Feuille de TP n°7
Vecteurs aléatoires et modèles linéaires gaussiens

1 Vecteurs aléatoires gaussiens

Définition 1.1. *Soit X un vecteur aléatoire défini sur un espace de probabilité $(\Omega, \mathcal{A}, \mathbb{P})$, à valeurs dans $\mathbb{R}^d$ avec $d \geqslant 1$. X est dit gaussien si toute combinaison linéaire de ses composantes est une v.a. gaussienne.*

Remarque 1.2.

Theorème 1.3. *La loi d'un vecteur aléatoire gaussien est entièrement déterminée par son espérance $m = \mathbb{E}(X) \in \mathbb{R}^d$ et sa matrice de covariance $\Gamma = \mathbb{E}((X-m)(X-m)^\top) \in \mathbb{S}_d^+$ où $\mathbb{S}_d^+$ est le cône convexe des matrices carrées $d \times d$ symétriques positives (pas forécment inversibles). Soit $X \sim \mathcal{N}_d(m, \Gamma)$. Alors, pour tout $u \in \mathbb{R}^d$, on a*

$$\Phi_X(u) = \mathbb{E}(\exp(i\langle u, X\rangle) = \exp\left(i\langle u, m\rangle - \frac{1}{2}\langle u, \Gamma u\rangle\right)$$

où $i := \sqrt{-1}$.

Theorème 1.4. *Soit $X \sim \mathcal{N}_d(m, \Gamma)$. X admet une densité f_X par rapport à la mesure de Lebesgue de $\mathbb{R}^d$ si et seulement si Γ est inversible et l'on a*

$$f_X(x) = ((2\pi)^d \det\Gamma)^{-1/2} \exp\left(-\frac{1}{2}\langle x-m, \Gamma^{-1}(x-m)\rangle\right).$$

La loi particulière $\mathcal{N}(0, I)$ où I est la matrice identité de $\mathbb{R}^d$ est appelée *gaussienne standard*. Dans tout ce texte, si x et y sont dans $\mathbb{R}^d$, alors $x^\top y := \langle x, y\rangle := x_1y_1 + \cdots + x_dy_d$. Si $X := (X_1, \ldots, X_d)$ est un vecteur aléatoire de $\mathbb{R}^d$, alors $\mathbb{E}(X) := (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d))$, idem pour l'espérence des matrices aléatoires comme $XX^\top$ qui sont à prendre composante par composante. Voici des propriétés fondamentales des vecteurs gaussiens.

P1. Soit $X = (X_1, \ldots, X_d)$ un vecteur gaussien de matrice de covariance Γ. On a les équivalences suivantes.
- (a) $(X_1, \ldots, X_d)$ sont deux à deux indépendantes.
- (b) $(X_1, \ldots, X_d)$ sont indépendantes dans leur ensemble.
- (c) La matrice Γ est diagonale.

P2. Soit $m \in \mathbb{R}^d$ et Γ une matrice réelle, carrée d'ordre d, symétrique et semi-définie positive. Soit A une matrice réelle, carrée d'ordre d, telle que $A^\top A = \Gamma$. Si $X \sim \mathcal{N}_d(O, I)$ et $Y = AX + m$, alors $Y \sim \mathcal{N}_d(m, \Gamma)$. Réciproquement, étant donnée une loi gaussienne $\mathcal{N}(m, \Gamma)$ et une matrice carrée A telle que $\Gamma = A^\top A$, alors si $X \sim \mathcal{N}(m, I)$, on a $AX + m \sim \mathcal{N}(m, \Gamma)$. La matrice A n'est pas unique, et peut être calculée par exemple par la méthode de Choleski (rapide) ou encore en considérant la racine carrée matricielle de Γ obtenue par diagonalisation en base orthonormée (lent).

