LSTAT 2040 - TP 1 : Solutions

Rappels sur les vecteurs aléatoires et introduction aux modèles paramétriques

Exercice 1

Notons $X^* = X_3 + aX_1 + bX_2$. Alors, il suit des propriétés de la distribution normale multivariée que

$$\begin{pmatrix} X_1 \\ X_2 \\ X^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a & b & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

suit une distribution normale multivariée. Il s'ensuit que

$$X^* \perp\!\!\!\perp (X_1, X_2) \iff \mathrm{Cov}[X^*, (X_1, X_2)] = (0, 0).$$

Cette dernière relation correspond à un système linéaire de deux équations en a et b donné par

$$\begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}.$$

L'unique solution de ce système d'équations est $a = 1/7$ et $b = -4/7$.

Exercice 2

(a) Il faut vérifier que $f(x_1, x_2) \geq 0$ et $\iint f(x_1, x_2)\, dx_1 dx_2 = 1$ pour toutes valeurs de $(x_1, x_2) \in \mathbb{R}^2$ et $\alpha \in [-1, 1]$. Pour la positivité, on peut se limiter à $(x_1, x_2) \in (0, 1)^2$ car $f \equiv 0$ en dehors de ce rectangle. Pour de telles valeurs, on observe, par définition de f, que l'équation est $f(x_1, x_2) \geq 0$ est équivalente à

$$\alpha(2x_1 - 1)(2x_2 - 1) \geq -1,$$

et cette inégalité est forcément satisfaite car le membre de gauche correspond au produit de trois nombres appartenant à l'intervalle $[-1, 1]$. Ensuite, pour l'intégrale, il suffit de remarquer que, peu importe la valeur de α,

$$\iint f(x_1, x_2)\, dx_1 dx_2 = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1 dx_2 = 1 + \alpha \int_0^1 2x_1 - 1\, dx_1 \int_0^1 2x_2 - 1\, dx_2 = 1,$$

car les deux dernières intégrales valent 0.

(b) Par définition on a

$$\mathrm{E}[X_1 | X_2 = x_2] = \int x_1 f(x_1|x_2)\, dx_1 = \int x_1 \frac{f(x_1, x_2)}{f_2(x_2)}\, dx_1,$$

où $f_2(x_2) = \int f(x_1, x_2)\, dx_1$ est la densité marginale de X_2. On calcule facilement que $f_2(x_2) = 1$ pour n'importe quelle valeur de $0 < x_2 < 1$ et $f_2 \equiv 0$ ailleurs, i.e., $X_2 \sim \mathrm{Unif}(0, 1)$. On déduit, après calcul

$$\mathrm{E}[X_1 | X_2 = x_2] = \int_0^1 x_1 f(x_1, x_2)\, dx_1 = \frac{1}{2} + \frac{\alpha}{6}(2x_2 - 1).$$

De manière similaire, on peut montrer que $\mathrm{E}[X_1^2 | X_2 = x_2] = \frac{1}{3} + \frac{\alpha}{6}(2x_2 - 1)$ et on déduit

$$\mathrm{Var}[X_1 | X_2 = x_2] = \mathrm{E}[X_1^2 | X_2 = x_2] - (\mathrm{E}[X_1 | X_2 = x_2])^2 = \frac{1}{12} - \frac{\alpha^2}{36}(2x_2 - 1)^2.$$

(c) Si $X_1 \perp\!\!\!\perp X_2$, on a trivialement que $\mathrm{Corr}[X_1, X_2] = 0$. Pour l'implication contraire, un calcul direct montre que

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2] = \frac{\alpha}{36}.$$

Notez que les espérances marginales valent trivialement $\mathrm{E}[X_1] = \mathrm{E}[X_2] = 1/2$ puisque $X_1, X_2 \sim \mathrm{Unif}(0,1)$ (le cas de X_1 est identique au cas de X_2 discuté au point (b)). On déduit que

$$\mathrm{Corr}[X_1, X_2] = 0 \implies \alpha = 0 \implies f \equiv 1$$

sur le rectangle ouvert $(0,1)^2$, i.e., $f = f_1 \times f_2$ et on a bien $X_1 \perp\!\!\!\perp X_2$.

