que les lignes d'isodensité de f_Z étaient des ellipses, dont les caractéristiques dépendent de μ, qui donne le centre de l'ellipse, et de Σ, qui donne l'orientation de cette ellipse ainsi que sa longueur.

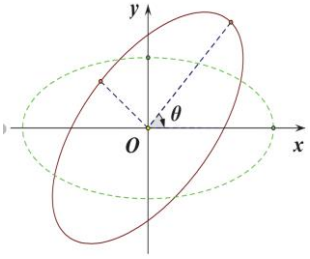


Figure 1 : Schéma rotation d'une ellipse.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

II. Calculer la probabilité qu'un point tiré selon la loi Z appartienne à la surface interne d'une ellipse d'isodensité K. En déduire l'ellipse d'isodensité K vérifiant $P(Z \in S_k)$.

Ensuite, nous cherchons à calculer la probabilité qu'un point tiré selon la loi Z appartienne à la surface intérieure S_k délimité par le contour d'isodensité K, qui est une ellipse.

Pour ce faire, on cherche à intégrer sur une surface, et calculer l'intégrale :

$$P(Z \in S_k) = \int \int_{S_k} f_Z(x,y)dxdy$$

On remplace $f_Z(x,y)$ par son expression, et on simplifie le contenu de l'exponentielle par le résultat obtenu précédemment.

$$P(Z \in S_k) = \int \int_{S_k} \frac{1}{2\pi\sqrt{\lambda_1\lambda_2}} \exp\left(-\left(\frac{\left(\cos(\theta)\,(x-\mu_1)+\sin(\theta)\,(y-\mu_2)\right)^2}{2\lambda_1}+\frac{\left(\sin(\theta)\,(x-\mu_1)-\cos(y-\mu_2)\right)^2}{2\lambda_2}\right)\right)dxdy$$

On utilise le changement de variable :

$$\begin{cases} x' = \dfrac{\cos(\theta)\,(x-\mu_x)+\sin(\theta)\,(y-\mu_y)}{\sqrt{\lambda_1}} \\ y' = \dfrac{\sin(\theta)\,(x-\mu_x)-\cos(\theta)\,(y-\mu_y)}{\sqrt{\lambda_2}} \end{cases} \Leftrightarrow \quad \begin{cases} x = \dfrac{\sqrt{\lambda_1}x' - \sin(\theta)(y-\mu_y)}{\cos(\theta)} - \mu_x \\ y = \dfrac{-\sqrt{\lambda_2}y' + \sin(\theta)(x-\mu_1)}{\cos(\theta)} + \mu_y \end{cases}$$

Et on injecte la 2ème équation dans la 1ère pour obtenir :

$$\begin{cases} x = \dfrac{\sqrt{\lambda_1}x'}{\cos(\theta)} + \dfrac{\tan(\theta)}{\cos(\theta)}\sqrt{\lambda_2}y' - \tan^2(\theta)(x-\mu_x) + \mu_x \\ y = \dfrac{-\sqrt{\lambda_2}y' + \sin(\theta)(x-\mu_x)}{\cos(\theta)} + \mu_y \end{cases}$$

$$\Leftrightarrow \begin{cases} x(1+\tan^2(\theta)) = \dfrac{\sqrt{\lambda_1}x'}{\cos(\theta)} + \dfrac{\tan(\theta)}{\cos(\theta)}\sqrt{\lambda_2}y' + \tan^2(\theta)\mu_x + \mu_x \\ y = \dfrac{-\sqrt{\lambda_2}y' + \sin(\theta)(x-\mu_x)}{\cos(\theta)} + \mu_y \end{cases} \qquad avec\ (1+\tan^2(\theta)) = \cos\text{^}2\ (\theta)$$

$$\Leftrightarrow \begin{cases} x = \dfrac{\sqrt{\lambda_1}x'}{\cos(\theta)}\cos^2(\theta) + \dfrac{\tan(\theta)}{\cos(\theta)}\cos^2(\theta)\sqrt{\lambda_2}y' + \tan^2(\theta)\cos^2(\theta)\mu_x + \cos^2(\theta)\mu_x \\ y = \dfrac{-\sqrt{\lambda_2}y' + \sin(\theta)(x-\mu_x)}{\cos(\theta)} + \mu_y \end{cases}$$

$$\Leftrightarrow \begin{cases} x = \sqrt{\lambda_1}\cos(\theta)\,x' + \sin(\theta)\sqrt{\lambda_2}y' + sin^2(\theta)\mu_x + \cos^2(\theta)\,\mu_x \\ y = \dfrac{-\sqrt{\lambda_2}y' + \sin(\theta)(x-\mu_x)}{\cos(\theta)} + \mu_y \end{cases} \Leftrightarrow \begin{cases} x = \sqrt{\lambda_1}\cos(\theta)\,x' + \sin(\theta)\sqrt{\lambda_2}y' + \mu_x \\ y = \dfrac{-\sqrt{\lambda_2}y' + \sin(\theta)(x-\mu_x)}{\cos(\theta)} + \mu_y \end{cases}$$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

On fait de même pour trouver y selon x' et y', et on obtient finalement

$$\begin{cases} x = \sqrt{\lambda_1}\cos(\theta)\,x' + \sqrt{\lambda_2}\sin(\theta)\,y' + \mu_x \\ y = \sqrt{\lambda_1}\sin(\theta)\,x' - \sqrt{\lambda_2}\cos(\theta)\,y' + \mu_y \end{cases}$$

Soit l'application φ :

$$\varphi : \mathbb{R}^2 \to \mathbb{R}^2$$
$$(x,y) \mapsto \left(\sqrt{\lambda_1}\cos(\theta)\,x' + \sqrt{\lambda_2}\sin(\theta)\,y' + \mu_y,\ \sqrt{\lambda_1}\sin(\theta)\,x' - \sqrt{\lambda_2}\cos(\theta)\,y' + \mu_y\right)$$

Montrer que φ est bijective.

On effectue les dérivées partielles :

$$\frac{\partial x}{\partial x'} = \sqrt{\lambda_1}\cos(\theta), \quad \frac{\partial x}{\partial y'} = \sqrt{\lambda_2}\sin(\theta)$$

$$\frac{\partial y}{\partial x'} = \sqrt{\lambda_1}\sin(\theta), \quad \frac{\partial y}{\partial x'} = -\sqrt{\lambda_2}\cos(\theta)$$

Donc φ est $\mathbb{C}^1$.

On obtient ensuite la matrice Jacobienne suivante :

$$J = \begin{pmatrix} \sqrt{\lambda_1}\cos(\theta) & \sqrt{\lambda_2}\sin(\theta) \\ \sqrt{\lambda_1}\sin(\theta) & -\sqrt{\lambda_2}\cos(\theta) \end{pmatrix}$$

Ainsi que le Jacobien : $\det(J) = -\sqrt{\lambda_1\lambda_2} \neq 0$ donc φ^{-1} est $\mathbb{C}^1$.

