Sorbonne Université
Master Mathématiques et Applications
MU4MA015 : Statistique

1er décembre 2023
Aucun document autorisé
Durée : 1 heure

Contrôle Continu

Rappels

- Il sera tenu compte du soin apporté à la présentation et à la rédaction.
- Si vous n'arrivez pas à démontrer un résultat, vous pouvez l'admettre pour la suite de l'exercice.
- Pour $\lambda > 0$ et $p > 0$, si $T \sim \Gamma(p, \lambda)$ alors $\mathbb{E}[T] = \frac{p}{\lambda}$ et $\mathrm{Var}(T) = \frac{p}{\lambda^2}$. De plus, si $p > 2$, alors

$$\mathbb{E}\left[\frac{1}{T}\right] = \frac{\lambda}{p-1}, \qquad \text{et} \qquad \mathbb{E}\left[\frac{1}{T^2}\right] = \frac{\lambda^2}{(p-1)(p-2)}.$$

Exercice 1 On part du modèle linéaire gaussien classique

$$Y = X\beta + \boldsymbol{\varepsilon},$$

où X est la matrice $(n \times p)$ du plan d'expérience, $\beta = [\beta_1, \ldots, \beta_p]'$ un vecteur de $\mathbb{R}^p$, Y est le vecteur $(n \times 1)$ des observations, et $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$. Les paramètres β et σ^2 sont supposés inconnus et on souhaite les estimer. Néanmoins, on n'observe pas comme d'habitude les Y_i, mais des moyennes par classe. Plus précisément, les n données sont réparties en L classes $C_1, \ldots, C_L$ d'effectifs respectifs **connus** $n_1, \ldots, n_L$ et on a seulement accès aux moyennes par classe, à savoir pour tout $\ell \in \{1, \ldots, L\}$ et pour tout $j \in \{1, \ldots, p\}$:

$$\overline{Y}_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} Y_i \qquad \text{et} \qquad \overline{x}_{\ell j} = \frac{1}{n_\ell} \sum_{i \in C_\ell} x_{ij},$$

avec $\overline{X} := [\overline{x}_{\ell j}]_{1 \leq \ell \leq L, 1 \leq j \leq p}$ supposée de rang p. En notant $\overline{\varepsilon}_\ell = \frac{1}{n_\ell} \sum_{i \in C_\ell} \varepsilon_i$, le modèle peut donc se mettre sous la forme

$$\overline{Y} = \overline{X}\beta + \overline{\boldsymbol{\varepsilon}}.$$

1. Donner la loi de $\overline{\boldsymbol{\varepsilon}}$.
2. Déterminer une matrice diagonale Δ telle que si l'on définit $X^* = \Delta\overline{X}$, $Y^* = \Delta\overline{Y}$ et $\boldsymbol{\varepsilon}^* = \Delta\overline{\boldsymbol{\varepsilon}}$, on ait $Y^* = X^*\beta + \boldsymbol{\varepsilon}^*$, avec $\boldsymbol{\varepsilon}^* \sim \mathcal{N}(0, \sigma^2 I_L)$.
3. Justifier que la matrice X^* est de rang p.
4. En déduire un estimateur $\widehat{\beta}$ de β et préciser sa loi.
5. Déterminer de même un estimateur sans biais $\widehat{\sigma}^2$ de σ^2 et préciser sa loi.
6. Quelle est la variance de $\widehat{\sigma}^2$? La comparer à celle de l'estimateur sans biais $\widetilde{\sigma}^2$ dans le modèle linéaire gaussien initial $Y = X\beta + \boldsymbol{\varepsilon}$ où l'on a accès à toutes les observations $(Y_i)_{1 \leq i \leq n}$. Interpréter le résultat.

Exercice 2 Soient $\lambda > 0$ et $n \geq 5$ entier. On considère le modèle bayésien suivant :

$$\boldsymbol{\theta} \sim \Pi = \mathcal{E}(\lambda),$$
$$\mathbf{X} = (X_1, \ldots, X_n) \,\big|\, \boldsymbol{\theta} \sim \mathcal{N}\left(0, \frac{1}{\boldsymbol{\theta}}\right)^{\otimes n}.$$

1. Déterminer la loi a posteriori $\Pi\left[\cdot \,\middle|\, \mathbf{X}\right]$. On notera $S_{\mathbf{X}}^2 = \sum_{i=1}^n X_i^2$.
2. Donner un estimateur de Bayes pour la perte quadratique.
3. On considère maintenant l'estimateur $\widehat{\theta}_n(\mathbf{X}) = \frac{n}{S_{\mathbf{X}}^2}$. L'objectif de cette question est de calculer son risque de Bayes pour Π et la perte quadratique.
   (a) Soit $\theta > 0$. Sachant $\boldsymbol{\theta} = \theta$, quelle loi suit $\theta S_{\mathbf{X}}^2$?
   (b) On rappelle que, pour tout entier non-nul d, la loi χ_d^2 est identique à la loi $\Gamma\left(\frac{d}{2}, \frac{1}{2}\right)$. En déduire, pour tout $\theta > 0$, $\mathbb{E}_\theta\left[\frac{1}{S_{\mathbf{X}}^2}\right]$ et $\mathbb{E}_\theta\left[\frac{1}{S_{\mathbf{X}}^4}\right]$.
   (c) En déduire que, pour tout $\theta > 0$, le risque ponctuel vaut
   $$\mathbf{R}(\theta, \widehat{\theta}_n(\mathbf{X})) = \frac{2(n+4)}{(n-2)(n-4)}\theta^2.$$
   (d) En déduire le risque de Bayes pour Π de l'estimateur $\widehat{\theta}_n(\mathbf{X})$.
4. (a) L'objectif est maintenant d'utiliser la méthode du Bayésien empirique pour calibrer λ. Montrer que, pour tout $\lambda > 0$, la densité marginale f_λ de la loi de $\mathbf{X}$ vérifie :
   $$f_\lambda(\mathbf{X}) \propto \frac{\lambda}{\left(\lambda + \frac{S_{\mathbf{X}}^2}{2}\right)^{\frac{n}{2}+1}}$$
   et déterminer l'estimateur du maximum de vraisemblance marginale de λ.
   (b) Donner la pseudo-loi a posteriori obtenue par la méthode du Bayésien empirique dans ce cas. Que retrouve-t-on en considérant la pseudo-moyenne a posteriori associée ?