Feuille d'exercices 3

Exercice 1 (Tests dans le modèle linéaire gaussien)
On se place dans le modèle linéaire gaussien :

$$Y = X\beta + \varepsilon\,,$$

où X est une matrice $n \times p$ de rang p, β un vecteur inconnu de $\mathbb{R}^p$, et $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. On note $\widehat{\beta} = ({}^t XX)^{-1}\, {}^t XY$ et $\widehat{\sigma}^2 = \frac{\|Y - X\widehat{\beta}\|^2}{n-p}$ les estimateurs usuels de β et de σ^2.

1. *Test de Student d'une relation affine.*
   Soient $c \in \mathbb{R}^p \setminus \{(0, \dots, 0)\}$ et $a \in \mathbb{R}$.
   (a) Déterminer la loi de la variable
   $$\frac{{}^t c\widehat{\beta} - {}^t c\beta}{\widehat{\sigma}\sqrt{{}^t c({}^t XX)^{-1} c}}\,.$$
   (b) En déduire un test de niveau α pour
   $$H_0 : {}^t c\beta = a \qquad \text{contre} \qquad H_1 : {}^t c\beta \neq a\,.$$
   (c) Donner un intervalle de confiance de niveau $1 - \alpha$ pour ${}^t c\beta$.
2. *Test de Fisher d'un sous-modèle.*
   Soient V le sous-espace vectoriel de $\mathbb{R}^n$ de dimension p engendré par les colonnes de X et W un sous-espace vectoriel de V, de dimension $q < p$. On note P_V et P_W les matrices de projection orthogonale sur V et W respectivement.

   (a) Montrer que si $X\beta \in W$, alors la statistique
   $$F = \frac{\|P_W Y - P_V Y\|^2/(p-q)}{\|Y - P_V Y\|^2/(n-p)}$$
   suit une loi de Fisher $\mathcal{F}(p-q, n-p)$, et est indépendante de $P_W Y$.
   (b) En déduire un test de niveau α pour
   $$H_0 : X\beta \in W \qquad \text{contre} \qquad H_1 : X\beta \in V \setminus W\,.$$
   (c) Pour $q = p - 1$, montrer que ce test est le même que celui déterminé en question 1.(b), pour certains $c \in \mathbb{R}^p$ et $a \in \mathbb{R}$ à déterminer.
3. *Test de Wald de plusieurs hypothèses affines.*
   Soient C une matrice $k \times p$, avec $k \leq p$, de rang plein, et $a \in \mathbb{R}^k$.
   (a) Quelle est la loi de $C\widehat{\beta} - C\beta$?
   (b) Montrer que la matrice $\Sigma = C({}^t XX)^{-1}\, {}^t C$ est symétrique définie positive.
   (c) Montrer que la variable
   $$\frac{{}^t(C\widehat{\beta} - C\beta)\Sigma^{-1}(C\widehat{\beta} - C\beta)/k}{\|Y - X\widehat{\beta}\|^2/(n-p)}$$
   suit une loi de Fisher $\mathcal{F}(k, n-p)$.
   (d) En déduire un test de niveau α pour
   $$H_0 : C\beta = a \qquad \text{contre} \qquad H_1 : C\beta \neq a\,.$$
   (e) Construire une ellipsoïde de confiance de niveau $1 - \alpha$ pour $C\beta$.

Exercice 2 (Régression Ridge - Régularisation de Tikhonov)
On considère le modèle de regression

$$\underset{(n,1)}{Y} = \underset{(n,k)}{X}\underset{(k,1)}{\theta} + \underset{(n,1)}{\xi}.$$

On suppose que X est une matrice déterministe, $\mathbb{E}[\xi] = 0$, $\mathbb{E}[\xi\xi^T] = \sigma^2 I_n$,

1. On suppose que $k > n$. Que dire de l'estimation par moindres carrés ?
2. On appelle estimateur `Ridge regression` de paramètre de régularisation $\lambda > 0$ l'estimateur

$$\widehat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^k} \left\{ \|Y - X\theta\|^2 + \lambda \|\theta\|^2 \right\}.$$

   Exprimez $\widehat{\theta}_\lambda$ en fonction de X, Y et λ. Cet estimateur est-il défini pour $k > n$?
3. Calculez la moyenne et la matrice de covariance de l'estimateur Ridge. Est-il sans biais ?

Exercice 3 (Maximum de gaussiennes corrélées)
On considère $\rho \in [0,1]$ et un vecteur gaussien $(X_1, \ldots, X_n)$ centré et tel que, pour tous $i, j \in [\![1, n]\!]$,

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} 1 & \text{si } i = j, \\ \rho \text{ si } i \neq j. \end{cases}$$

1. Dans le cas $n = 2$, calculer $\mathbb{E}[\max(X_1, X_2)]$.
2. Dans le cas général, on note $a_n(\rho) = \mathbb{E}[\max(X_1, \ldots, X_n)]$. Exprimer $a_n(\rho)$ en fonction de $a_n(0)$.
3. On se place dans le cas $\rho = 0$. Montrer que pour tout $\lambda \in \mathbb{R}_+$,

$$e^{\lambda \mathbb{E}[t \max(X_1, \ldots, X_n)]} \leq n e^{\lambda^2/2}$$

   et en déduire une borne sur $a_n(0)$.