φ est donc bijective, $\mathbb{C}^1$, et φ^{-1} est $\mathbb{C}^1$, on a montré que φ est un $\mathbb{C}^1$ difféomorphisme, on peut donc appliquer le théorème de changement de variable sur une intégrale double :

$$P(Z \in S_k) = \frac{1}{2\pi\sqrt{\lambda_1\lambda_2}} \int\int_{S_k} \exp\left(-\frac{x'^2 + y'^2}{2}\right) \left|-\sqrt{\lambda_1\lambda_2}\right| dx'dy'$$

On effectue un autre changement de variable pour passer aux coordonnées polaires : $\begin{cases} x' = r\cos(\theta) \\ y' = r\sin(\theta) \end{cases}$

On effectue de nouveau les dérivées partielles :

$$\frac{\partial x'}{\partial r} = \cos(\theta), \quad \frac{\partial x'}{\partial \theta} = -r\sin(\theta)\ \frac{\partial y'}{\partial r} = \sin(\theta), \quad \frac{\partial y'}{\partial \theta} = \mathrm{r}\cos(\theta)$$

Et on obtient la matrice jacobienne suivante : $J' = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & \mathrm{r}\cos(\theta) \end{pmatrix}$

Ainsi que le jacobien $\det(J') = r \neq 0$

L'application linéaire : $\varphi' : R * [0\,;2\pi] \longrightarrow \mathbb{R}^2$
$(r,\theta) \mapsto (r\cos(\theta), r\sin(\theta))$

Est bijective, $\mathbb{C}^1$, et φ'^{-1} est $\mathbb{C}^1$. Ainsi, φ' est un $\mathbb{C}^1$ difféomorphisme, et on peut appliquer le théorème de changement de variable, on obtient :

$$P(Z \in S_k) = \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\mathrm{R}} \exp\left(-\frac{(r\cos(\theta))^2 + (r\sin(\theta))^2}{2}\right)|r|drd\theta$$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Soit en simplifiant : $P(Z \in S_k) = \frac{1}{2\pi}\int_0^{2\pi}\left[-\exp\left(\frac{-r^2}{2}\right)\right]_0^R d\theta$

On avait : $\frac{(\ (x-\mu_x)\cos(\theta)+(y-\mu_y)\sin(\theta)\ )^2}{2\lambda_1} + \frac{(\ (y-\mu_y)\cos(\theta)-(x-\mu_x)\sin(\theta)\ )^2}{2\lambda_2} = -\ln\left(2\pi K\sqrt{\lambda_1\lambda_2}\ \right)$

Et en utilisant nos 2 changements de variables comme précédemment, on a :

$$\frac{r^2}{2} = -\ln\left(2\pi K\sqrt{\lambda_1\lambda_2}\ \right) \Leftrightarrow r = \sqrt{-2\ln\left(2\pi K\sqrt{\lambda_1\lambda_2}\ \right)}$$

Et comme on intègre sur la surface d'un cercle, on a $R = \sqrt{-2\ln(2\pi K\sqrt{\lambda_1\lambda_2})}$.

On obtient alors : $P(Z \in S_k) = \frac{1}{2\pi}\int_0^{2\pi}\left(-\exp\left(\frac{-\sqrt{-2\ln\left(2\pi K\sqrt{\lambda_1\lambda_2}\right)}^2}{2}\right) - (-1)\right)d\theta$

Et finalement, nous avons en intégrant sur θ la probabilité qu'un point tiré selon la loi Z appartienne à la surface interne délimitée par l'isodensité K :

$$P(Z \in S_k) = 1 - 2\pi K\sqrt{\lambda_1\lambda_2}$$

On cherche maintenant à déterminer l'ellipse d'isodensité K vérifiant $P(Z \in S_k) = p$, avec $p \in [0,1]$.

Soit : $P(Z \in S_k) = p \Leftrightarrow \left(1 - 2\pi K\sqrt{\lambda_1\lambda_2}\right) = p \Leftrightarrow -2\pi K\sqrt{\lambda_1\lambda_2} = p - 1 \Leftrightarrow K = \frac{1-p}{2\pi\sqrt{\lambda_1\lambda_2}}$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

III. Déterminer un estimateur de Σ et de μ

Ceci fait, attardons nous désormais sur la détermination d'estimateurs des paramètres de Z dont nous justifierons plus tard la pertinence en confrontant nos résultats avec simulations numériques.

Estimateur de μ: $\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$: moyenne empirique ou expérimentale.

Estimateur de $\sum$: $\bar{\bar{\Sigma}} = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{u}_X)^2 & \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{u}_X)(Y_i - \bar{u}_Y) \\ \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{u}_X)(Y_i - \bar{u}_Y) & \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{u}_Y)^2 \end{pmatrix}$ matrice de covariance expérimentale .

<u>Détermination d'estimateurs par la méthode du maximum de vraisemblance.</u>

On souhaite estimer les paramètres $\theta = \mu$ et $\Theta = \sum$ d'une loi normale bidimensionnelle Z. Soit Z = $(Z_1, Z_2, \ldots, Z_N)$ un échantillon de la loi jointe $f_Z(m;\bar{\theta})$ où $\bar{\theta}$ est un couple (θ, Θ) est un couple de paramètre à estimer. Soit m = $(m_1, m_2, \ldots, m_N)$ une réalisation de cet échantillon. Notons $L(\bar{\theta} \mid m) = f_Z(m; \bar{\theta})$ la fonction de vraisemblance de $\bar{\theta}$. Ainsi sa fonction log vraisemblance associée s'écrit :

$$\log L(\bar{\theta} \mid m) = \sum_{i=1}^{N} \log f_{m_i}(m_i; \bar{\theta})$$

Donc : $\log L(\bar{\theta} \mid m) = \sum_{i=1}^{N} \log \frac{1}{\sqrt{(2\pi)^2 \det\Theta}} \exp\left(-\frac{1}{2}\, {}^t(m_i - \theta)\, \Theta^{-1}(m_i - \theta)\right)$

(Notons que le vecteur θ s'écrit $\binom{\theta_{11}}{\theta_{21}}$, *le vecteur* m_i *s'écrit* $\binom{X_i}{Y_i}$, et que la matrice Θ s'écrit $\begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix}$.)

$$= \cdots = -N\log(2\pi) - \frac{N}{2}\log((\det(\det\Theta)) - \frac{1}{2}\sum_{i=1}^{N} \Theta_{11}(X_i - \theta_{11})^2 - \frac{1}{2}\sum_{i=1}^{N} (\Theta_{21+}\Theta_{12})(X_i - \theta_{11})(Y_i - \theta_{21}) - \frac{1}{2}\sum_{i=1}^{N} \Theta_{22}(Y_i - \theta_{21})^2$$

Cherchons maintenant les maximums de la fonction de log vraisemblance en pour chaque paramètre et θ_{ij} .

Avant de poursuivre **<u>admettons</u>** quelques théorèmes de dérivation matricielle :

Proposition 1 : Soit M ∈ $\mathbb{R}^{k\times k}$ alors,

$$\frac{\partial(\log(\det(M)))}{\partial M} = M^{-1}$$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Proposition 2 : Soit X ∈ $\mathbb{R}^{k\times k}$ et inversible et a une forme scalaire alors,

$$\frac{\partial((a^t)(X^{-1})a)}{\partial X} = -X^{-1}aa^t\,X^{-1}$$

Définition 1 : Soit un scalaire a, fonctionnellement dépendant des variables x_1,……., x_p.C'est à dire a=h (x_1,……., x_p). Si x = (x_1,……., x_p)' est le vecteur de ces variables alors,

$$\frac{\partial\, a}{\partial x} = \begin{pmatrix} \frac{\partial\, a}{\partial x_1} \\ \cdots \\ \frac{\partial\, a}{\partial x_n} \end{pmatrix} \quad et \quad \frac{\partial\, a}{\partial x'} = \left(\frac{\partial\, a}{\partial x_1}, \ldots, \frac{\partial\, a}{\partial x_n}\right) = \left(\frac{\partial\, a}{\partial x_n}\right)'$$