Exercice 3

La distribution de T est caractérisée par sa fonction de répartition. On calcule pour tout $t \geq 0$,

$$\begin{aligned}\Pr(T \leq t) &= \Pr(\max(X_1, X_2) \leq t) = \Pr(X_1 \leq t, X_2 \leq t) = \Pr(X_1 \leq t)\Pr(X_2 \leq t) \\ &= (1 - \exp(\theta_1 t))(1 - \exp(\theta_2 t)),\end{aligned}$$

et $\Pr(T \leq t) = 0$ pour tout $t < 0$ car T est le maximum de deux variables aléatoires positives.

Exercice 4

Notons

$$C = \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Alors, on a

$$\begin{pmatrix} AX \\ BX \end{pmatrix} = CX,$$

et l'on déduit que le vecteur aléatoire $(AX, BX)^t$ admet également une distribution normale de dimension 2 avec moyenne $C\mu = (4, 0)^t$ et variance

$$C\Sigma C^t = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Cela implique $\mathrm{Cov}[AX, BX] = 0$ et, comme le vecteur aléatoire $(AX, BX)^t$ est normal, cela montre $AX \perp\!\!\!\perp BX$.

Exercice 5

(a) Pourvu que $k \geq 0$, on a directement $f_T \geq 0$. On cherche donc $k \geq 0$ tel que $\iiint f_T(x, y, z)\, dxdydz = 1$. Après un calcul explicite, on trouve qu'il faut prendre $k = 4/3$.

(b) On calcule, pour tout $0 \leq x \leq 1$,

$$f_X(x) = \iint f_T(x, y, z)\, dydz = \frac{1}{3}(4x + 1),$$

et $f_X \equiv 0$ ailleurs.

(c) On calcule

$$\mathrm{Cov}[Y, Z] = \mathrm{E}[YZ] - \mathrm{E}[Y]\,\mathrm{E}[Z] = \frac{1}{162}.$$

(d) La distribution de $(Y, Z|X)$ est déterminée par sa densité de probabilité :

$$f_{Y,Z|X}(y, z|x) = \frac{f_T(x, y, z)}{f_X(x)} = \frac{4(x + yz)}{4x + 1}, \quad x, y, z \in [0, 1],$$

et $f_{Y,Z|X} \equiv 0$ ailleurs.

Exercice 6

(a) On a

$$\overline{X}_2 = \frac{1}{2}(X_1 + X_2) \quad \text{et} \quad S_2^2 = \frac{1}{2}(X_1 - X_2)^2.$$

Dès lors, pour montrer que $\overline{X}_2 \perp\!\!\!\perp S_2^2$, il suffit de montrer que $X_1 + X_2$ et $X_1 - X_2$ sont indépendants. Cela est une conséquence direct du fait que le vecteur aléatoire

$$\begin{pmatrix} X_1 + X_2 \\ X_1 - X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

est une transformation linéaire du vecteur aléatoire (X_1, X_2) qui suit une distribution normale bivariée, donc $(X_1 + X_2, X_1 - X_2)$ est également normal. Un calcul direct donne

$$\mathrm{Cov}[X_1 + X_2, X_1 - X_2] = 0$$

et l'on déduit que $X_1 + X_2 \perp\!\!\!\perp X_1 - X_2$.