Exercice 4 (Maximum de variables de Poisson indépendantes)
Soient $X_1, \ldots, X_n$ des variables i.i.d. de loi de Poisson $\mathcal{P}(1)$.

1. Montrer que $\mathbb{E}[\max_{1 \leq i \leq n}(X_i)] \leq \frac{2\ln(n)}{\ln\ln(n)}$.
2. En utilisant l'inégalité de Markov, trouver une borne inférieure du même ordre.

Exercice 5 (Conditionnement linéaire gaussien)
Soit $X = (X_1, \ldots, X_n)$ un vecteur gaussien. On se donne une matrice A de taille $m \times n$, que l'on suppose de rang m. Donner l'espérance et loi conditionnelle de X sachant AX.

Exercice 6 (Statistique exhaustive, statistique complète)
Soit $\Theta \subset \mathbb{R}^d$, $d \geq 1$, et $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$ un modèle statistique dominé par une mesure σ-finie μ sur $(E, \mathcal{E})$, i.e. pour tout $\theta \in \Theta$, et pour tout $x \in E$, $dP_\theta(x) = f_\theta(x) d\mu(x)$. On considère une variable aléatoire X de loi P_θ, pour $\theta \in \Theta$.

1. On dit qu'une statistique $T(X)$ est *exhaustive pour* θ si la loi de X conditionnellement à $T(X)$ ne dépend pas de θ. Le théorème de factorisation de Fisher-Neyman assure que $T(X)$ est exhaustive ssi il existe des fonctions $h : E \to \mathbb{R}_+$ et $g : \Theta \times T(E) \to \mathbb{R}_+$ telles que

$$\forall \theta \in \Theta,\ \forall x \in E,\ f_\theta(x) = h(x) \cdot g(\theta, T(x)).$$

   (a) Cas Poissonien. On suppose que $X = (X_1, \ldots, X_n)$ est une suite i.i.d. de loi de Poisson de paramètre $\lambda \geq 0$ (le cas $\lambda = 0$ correspondant à une loi de Dirac en 0), et on cherche à estimer $\exp(-\lambda)$. Donner une statistique exhaustive $T(X)$ et rappeler sa loi.

   (b) Cas général. On suppose qu'il existe une statistique exhaustive $T(X)$. Soit un estimateur $\widehat{\theta}(X)$ de θ. On considère la variable aléatoire $\theta^*(X)$ définie par

$$\theta^*(X) = \mathbb{E}\left[\widehat{\theta}(X) \,|\, T(X)\right].$$

Pourquoi peut-on dire que $\theta^*(X)$ est un estimateur ?
Prouver le théorème de Rao-Blackwell :

$$\mathbb{E}\left[\|\theta^*(X)-\theta\|^2\right] \leq \mathbb{E}\left[\left\|\widehat{\theta}(X)-\theta\right\|^2\right].$$

On dit que $\theta^*(X)$ est la version Rao-Blackwelliséé de $\widehat{\theta}(X)$. Ce théorème garantit qu'il est meilleur que l'estimateur d'origine $\widehat{\theta}(X)$ au sens du risque quadratique.

(c) Cas Poissonien. On considère l'estimateur naïf W défini par

$$W = \begin{cases} 1 & \text{si } X_1 = 0 \\ 0 & \text{sinon.} \end{cases}$$

Trouver une forme simple pour la version Rao-Blackwellisée de W.

2. On dit qu'une statistique $T(X)$ est *complète* (ou *totale*) si pour toute fonction mesurable $g : T(E) \to \mathbb{R}$,

$$\{\forall \theta \in \Theta,\ \mathbb{E}_\theta\left[g(T(X))\right] = 0\} \implies \{\forall \theta \in \Theta,\ \mathbb{P}_\theta\left(g(T(X)) = 0\right) = 1\}.$$

(a) Cas Poissonien. Montrer que la statistique exhaustive considérée à la question 1(a) est complète.

(b) Cas général. Soit $\widehat{\theta}(X)$ un estimateur sans biais de θ. On suppose que $T(X)$ est une statistique à la fois exhaustive et complète et l'on considère $\theta^*(X) = \mathbb{E}\left[\widehat{\theta}(X)\,|\,T(X)\right]$. Montrer que $\theta^*(X)$ est l'unique estimateur sans biais de risque quadratique minimal, i.e. si un estimateur sans biais a un risque quadratique inférieur ou égal à celui de $\theta^*(X)$, alors il est presque sûrement égal à $\theta^*(X)$. Ce résultat s'appelle le théorème de Lehmann-Scheffé.

(c) Cas Poissonien. Conclure sur l'estimateur obtenu à la question 1(c).