Donc :

$$\frac{\partial\ \log L(\overline{\theta}\,|\,m)}{\partial\Theta} = \frac{\partial}{\partial\Theta}\left(-N\log(2\pi) - \frac{N}{2}\log(\det\Theta) - \frac{1}{2}\sum_{i=1}^{N}\ {}^t(m_i-\theta)\,\Theta^{-1}(m_i-\theta)\right)$$

$$= -\frac{N}{2}\,\Theta^{-1} + \frac{1}{2}\sum_{i=1}^{N}\Theta^{-1}\,(m_i-\theta)\,{}^t(m_i-\theta)\,\Theta^{-1}$$

$$= -\frac{N}{2(\Theta_{11}\Theta_{22}-\Theta_{12}\Theta_{21})}\begin{pmatrix}\Theta_{22} & -\Theta_{12}\\ -\Theta_{21} & \Theta_{11}\end{pmatrix} + \frac{1}{2}\sum_{i=1}^{N}\frac{1}{(\Theta_{11}\Theta_{22}-\Theta_{12}\Theta_{21})^2}$$

$$\begin{pmatrix} {\Theta_{22}}^2(X_i-\theta_{11})^2 - (\Theta_{21}\Theta_{22}+\Theta_{12}\Theta_{22})\,(X_i-\theta_{11})(Y_i-\theta_{21}) + \Theta_{21}\Theta_{12}\,(Y_i-\theta_{21})^2 \\ -\Theta_{22}\Theta_{12}(X_i-\theta_{11})^2 + (\Theta_{11}\Theta_{22}+{\Theta_{12}}^2)\,(X_i-\theta_{11})(Y_i-\theta_{21}) - \Theta_{11}\Theta_{12}\,(Y_i-\theta_{21})^2 \\ -\Theta_{21}\Theta_{22}(X_i-\theta_{11})^2 + ({\Theta_{21}}^2+\Theta_{11}\Theta_{22})\,(X_i-\theta_{11})(Y_i-\theta_{21}) - \Theta_{11}\Theta_{21}\,(Y_i-\theta_{21})^2 \\ \Theta_{21}\Theta_{12}(X_i-\theta_{11})^2 - (\Theta_{21}\Theta_{11}+\Theta_{11}\Theta_{12})\,(X_i-\theta_{11})(Y_i-\theta_{21}) + {\Theta_{11}}^2(Y_i-\theta_{11})^2 \end{pmatrix}$$

$$= -\frac{N}{2(\Theta_{11}\Theta_{22}-\Theta_{12}\Theta_{21})}\begin{pmatrix}\Theta_{22} & -\Theta_{12}\\ -\Theta_{21} & \Theta_{11}\end{pmatrix} + \frac{1}{2}\frac{1}{(\Theta_{11}\Theta_{22}-\Theta_{12}\Theta_{21})^2}$$

$$\begin{pmatrix} {\Theta_{22}}^2\sum_{i=1}^{N}(X_i-\theta_{11})^2 - (\Theta_{21}\Theta_{22}+\Theta_{12}\Theta_{22})\sum_{i=1}^{N}(X_i-\theta_{11})(Y_i-\theta_{21}) + \Theta_{21}\Theta_{12}\sum_{i=1}^{N}(Y_i-\theta_{21})^2 \\ -\Theta_{22}\Theta_{12}\sum_{i=1}^{N}(X_i-\theta_{11})^2 + (\Theta_{11}\Theta_{22}+{\Theta_{12}}^2)\sum_{i=1}^{N}(X_i-\theta_{11})(Y_i-\theta_{21}) - \Theta_{11}\Theta_{12}\sum_{i=1}^{N}(Y_i-\theta_{21})^2 \end{pmatrix}$$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

$$-\Theta_{21}\Theta_{22}\sum_{i=1}^{N}(X_i-\theta_{11})^2+({\Theta_{21}}^2+\Theta_{11}\Theta_{22})\sum_{i=1}^{N}(X_i-\theta_{11})(Y_i-\theta_{21})-\Theta_{11}\Theta_{21}\sum_{i=1}^{N}(Y_i-\theta_{21})^2$$
$$\Theta_{21}\Theta_{12}\sum_{i=1}^{N}(X_i-\theta_{11})^2-(\Theta_{21}\Theta_{11}+\Theta_{11}\Theta_{12})\sum_{i=1}^{N}(X_i-\theta_{11})(Y_i-\theta_{21})+{\Theta_{11}}^2\sum_{i=1}^{N}(Y_i-\theta_{11})^2\Bigg)$$

$$\log \mathrm{L}(\bar{\theta} \mid \mathrm{m}) = -N\log(2\pi)-\frac{N}{2}\log((\det(\det\Theta))-\frac{1}{2}\sum_{i=1}^{N}\Theta_{11}(X_i-\theta_{11})^2$$
$$-\frac{1}{2}\sum_{i=1}^{N}(\Theta_{21+}\Theta_{12})(X_i-\theta_{11})(Y_i-\theta_{21})-\frac{1}{2}\sum_{i=1}^{N}\Theta_{22}(Y_i-\theta_{21})^2$$

C'est une fonction concave de θ_{11} et θ_{21} (car somme de fonctions concaves), de

Dérivée :

$$\frac{\partial \log \mathrm{L}(\bar{\theta}\mid \mathrm{m})}{\partial\theta}=\Big(\sum_{i=1}^{N}\Theta_{11}(X_i-\theta_{11})+\frac{1}{2}\sum_{i=1}^{N}(\Theta_{21+}\Theta_{12})(Y_i-\theta_{21}),\sum_{i=1}^{N}\Theta_{22}(Y_i-\theta_{21})+\frac{1}{2}\sum_{i=1}^{N}(\Theta_{21+}\Theta_{12})(X_i-\theta_{11})\Big)$$

Elle atteint donc son maximum lorsque $\frac{\partial \log \mathrm{L}(\bar{\theta}\mid \mathrm{m})}{\partial\theta}=(0,0)$ c'est-à-dire lorsque

$$\Bigg(\theta_{11}=\frac{1}{\Theta_{11}N}\sum_{i=1}^{N}X_i-\frac{1}{2\Theta_{11}N}\sum_{i=1}^{N}(\Theta_{21+}\Theta_{12})(Y_i-\theta_{21}),\theta_{21}$$
$$=\frac{1}{\Theta_{22}N}\sum_{i=1}^{N}Y_i-\frac{1}{2\Theta_{22}N}\sum_{i=1}^{N}(\Theta_{21+}\Theta_{12})(X_i-\theta_{11})\Bigg)$$

Malgré l'obtention de quelques résultats, nous n'avons pas pu déterminer de formule exploitable.

Avant de poursuivre rappelons quelles sont les qualités qui font qu'un estimateur est pertinent.

1.Son biais : Défini par b ($\hat{\Theta}_N$)= E [$\hat{\Theta}_N$]- θ.

Un estimateur est dit sans biais lorsque :

Soit (X_1, . . .,X_N) un échantillon. On s'intéresse au paramètre θ dont un estimateur est $\hat{\Theta}_N$.
L'estimateur ΘN est dit sans biais, ou non biaisé, lorsque E ($\hat{\Theta}_N$) = θ.

2.Sa consistance :

Soit (X_1, . . .,X_N) un échantillon. On s'intéresse au paramètre θ dont un estimateur est $\hat{\Theta}_N$.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

L'estimateur $\hat{\Theta}_N$ est dit fortement consistant ou convergeant si P ($\lim_{N\to+\infty} \hat{\Theta}_N = \theta$) = 1. On dit que $\hat{\Theta}_N$ tend vers θ presque sûrement lorsque N est grand. Ce qui en d'autres termes veut dire que plus l'échantillon est grand plus la qualité de l'estimateur augmente.