(b) On observe que

$$\begin{pmatrix} \overline{X}_n \\ X_1 - \overline{X}_n \\ \vdots \\ X_n - \overline{X}_n \end{pmatrix} = B \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

pour

$$B = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1 - \frac{1}{n} \end{pmatrix}.$$

Ensuite, par définition du produit matriciel, on calcule

$$(BB^t)_{11} = \Big(\frac{1}{n}, \ldots, \frac{1}{n}\Big)^t \Big(\frac{1}{n}, \ldots, \frac{1}{n}\Big) = \frac{1}{n}$$

et, pour n'importe quel $2 \leq k \leq n+1$,

$$(BB^t)_{1k} = (BB^t)_{k1} = \Big(\frac{1}{n}, \ldots, \frac{1}{n}\Big)^t \Big(-\frac{1}{n}, \ldots, 1 - \frac{1}{n}, \ldots, -\frac{1}{n}\Big) = 0,$$

ce qui montre bien

$$BB^t = \begin{pmatrix} \frac{1}{n} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & A & \\ 0 & & & \end{pmatrix},$$

où A est une certaine matrice de dimension $n \times n$.

(c) On déduit de la relation

$$\begin{pmatrix} \overline{X}_n \\ X_1 - \overline{X}_n \\ \vdots \\ X_n - \overline{X}_n \end{pmatrix} = B \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$$

que le vecteur aléatoire $(\overline{X}_n, X_1 - \overline{X}_n, \ldots, X_n - \overline{X}_n)$ admet une distribution normale de dimension $n+1$ de variance

$$B\sigma^2 I_n B^t = \sigma^2 BB^t,$$

où I_n est la matrice identité de dimension $n \times n$. Au vue de notre calcul pour BB^t, on déduit que

$$\overline{X}_n \perp\!\!\!\perp X_i - \overline{X}_n, \qquad 1 \leq i \leq n.$$

Puisque $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$, on déduit $\overline{X}_n \perp\!\!\!\perp S_n^2$.

Exercice 7

Considérons les deux paramètres

$$\theta_1 = (p = 0.5, \mu_1 = \mu, \mu_2 = 0, \sigma_1^2 = \sigma^2, \sigma_2^2 = \sigma^2) \qquad \theta_2 = (p = 0.5, \mu_1 = 0, \mu_2 = \mu, \sigma_1^2 = \sigma^2, \sigma_2^2 = \sigma^2)$$

pour un certain $\mu \neq 0$ et $\sigma > 0$. Alors, on a clairement que $\theta_1 \neq \theta_2$, pourtant, il est immédiat de voir que $f(x;\theta_1) = f(x;\theta_2)$. Cela montre que ce modèle n'est pas identifiable.

Exercice 8

Rappelons qu'une famille de distributions $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ forme une famille exponentielle (à 1 paramètre) si l'on peut réécrire leur fonction de densité/masse $f(\cdot;\theta)$ comme

$$f(x;\theta) = h(x)\exp(g(\theta)T(x) - B(\theta)), \qquad \forall x,$$

où h et T sont des fonctions de x seulement, g et B sont des fonctions de θ seulement.

(a) Les densités de la famille de distributions $\{P_\sigma = \mathrm{N}(1,\sigma^2) : \sigma > 0\}$ peuvent se réécrire comme

$$f(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-1)^2\right) = \exp\left(-\frac{1}{2\sigma^2}(x-1)^2 - \frac{1}{2}\log(2\pi\sigma^2)\right).$$

En posant $h(x) = 1$, $g(\sigma) = -\frac{1}{2\sigma^2}$, $T(x) = (x-1)^2$ et $B(\sigma) = \frac{1}{2}\log(2\pi\sigma^2)$, on trouve bien que la famille de distributions $\{P_\sigma = \mathrm{N}(1,\sigma^2) : \sigma > 0\}$ est une famille exponentielle.