P3. Soit $Z = (X, Y)$ un vecteur gaussien de $\mathbb{R}^{d+1}$ avec $X = (X_1, \ldots, X_d)$ d'espérance m et de matrice de covariance inversible Γ. Alors, la loi conditionnelle de Y sachant X est gaussienne d'espérance affine en X $\mathbb{E}(Y|X) = a + \langle b, X\rangle = \mathbb{E}(Y) + \langle b, X - m\rangle$ et de variance $\mathbf{Var}(Y|X) = \mathbf{Var}(Y) - \langle b, \Gamma b\rangle$ avec $a = \mathbb{E}(Y) - \langle b, m\rangle$ et $b = \Gamma^{-1}\mathbf{Cov}(X, Y)$. De plus, $\varepsilon = Y - \mathbb{E}(Y|X)$ est indépendante de X.

Exercice 1.5 (Algorithme de Box-Muller). Soit (X, Y) un vecteur aléatoire de $\mathbb{R}^2$. Montrer que (X, Y) suit la loi normale $\mathcal{N}_2(0, I)$ si et seulement si $X = r\cos\theta$ et $Y = r\sin\theta$ où r et θ sont deux variables aléatoires indépendantes avec r^2 de loi exponentielle $\mathcal{E}(1/2)$ et θ de loi uniforme $\mathcal{U}([0, 2\pi])$. En déduire un code Matlab permettant de générer N réalisations de variables aléatoires indépendantes et de loi normale $\mathcal{N}(m, \sigma^2)$ où le nombre de réalisations N, la moyenne $m \in \mathbb{R}$ et la variance $\sigma^2 > 0$ sont affectées par l'utilisateur. Tracer l'histogramme associé à vos réalisations et le comparer à la fonction `dnorm` de Matlab.

Exercice 1.6. Soit (X,Y) un vecteur aléatoire de $\mathbb{R}^2$, de loi uniforme sur le disque unité

$$\mathcal{D}=\big\{(x,y)\in\mathbb{R}^2,\, x^2+y^2<1\big\}.$$

Soit (r,θ) le couple de coordonnées polaires associé à (X,Y), $X=r\cos\theta$ et $Y=r\sin\theta$. Si $R=2\sqrt{-\log r}/r$, montrer que (RX,RY) suit la loi normale $\mathcal{N}_2(0,I)$. Reprendre l'exercice 1.5 avec le code Matlab suivant.

```
N=input('Entrez la taille de l'échantillon N : ');
max=round(3*N/2);
m=input('Précisez la valeur de la moyenne m : ');
sigma=input('Précisez la valeur de écart type : ');
X=2*rand(max,1)-ones(max,1); Y=2*rand(max,1)-ones(max,1);
S=X.^2+Y.^2; X=X(find(S<1)); Y=Y(find(S<1));
r=sqrt(X.^2+Y.^2); R=2*sqrt(-log(r))./r; Z=R(1:N).*X(1:N);
T=m*ones(N,1)+sqrt(sigma^2)*Z;
```

2 Modèles linéaires gaussiens

Exercice 2.1. On considère le modèle de régression linéaire gaussienne multiple défini, pour $p\geqslant 1$ et $n>p+1$, par

$$Y_i=a+b_1x_{i,1}+b_2x_{i,2}+\ldots+b_px_{i,p}+\varepsilon_i,\ i=1,\ldots,n$$

où pour $j=1,\ldots,p$, $(x_{i,j})_{i,j}$ est une suite de nombres réels connus et où $(\varepsilon_i)_i$ est une suite i.i.d. de loi $\mathcal{N}(0,\sigma^2)$. Montrer que le modèle peut s'écrire sous la forme matricielle $Y=X\theta+\varepsilon$ où X est une matrice rectangulaire de dimension $n\times(p+1)$ à déterminer. On suppose dans toute la suite que le modèle est identifiable i.e. la matrice X est de rang plein égale à $p+1$. Déterminer les estimateurs des moindres carrés $\widehat{\theta}=(\widehat{a},\widehat{b})$ et $\widehat{\sigma}^2$ de $\theta=(a,b)$ et σ^2. Montrer que $\widehat{\theta}$ et $\widehat{\sigma}^2$ sont indépendants, $\widehat{\theta}\sim\mathcal{N}(\theta,\sigma^2(XX^\top)^{-1})$ et $(n-p-1)\widehat{\sigma}^2\sim\sigma^2\chi^2(n-p-1)$. En déduire que, pour $j=1,\ldots,p$