Dans le cas de la moyenne empirique :

Ainsi on constate directement que la moyenne empirique est un estimateur ponctuel sans biais de μ car on sait que dans le cas d'une loi normale on a E (X_i)=μ.

$$\text{E}(\bar{x}) = \text{E}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu .$$

De plus cet estimateur est fortement consistant, c'est un résultat de la loi des grands nombres.

Dans le cas de la matrice des variances et covariance empiriques :

Son biais :

D'après le cours les statistiques de cette matrice sont des estimateurs biaisés dont on connaît une variante débiaisée :

$$\bar{\bar{\Sigma}}_* = \begin{pmatrix} \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{u}_X)^2 & \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{u}_X)(Y_i - \bar{u}_Y) \\ \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{u}_X)(Y_i - \bar{u}_Y) & \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{u}_Y)^2 \end{pmatrix}$$ ($\bar{\bar{\Sigma}}_*$ sera noté $\bar{\bar{\Sigma}}$ dans la suite.)

De plus les statistiques contenues dans cette matrice sont fortement convergentes car lorsque N est grand on retrouve la matrice de covariance tel qu'elle est définie

$$\lim_{N\to+\infty} \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y) = COV(X,Y)$$

$$\lim_{N\to+\infty} \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \hat{\mu}_Y)^2 = Var(Y)$$

$$\lim_{N\to+\infty} \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu}_X)^2 = \text{Var(X)}$$

Plus tard nous calculerons la valeur de ces statistiques pour des échantillon tiré de façon aléatoire selon la loi Z, pour illustrer nos propos.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

III.1. Bonus : que se passe-t-il si une seule valeur propre de Σ est nulle ?

Soit la densité de probabilité de la loi normale bidimensionnelle :

$$f_Z(m) = \frac{1}{\sqrt{(2\pi)^2 \det \sum}} \exp\left(-\frac{1}{2}{}^t(m-\mu)\sum^{-1}(m-\mu)\right) \quad \forall\, m \in \mathbb{R}^2$$

Si une des valeurs propres de Σ est nulle, d'une part, Σ n'est plus inversible, et de ce fait $\det \Sigma = 0$.

L'expression de la densité de probabilité de Z n'est donc plus valide si une des deux valeurs propres est nulle, dans la mesure où l'on divise par 0 d'un côté, et de l'autre côté Σ n'est pas inversible.

Néanmoins, au vu de l'équation d'ellipse que nous avions trouvée précédemment comme ci :

$$\frac{\left(\ (x-\mu_x)\cos(\theta) + (y-\mu_y)\sin(\theta)\ \right)^2}{2\lambda_1 \ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)} + \frac{\left(\ (y-\mu_y)\cos(\theta) - (x-\mu_x)\sin(\theta)\ \right)^2}{2\lambda_2 \ln\left(\frac{1}{2\pi K\sqrt{\lambda_1\lambda_2}}\right)} = 1$$

On peut prédire la forme du contour d'isodensité de la loi. En effet, pour une valeur propre nulle, soit le demi-grand axe, soit le demi-petit axe devient nul, l'ellipse s'aplatit donc, et se réduit à une ligne.

Pour être plus précis, on va étudier la matrice de covariance ainsi que ses valeurs propres et vecteurs propres associés.

Les vecteurs propres d'une matrice de covariance représentent les directions dans lesquelles les données varient le plus, et sa valeur propre associée indique son amplitude. De plus, la plus grande valeur propre est associée au vecteur propre qui est dirigée vers là où les données s'étendent le plus. Pour une matrice de covariance symétrique, les vecteurs propres seront orthogonaux.

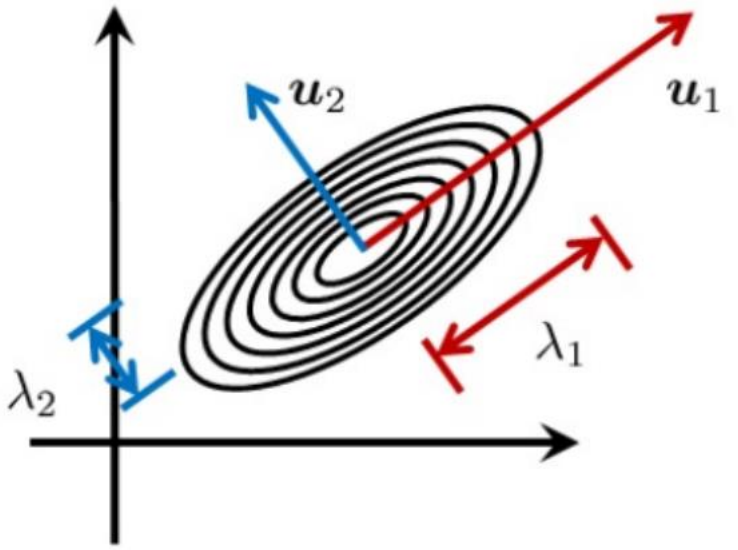


Figure 2, source (8) – Interprétation géométrique de la décomposition spectrale de Σ – Orientation d'une densité de probabilité d'une variable aléatoire à densité

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Ainsi, pour la loi normale bidimensionnelle par exemple, si on tire un échantillon de points aléatoire, la plus grande valeur propre aura son vecteur propre qui va suivre la plus grande dispersion de données selon une certaine amplitude, et son autre valeur propre aura son vecteur propre orthogonal au premier, qui va suivre la seconde plus grande dispersion de points selon une autre amplitude. On distingue ainsi une forme elliptique de la dispersion de données, d'où le but de la question 1.a du développement mathématique, qui demande de montrer que les lignes d'isodensité de f_z sont des ellipses.

Les ellipses d'isodensité dépendent donc directement de la matrice de covariance Σ qui va donner la direction et l'amplitude de l'ellipse, et de μ qui va donner le centre de cette ellipse. Le demi-grand axe de cette ellipse dépend de la plus grande valeur propre et son vecteur propre associée, et le demi-petit axe dépend de la seconde valeur propre et de son vecteur propre associée.

Si une de ces valeurs propres venait à être nulle, la distribution de données ne suivrait alors plus qu'une seule direction, et formerait une ligne, d'où l'aplatissement de l'ellipse d'isodensité K, selon le grand ou petit axe.