(b) Les densités de la famille de distributions $\{P_\mu = \mathrm{N}(\mu,\mu) : \mu > 0\}$ peuvent se réécrire comme

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\mu}}\exp\left(-\frac{1}{2\mu}(x-\mu)^2\right) = \exp(x)\exp\left(-\frac{1}{2\mu}x^2 - \frac{1}{2}(\mu + \log(2\pi\mu))\right).$$

En posant $h(x) = \exp(x)$, $g(\mu) = -\frac{1}{2\mu}$, $T(x) = x^2$ et $B(\mu) = \frac{1}{2}(\mu + \log(2\pi\mu))$, on trouve bien que la famille de distributions $\{P_\mu = \mathrm{N}(\mu,\mu) : \mu > 0\}$ est une famille exponentielle.

(c) Les densités de la famille de distributions $\{P_\pi = \mathrm{Geo}(\pi) : 0 < \pi \leq 1\}$ peuvent se réécrire comme

$$f(x;\pi) = (1-\pi)^{x-1}\pi = \exp\left(x\log(1-\pi) - \log\left(\frac{1-\pi}{\pi}\right)\right), \qquad x \in \{1,2,\ldots\}.$$

En posant $h(x) = 1$, $g(\pi) = \log(1-\pi)$, $T(x) = x$ et $B(\pi) = \log\left(\frac{1-\pi}{\pi}\right)$, on trouve bien que la famille de distributions $\{P_\pi = \mathrm{Geo}(\pi) : 0 < \pi \leq 1\}$ est une famille exponentielle.

(d) Les densités de la famille de distributions $\{P_\theta = \mathrm{Unif}[0,\theta] : \theta > 0\}$ peuvent se réécrire comme

$$f(x;\theta) = \frac{1}{\theta}\mathrm{I}(0 \leq x \leq \theta) = \exp(\log(\mathrm{I}(0 \leq x \leq \theta)) - \log\theta).$$

Observons que l'expression $\log(\mathrm{I}(0 \leq x \leq \theta))$ ne pourra jamais se factoriser en un produit de termes ne dépendant que de x et θ. Dès lors, la famille $\{P_\theta = \mathrm{Unif}[0,\theta] : \theta > 0\}$ n'est pas une famille exponentielle.

(e) Les densités de la famille de distributions $\{P_\sigma = \mathrm{Rayleigh}(\sigma) : \sigma > 0\}$ peuvent se réécrire comme

$$f(x;\mu) = \frac{2}{\sigma}x\exp\left(-\frac{x^2}{2\sigma}\right)\mathrm{I}(x \geq 0) = x\mathrm{I}(x \geq 0)\exp\left(-\frac{x^2}{2\sigma} - \log\left(\frac{\sigma}{2}\right)\right).$$

En posant $h(x) = x\mathrm{I}(x \geq 0)$, $g(\sigma) = -\frac{1}{2\sigma}$, $T(x) = x^2$ et $B(\sigma) = \log\left(\frac{\sigma}{2}\right)$, on trouve bien que la famille de distributions $\{P_\sigma = \mathrm{Rayleigh}(\sigma) : \sigma > 0\}$ est une famille exponentielle.

(f) Rappelons qu'une famille de distributions $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^2\}$ forme une famille exponentielle de dimension 2 si l'on peut réécrire leur fonction de densité/masse $f(\cdot;\theta)$ comme

$$f(x;\theta) = h(x)\exp\left(g_1(\theta)T_1(x) + g_2(\theta)T_2(x) - B(\theta)\right), \qquad \forall x,$$

où h et T_i sont des fonctions de x seulement, g_i et B sont des fonctions de θ seulement, pour $i \in \{1, 2\}$. Les densités de la famille de distributions $\{P_{(\lambda,k)} = \text{Weibull}(\lambda, k) : (\lambda, k) \in (0, \infty)^2\}$ peuvent se réécrire comme

$$\begin{aligned} f(x; \lambda, k) &= \frac{k}{\lambda}(x/\lambda)^{k-1} \exp\left(-(x/\lambda)^k\right) \mathrm{I}(x \geq 0) \\ &= \mathrm{I}(x \geq 0) \exp\left(-\frac{x^k}{\lambda^k} + (k-1)\log(x) - (k-1)\log\lambda + \log(k/\lambda)\right). \end{aligned}$$