$$\frac{\widehat{a}-a}{\widehat{\sigma}(\widehat{a})}\sim t(n-p-1)\quad\text{et}\quad\frac{\widehat{b}_j-b_j}{\widehat{\sigma}(\widehat{b}_j)}\sim t(n-p-1)$$

avec $\widehat{\sigma}^2(\widehat{a})=\widehat{\sigma}^2(XX^\top)^{-1}_{1,1}$ et $\widehat{\sigma}^2(\widehat{b_j})=\widehat{\sigma}^2(XX^\top)^{-1}_{j+1,j+1}$. On peut ainsi effectuer des tests sur les valeurs a, b_j et σ^2 et obtenir des intervalles de confiance pour a, b_j et σ^2. Montrer que, si $b=0$, $\sum_{i=1}^n(\widehat{a}+\widehat{b}_1x_{i,1}+\cdots+\widehat{b}_px_{i,p}-\overline{Y})^2/p\widehat{\sigma}^2\sim F(p,n-p-1)$. On peut ainsi tester H_0 : « $b=0$ » contre H_1 : « $b\neq 0$ » donc vérifier la significativité des variables explicatives $(x_{ij})_{i,j}$ avec $j=1,\ldots,p$.

Exercice 2.2. Créer un code Matlab permettant de générer une régression linéaire gaussienne multiple où les valeurs n, p, a, b et σ^2 sont affectées par l'utilisateur et où, pour $j=1,\ldots,p$, $(x_{ij})_{i,j}$ est une réalisation d'un n-échantillon de loi uniforme sur $[0,1]$. Calculer les estimateurs des moindres carrés $\widehat{\theta}=(\widehat{a},\widehat{b})$ et $\widehat{\sigma}^2$. Donner pour chaque paramètre a, b et σ^2 un intervalle de confiance de risque $\alpha=5\%$. Représenter graphiquement les Y_i ainsi que les $\widehat{Y}_i=\widehat{a}+\widehat{b}_1x_{i,1}+\cdots+\widehat{b}_px_{i,p}$. Reprendre cet exercice en faisant varier n, p, a, b et σ^2 ainsi que la loi associée à $(x_{ij})_{i,j}$.

Exercice 2.3. On souhaite étudier la variation du taux d'hémoglobine dans le sang au cours d'une opération chirurgicale en fonction de la durée de l'opération et du volume de sang perdu pendant l'opération. On dispose des résultats suivants où y_i représente la valeur observée en pourcentage de la variation du taux d'hémoglobine, $x_{i,1}$ est la durée de l'opération en heures décimales et $x_{i,2}$ est le volume en litres de sang perdu.

y_i	-1.70	-4.61	-5.82	-1.17	-4.23	-3.31	+0.42	-2.98
$x_{i,1}$	1.75	1.33	1.43	1.86	1.81	1.66	1.60	2.00
$x_{i,2}$	0.52	0.59	0.61	0.50	0.54	0.49	0.27	0.47

On suppose que y_i est une réalisation d'une variable aléatoire Y_i de loi $\mathcal{N}(a + b_1 x_{i,1} + b_2 x_{i,2}, \sigma^2)$. Etudier cette régression linéaire multiple grâce à `lsfit` de Matlab. Tester l'hypothèse suivant laquelle la variation du taux d'hémoglobine ne dépend ni de la durée de l'opération ni du volume de sang perdu ou encore l'hypothèse suivant laquelle la variation du taux d'hémoglobine ne dépend pas de la durée de l'opération.