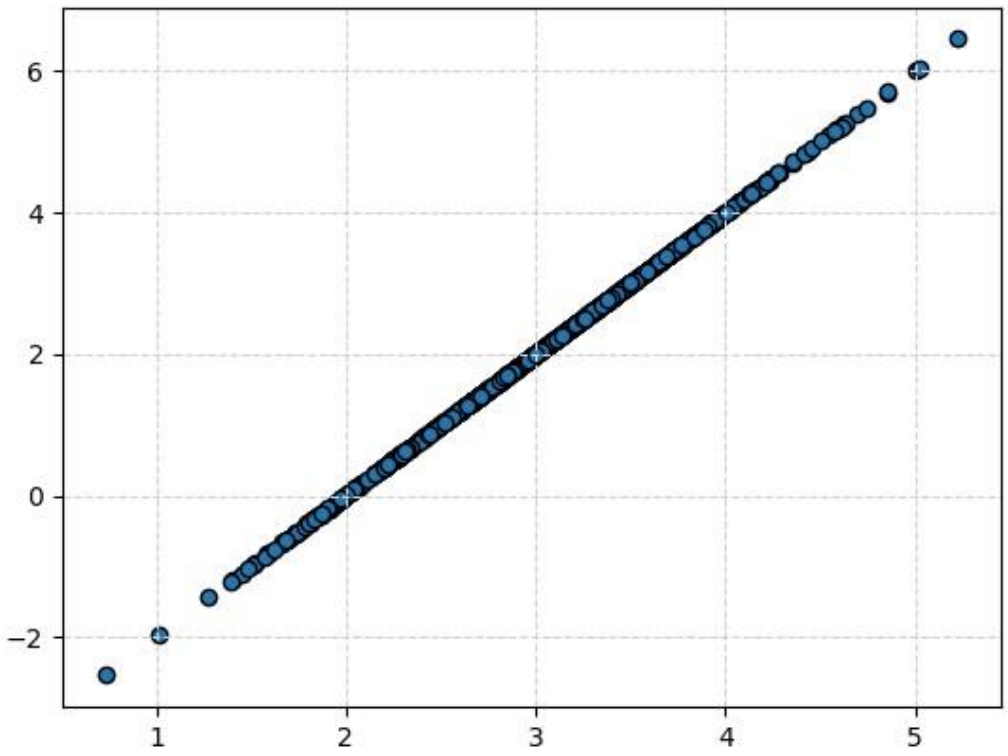


Figure 3 – Simulation d'une distribution de données selon la loi normale bidimensionnelle avec une valeur propre nulle.

Pour expliciter les explications précédentes, nous avons numérisé notre distribution de points avec une valeur propre nulle.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Pour ce faire, on utilise le même code python utilisé dans l'étude numérique pour la question 1.a de l'étude mathématiques, sauf qu'ici, nous allons changer la matrice de covariance. On utilise la fonction python stats.multivariate_normal.rvs pour obtenir un échantillon de données.
Dans notre cas, nous avons choisi comme matrice de covariance :

$$\Sigma = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$$

Cette matrice à pour valeurs propres $\lambda_1 = 1$, $\lambda_2 = 0$. Et on observe en effet une distribution suivant une ligne. Elle est ainsi car la plus grande valeur propre, soit λ_1, à son vecteur propre dirigé selon la plus grande étendue des points distribués, et le 2nd vecteur propre est nul.

Néanmoins, comme indiqué précédemment, la densité de probabilité de la loi normale bidimensionnelle n'est plus valable car on divise par 0, et on ne peut afficher les contours d'isodensité de cette densité de probabilité. Mais au vu de nos hypothèses on pourrait supposer que le « contour » de cette densité de probabilité est un contour de « dimension 1 » c'est-à-dire un segment dont la longueur dépend de λ_1.

III.2. Bonus : que se passe-t-il si les deux valeurs propres de Σ est nulle ?

De la même manière que si une valeur propre était nulle, l'expression de la densité de probabilité de la loi normale bidimensionnelle n'est plus valable.

Dans le cas où les 2 valeurs propres sont nulles, les deux vecteurs propres de la matrice de covariance sont aussi nuls. Ainsi la distribution de données ne suivrait aucune direction, et serait donc un point, au centre de l'ellipse indiqué par μ. Le contour d'isodensité de la densité de probabilité de la loi normale bidimensionnelle se réduirait aussi à un point (« dimension 0 »).

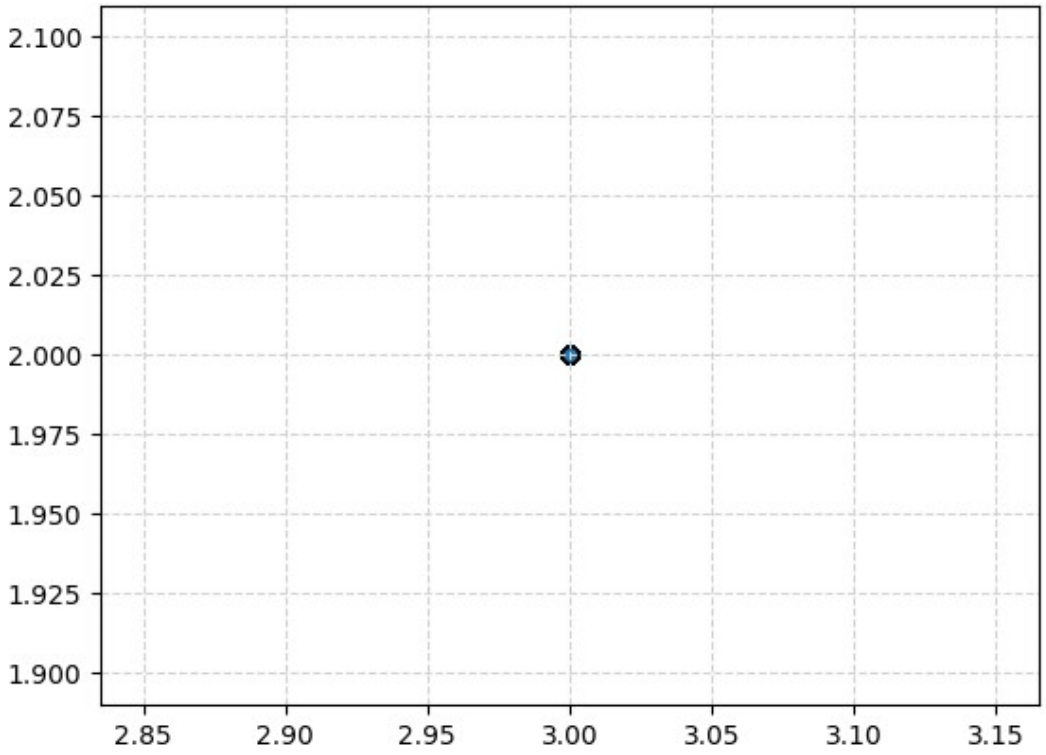


Figure 4 – Simulation d'une distribution de données selon la loi normale bidimensionnelle avec deux valeurs propres nulles.

Dans ce cas, nous avons comme matrice de covariance :

$$\Sigma = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

Cette matrice à pour valeurs propres $\lambda_1 = 0$, $\lambda_2 = 0$, et on observe effectivement une distribution des données sur un point seulement, centré en μ.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Étude Numérique

I. Représenter un échantillon de points tirés selon la loi Z ainsi que les ellipses d'isodensité associées à des probabilités p au choix.