Le premier terme dans l'exponentielle ne pourra jamais se réécrire comme un produit $g_1(\lambda, k)T_1(x)$ pour une fonction g_1 ne dépendant uniquement de (λ, k) et une fonction T_1 ne dépendant uniquement de x. Dès lors, la famille $\{P_{(\lambda,k)} = \text{Weibull}(\lambda, k) : (\lambda, k) \in (0, \infty)^2\}$ n'est pas une famille exponentielle de dimension 2.
Observons tout de même que, si on avait supposé k connu et fixé, la famille $\{P_\lambda = \text{Weibull}(\lambda, k) : \lambda > 0\}$ à un paramètre aurait bien formé une famille exponentielle de dimension 1 avec statistique naturelle $T(x) = x^k$.

Exercice 9

La densité jointe de X peut s'écrire comme

$$f(x; \lambda_1, \lambda_2) = \prod_{i=1}^{n_1} \lambda_1 \exp(-\lambda_1 x_{1i}) \prod_{j=1}^{n_2} \lambda_2 \exp(-\lambda_2 x_{2j}) = \exp\left(-\lambda_1 \sum_{i=1}^{n_1} x_{1i} - \lambda_2 \sum_{j=1}^{n_2} x_{2j} - \Big(n_1 \log\lambda_1 + n_2 \log\lambda_2\Big)\right)$$

En posant $h(x) = 1$, $g_1(\theta) = -\lambda_1$, $T_1(x) = \sum_{i=1}^{n_1} x_{1i}$, $g_2(\theta) = -\lambda_2$, $T_2(x) = \sum_{j=1}^{n_2} x_{2j}$ et $B(\theta) = n_1 \log\lambda_1 + n_2 \log\lambda_2$, on voit que la famille $\{f(\cdot; \theta) : \theta = (\lambda_1, \lambda_2) \in (0, \infty)^2\}$ forme une famille exponentielle de dimension $k = 2$.

Exercice 10

Les fonctions de masse dans $\mathcal{P}$ sont

$$f_n(x; \pi) = \frac{n!}{x_1! \cdots x_k!} \pi_1^{x_1} \cdots \pi_m^{x_m}, \qquad x_1, \ldots, x_m \in \{0, \ldots, n\},$$

où $\sum_{j=1}^m x_j = n$ et $\sum_{j=1}^m \pi_j = 1$. Notons $h(x) = \frac{n!}{x_1! \cdots x_k!}$. Alors, on calcule,

$$\begin{aligned} f_n(x; \pi) &= h(x) \exp\left(\sum_{j=1}^m x_j \log(\pi_j)\right) \\ &= h(x) \exp\left(\sum_{j=1}^{m-1} x_j \log(\pi_j) + \Big(n - \sum_{i=1}^{m-1} x_i\Big) \log\Big(1 - \sum_{i=1}^{m-1} \pi_i\Big)\right) \\ &= h(x) \exp\left(\sum_{j=1}^{m-1} x_j \log\Big(\frac{\pi_j}{1 - \sum_{i=1}^{m-1} \pi_i}\Big) - n \log\Big(\frac{1}{1 - \sum_{i=1}^{m-1} \pi_i}\Big)\right). \end{aligned}$$

En posant, pour $j \in \{1, \ldots, m-1\}$, $g_j(\pi_1, \ldots, \pi_{m-1}) = \log\Big(\frac{\pi_j}{1 - \sum_{i=1}^{m-1} \pi_i}\Big)$, $T_j(x) = x_j$ et $B(\pi_1, \ldots, \pi_{m-1}) = n \log\Big(\frac{1}{1 - \sum_{i=1}^{m-1} \pi_i}\Big)$, on voit que la famille $\mathcal{P}$ forme une famille exponentielle de dimension $k = m - 1$.