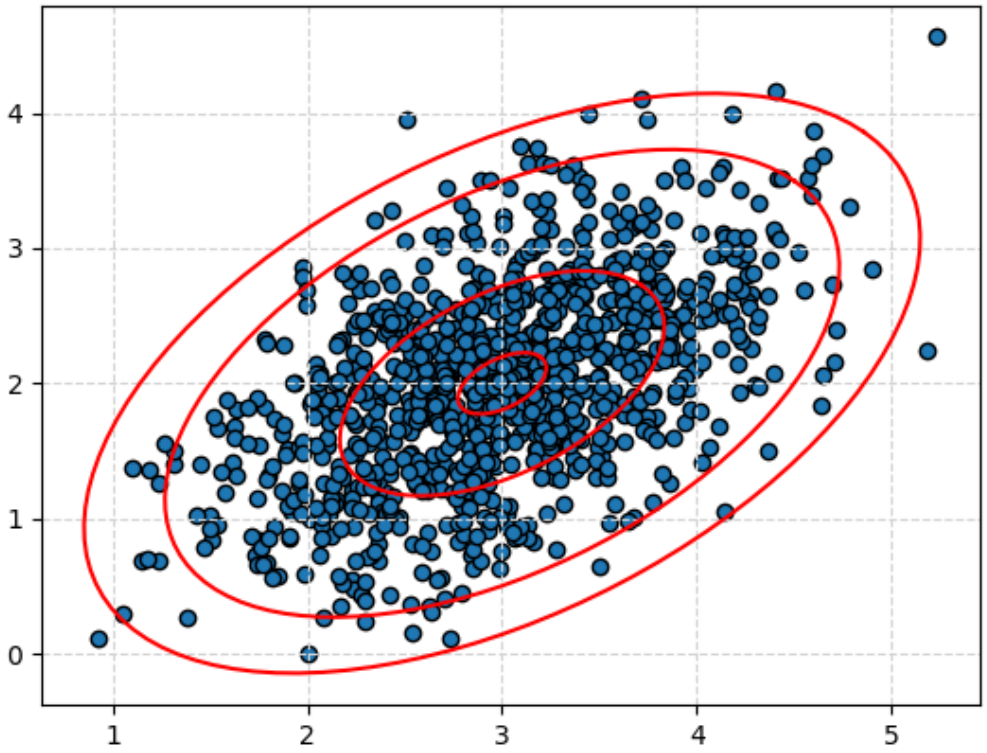


Figure 1 – Simulation d'un tirage de 1000 points selon la loi normale bidimensionnelle et tracé de différents contours gaussiens de probabilité respectives de l'extérieur vers l'intérieur 0.99, 0.95, 0.50 et 0.05

Dans cette figure, nous observons un échantillon de 1000 points tirés selon la loi normale bidimensionnelle avec comme paramètres une matrice μ et Σ donnée. On trace ensuite les différentes ellipses d'isodensité K grâce à notre équation d'ellipse trouvée précédemment, qu'on met sous forme d'équation paramétrique, avec la même matrice μ et Σ. Les différentes ellipses sont tracées en fonction de la probabilité que l'on veuille qu'un point se trouve dans l'ellipse.

Cela montre l'efficacité des estimateurs choisis pour la matrice de covariance Σ et μ.

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Code Python donnant le graphique précédent.

```
import scipy.stats as stats
import numpy as np
from math import pi
from matplotlib import pyplot as plt

sigma=np.array([[0.5,0.25] ,[0.25,0.5]]) #on choisit la matrice de covariance
mu=np.array([3 ,2]) #on choisit le centre de l'ellipse
z = stats.multivariate_normal.rvs(mu,sigma,1000) #on génère les points
aléatoires selon la loi normale bidimensionnelle

lambdas,vecs=np.linalg.eig(sigma) #on calcule les valeurs propres et les
vecteurs propres de la matrice de covariance
lambda1=lambdas[0]
lambda2=lambdas[1]
theta=np.arccos(vecs[0][0]) #on calcule l'angle de rotation de l'ellipse

K_99=(0.01)/(2*pi*np.sqrt(lambda1*lambda2)) #on calcule les valeurs de K pour
différentes probabilités
K_95=(0.05)/(2*pi*np.sqrt(lambda1*lambda2))
K_50=(0.5)/(2*pi*np.sqrt(lambda1*lambda2))
K_05=(0.95)/(2*pi*np.sqrt(lambda1*lambda2))
K=[K_99,K_95,K_50,K_05]
for i in range(4) :
    a=np.sqrt((2*lambda1*np.log((1/(2*pi*K[i]*np.sqrt(lambda1*lambda2))))))
#on calcule les demi-axes de l'ellipse
    b=np.sqrt((2*lambda2*np.log((1/(2*pi*K[i]*np.sqrt(lambda1*lambda2))))))
    t=np.linspace(0,2*pi,1000) #on crée un vecteur de 1000 points entre 0 et
2pi
    plt.plot(mu[0]+a*np.cos(theta)*np.cos(t)-b*np.sin(theta)*np.sin(t),
             mu[1]+a*np.sin(theta)*np.cos(t)+b*np.cos(theta)*np.sin(t),"red")

plt.grid(color='lightgray',linestyle='--')
plt.scatter (z [: ,0] , z [: ,1] , edgecolors ="black") #on affiche les points
aléatoires
plt.show()
```


COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

II. Calculer la valeur des estimateurs de Σ et de μ sur des échantillons de taille variable et illustrer leur convergence.

Nous allons maintenant générer des échantillons aléatoires selon la loi normale bidimensionnelle et vérifier si nos estimateurs permettent de déterminer les paramètres de la loi.

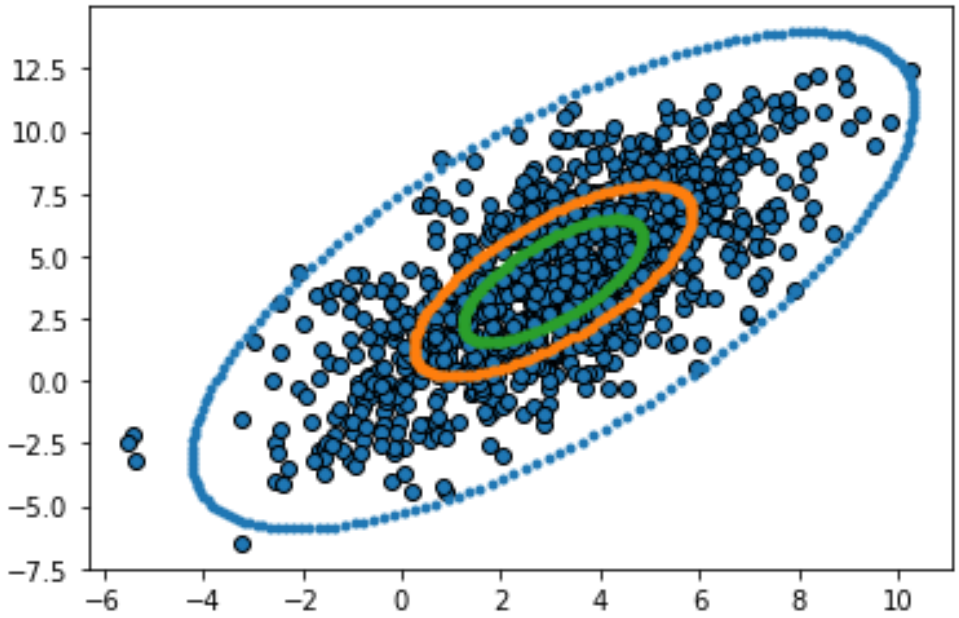


Figure A. Échantillon aléatoire de 1000 points.

$\mu = \begin{pmatrix} 3.035 \\ 4.018 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 5.743 & 5.551 \\ 5.551 & 10.73 \end{pmatrix}$

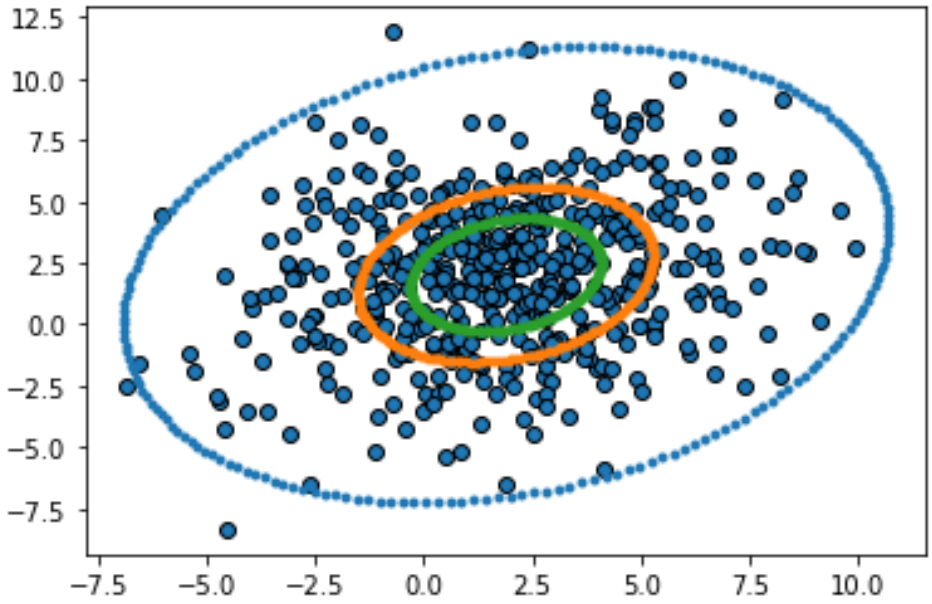


Figure B. Échantillon aléatoire de 500 points

$\mu = \begin{pmatrix} 1.885 \\ 2.024 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 8.454 & 1.910 \\ 1.910 & 9.307 \end{pmatrix}$

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

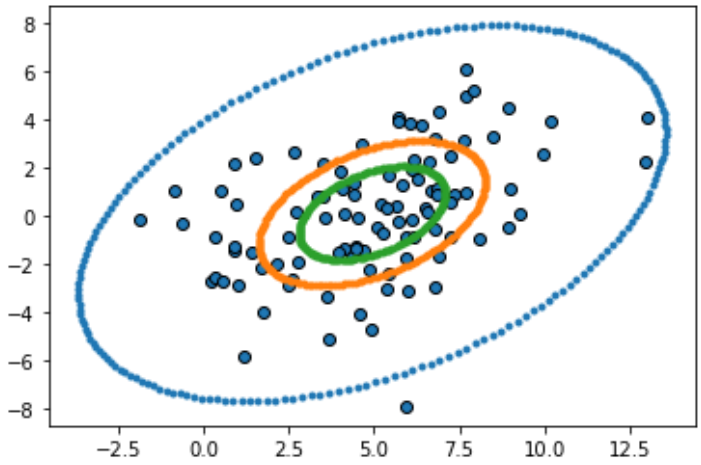


Figure C. Échantillon aléatoire de 100 points

$\mu = \begin{pmatrix} 4.935 \\ 0.113 \end{pmatrix}$ $\quad \sum = \begin{pmatrix} 8.102 & 3.140 \\ 3.140 & 6.646 \end{pmatrix}$

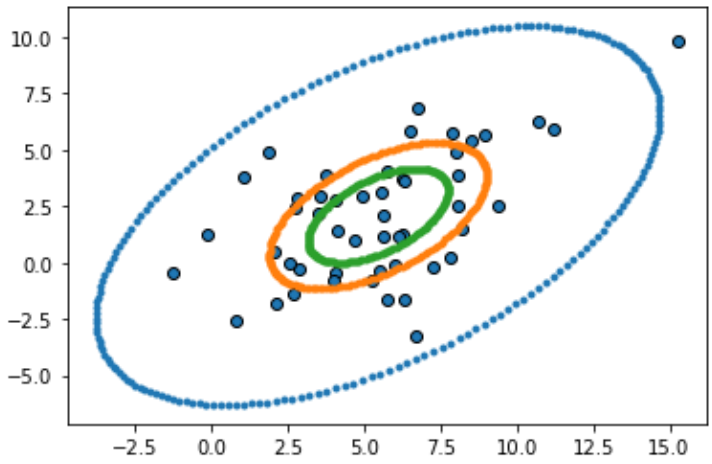


Figure D. Échantillon aléatoire de 50 points

$\mu = \begin{pmatrix} 5.448 \\ 2.089 \end{pmatrix}$ $\quad \sum = \begin{pmatrix} 9.258 & 4.548 \\ 4.548 & 7.641 \end{pmatrix}$

Dans chaque simulation, l'ellipse bleue est sensée contenir 99% des points, celle de couleur orange environ 50% et celle de couleur verte environ 25%. Ce qui est en effet le cas, de plus le résultant est même quasiment indiscutable visuellement. On peut en conclure que les statistiques choisis pour μ et ∑ sont pertinents pour déterminer les paramètres d'une loi normale bidimensionnelle.

On constate de plus que plus le nombre de point est élevé, plus le résultat est précis, ce qui correspond à ce qu'on avait prédit théoriquement.

Code Python des tests ci-dessus.

```
def estimateur_1_echantillon_aléatoire(nb_points,proba):
    def test_aleatoire(nb_points):
        #génère un écghantillonaléatoire pour une population de nb_points
entier
        absc=randint(0,5);ordo=randint(0,5)
        sig11=randint(1,10);sig22=randint(1,10)
        sig12=sigma11=randint(0,10)
        z=stats.multivariate_normal.rvs([absc ,ordo],
                            [[sig11,sig12],[sig12,sig22]],nb_points)
        plt.scatter(z[:,0],z[:,1] , edgecolors ="black")
        return(z)
    def printellipse(K,lambda1,lambda2,X0,Y0):
        a=np.sqrt(np.abs(2*lambda1*np.log(np.abs(2*pi*K*np.sqrt(lambda1*lambda
2)))))
        b=np.sqrt(np.abs(2*lambda2*np.log(np.abs(2*pi*K*np.sqrt(lambda1*lambda
2)))))
        theta=np.arccos(vecs[0][0])
        t = np.linspace(0,2*pi,200)
        plt.plot( mu[0]+a*np.cos(theta)*np.cos(t)-b*np.sin(theta)*np.sin(t) ,
                 mu[1]+a*np.sin(theta)*np.cos(t)+b*np.cos(theta)*np.sin(t),"."
)

    #génération d'un échantillon aléatoire
    z=test_aleatoire(nb_points)
    N=len(z[:,0])

    #construction de l'estimateur du centre de l'échantillon
    S=sum(z);  X=S[0]/N ;  Y=S[1]/N;  mu=[X,Y]

    #construction de l'estimateur de la matrice de covarience
    sig11=0; sig12=0; sig22=0
    for i in range(N):
        sig11+= (z[i,0]-X)**2
        sig22+= (z[i,1]-Y)**2
        sig12+= (z[i,0]-X)*(z[i,1]-Y)
    sig11=sig11/(N-1); sig22=sig22/(N-1); sig12=sig12/(N-1)
    sigma=[[sig11,sig12] , [sig12,sig22]]

    #parametres de léchantillon solon en
    #vue de modéliser une loi normale bidimentionnelle
    lambdas,vecs=np.linalg.eig(sigma)
    lambda1=lambdas[0]; lambda2=lambdas[1]
    theta=np.arccos(vecs[0][0])

    #Construction d'une ellipse d'isodensité K sensée contenir (proba*100)%
    #des valeurs de l'échantillon
```


COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

```
    K=(1-proba)/(2*pi*sqrt(lambda1*lambda2))
    printellipse(K,lambda1,lambda2,mu[0],mu[1])

    #Construction d'une ellipse d'isodensité K sensée contenir
((proba/2)*100)%
    #des valeurs de l'échantillon
    K=(1-(proba/2))/(2*pi*sqrt(lambda1*lambda2))
    printellipse(K,lambda1,lambda2,mu[0],mu[1])

    #Construction d'une ellipse d'isodensité K sensée contenir
((proba/4)*100)%
    #des valeurs de l'échantillon
    K=(1-(proba/4))/(2*pi*sqrt(lambda1*lambda2))
    printellipse(K,lambda1,lambda2,mu[0],mu[1])
    return(mu,sigma)
mu,sigma=estimateur_1_echantillon_aléatoire(1000,0.99)
print(mu)
print(sigma)
```


COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Conclusion

Pour conclure, nous avons donc étudié la loi normale bidimensionnelle ainsi que ses propriétés, et démontré d'une part, que les lignes d'isodensité K de f_Z sont des ellipses dépendant de μ et Σ, et d'autre part la probabilité qu'un point tiré selon cette loi appartienne à la surface interne de ellipses d'isodensité K, puis l'ellipse d'isodensité K vérifiant $P(Z \in S_k) = p$, avec $p \in [0,1]$. On peut aussi remarquer que cette égalité permet de « contrôler la précision » que l'on souhaite avoir sur un échantillon.

S'en suit l'étude numérique dans laquelle nous avons tracé des ellipses selon l'équation que nous avons trouvé précédemment, pour ensuite montrer que des échantillons de points tirés selon la loi normale bidimensionnelle se trouvent bien dans les ellipses d'isodensité K selon différentes probabilités.

Nous avons pris en compte les cas dégénérés, en considérant dans un premier temps qu'une valeur propre de Σ est nulle, puis dans un second temps en considérant les deux valeurs propres de Σ nulles. L'étude de ces cas dégénérés nous a permis de bien mieux comprendre les résultats de l'étude mathématiques, et donc la relation entre la matrice de covariance Σ, et la manière dont les données se propagent.

Ensuite, nous avons déterminé des estimateurs de μ et Σ, puis montré leur pertinence, et illustré à travers des simulations numériques le fait qu'ils permettent d'approcher les paramètres d'une loi normale bidimensionnelle et que plus l'échantillon considérer était grand plus le résultant était précis.

Par ailleurs étant donnée notre étude, on pourrait essayer de généraliser les résultats obtenus pour une loi normale n-dimensionnelle (n ≥3) et conjecturer que les contours ou « surfaces » d'isodensité de telles lois sont des ellipsoïdes dépendant de μ_n (vecteur coordonnées de dimension n) et Σ_n matrice de covariance de dimension n).

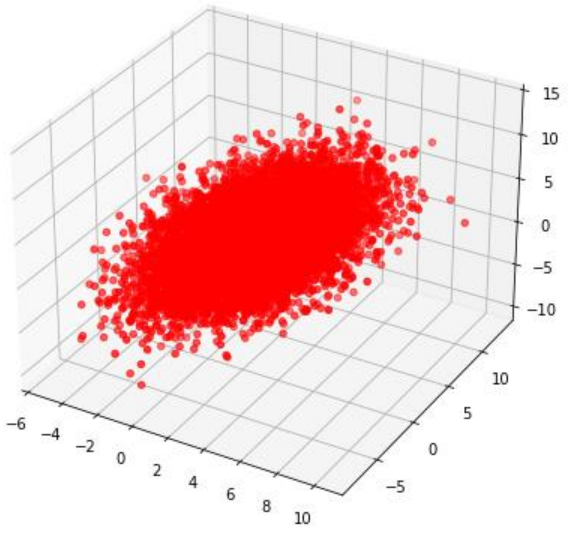
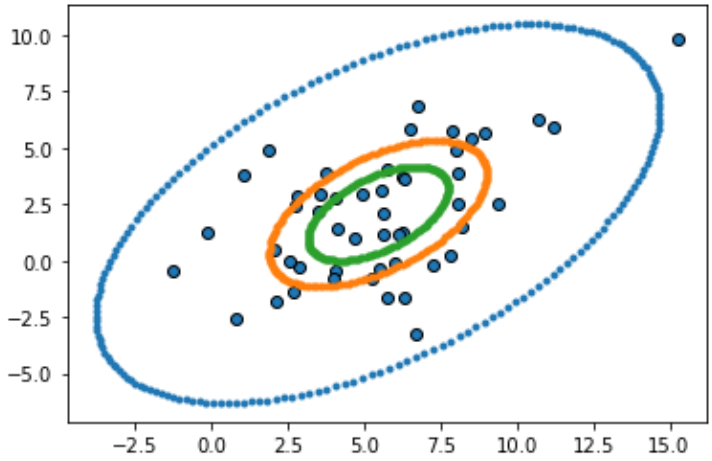


Figure E : échantillon tiré selon une loi normale tri-dimensionnelle quelconque, script dans le rendu

COMBO Houmadi
LIN Peter
ENSISA UHA MULHOUSE

Enfin pour pouvoir se représenter ce qu'est cette loi normale bidimensionnelle et s'en faire une image mentale, considérons l'exemple simple suivant.



Reprise figure D

Un quidam cueille des champignons et répertorie sur une carte l'emplacement de chaque point de cueillette par un point (schéma ci-dessus). En voyant la répartition des points il se dit que celle-ci n'est pas banale et décide de la modéliser selon une loi de probabilité. Il choisit la loi normale bi-dimensionnelle et après calcul, il trace des ellipse contenant 99% ,50% et 25% des champignon. Il parvient à la conclusion que, en moyenne, parmi les champignons qu'il ramasse, 50% sont concentrés dans une surface assez restreinte, alors que les 50 autres pourcents sont répartis sur une bien plus grande zone et en conclue qu'à l'avenir il serait bien plus rentable de concentrer sa cueillette dans une zone délimité par le contour orange plutôt que parcourir une plus grande zone pour qui contient statistiquement moins de champignon par unité de surface.

On peut néanmoins s'interroger sur la représentativité de nos résultats.

On a déterminé la corrélation entre la matrice de covariance et une distribution de données. On aurait donc pour différentes matrices de covariances, différentes formes de distributions de données, ce qui sera toujours vrai.
On a aussi montré que les estimateurs choisis étaient pertinents car non biaisé et convergents. Ainsi, les estimateurs de μ et Σ sont représentatifs des véritables paramètres de la distribution de données à condition que la taille de l'échantillon soit suffisamment grande.
Nous avons choisi une taille d'échantillon de 1000, qu'on considère comme assez grand pour que les estimateurs soient pertinents. De plus, au vu des résultats que nous avons à la figure (1), on peut conclure que les estimateurs de μ et Σ choisis pour cet exemple sont tout à fait pertinents.

Annexe

Source (1) image : https://www.researchgate.net/figure/An-ellipse-rotated-with-an-angle-th-in-clockwise-direction_fig10_273955667

Source (2) équation d'ellipse :

https://fr.wikipedia.org/wiki/Ellipse_(math%C3%A9matiques)

Source (3) – Changement de variables dans une intégrale multiple – Université de Toulouse – L2 Maths :

https://www.math.univ-toulouse.fr/~jroyer/TD/2014-15-L2PS/L2PS-Ch10.pdf

Source (4) détermination des estimateur de la matrice de covariance et de la moyenne :

[BOG] Patrick Bogaert, Probabilités pour scienti_ques et ingénieurs,
éditions De Boeck Supérieur, 2020.

Dérivé d'une expression matricielle :

https://www.di.ens.fr/~fbach/courses/fall2009/formulaire.pdf

https://cs.nyu.edu/~roweis/notes/matrixid.pdf

Source (5), (6) et (7) Matrice de covariance, valeurs propres, vecteurs propres et distribution de données :

https://math.stackexchange.com/questions/23596/why-is-the-eigenvector-of-a-covariance-matrix-equal-to-a-principal-component

https://fr.mathworks.com/matlabcentral/answers/73298-what-does-eigenvalues-expres-in-the-covariance-matrix

https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-des-multi.pdf

Source (8) – CM3 : Lois jointes (II) et inférence statistique classique (I), page 15 – Joël DION, [CHAN] Stanley H. Chan, Introduction to probability for data science, Michigan Publishing, 2021.