

2. Information de Fisher. Estimation par maximum de vraisemblance

Objectifs : Calculer une information de Fisher, l'interpréter. Savoir donner la borne de Cramer-Rao. Pratiquer l'estimation par maximum de vraisemblance. Les exercices 2.1 et 2.2 sont à faire pendant le TD, les autres sont à chercher de votre côté.

Exercice 2.1 (Information de Fisher dans un modèle à deux paramètres). On considère le modèle statistique où X_1, \dots, X_n sont des variables réelles indépendantes, avec pour tout $1 \leq i \leq n$, $X_i \sim \mathcal{N}(\alpha + \beta t_i, 1)$, où les constantes $(t_i)_{1 \leq i \leq n}$ sont connues et $\alpha, \beta \in \mathbb{R}$ sont des paramètres inconnus. On admettra que ce modèle est régulier.

1. Montrer que la matrice d'information de Fisher $I(\alpha, \beta)$ s'écrit

$$I(\alpha, \beta) = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

Solution. On écrit la log-vraisemblance qui est bien définie et vaut $\ell(\alpha, \beta, x) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^n (x_i - \alpha - t_i \beta)^2$, différentiable sur \mathbb{R}^2 et de gradient donné par $\nabla_{\alpha, \beta} \ell(\alpha, \beta, x) = \begin{pmatrix} \sum_{i=1}^n (x_i - \alpha - t_i \beta) \\ \sum_{i=1}^n t_i (x_i - \alpha - t_i \beta) \end{pmatrix}$. La matrice de covariance de $\nabla_{\alpha, \beta} \ell(\alpha, \beta, X)$ existe et vaut

$$I(\alpha, \beta) = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

2. Déterminer une condition nécessaire et suffisante sur les $(t_i)_{1 \leq i \leq n}$ pour que $I(\alpha, \beta)$ soit inversible. Interpréter cette condition. *Solution.* A noter que l'information de Fisher est ici constante, et la matrice $I(\alpha, \beta)$ est (symétrique) et inversible si et seulement si $\det I(\alpha, \beta) \neq 0$ i.e.

$$n \sum_{i=1}^n t_i^2 \neq \left(\sum_{i=1}^n t_i \right)^2.$$

Mais par Cauchy-Schwarz, $n \sum_{i=1}^n t_i^2 \geq \left(\sum_{i=1}^n t_i \right)^2$ avec égalité si et seulement si (t_1, \dots, t_n) et $(1, \dots, 1)$ sont colinéaires c'est-à-dire si les t_i sont constants. Dans ce cas, l'identifiabilité est alors mise en défaut : toutes les variables sont i.i.d. $\mathcal{N}(\alpha + \beta t, 1)$ et on a une symétrie $(\alpha, \beta) \mapsto (0, \alpha + \beta t)$. On peut retrouver géométriquement ce problème d'identifiabilité avec la matrice d'information de Fisher : si $I(\alpha, \beta)$ n'est pas inversible, elle l'est en aucun (α, β) (car elle est constante), et donc la log-vraisemblance reste localement dans un sous-espace de dimension 1. Il existe donc un sous-espace non trivial sur lequel la log-vraisemblance ne bouge pas quand bien même α et β bougent. Cela entraîne bien un problème d'identifiabilité.

3. Sous la condition de la question 2, donner une borne inférieure sur la variance de tout estimateur non biaisé de α . (On prendra donc $h(\alpha, \beta) = \alpha$). *Solution.* On vérifie aisément que $\mathbb{E}[\nabla_{\alpha, \beta} \ell(\alpha, \beta, X)] = 0$, la première condition du Théorème de Cramer-Rao est satisfaite. Pour l'autre condition, on doit imposer à l'estimateur T non biaisé de vérifier $\mathbb{E}[T(X) \nabla_{\alpha, \beta} \ell(\alpha, \beta, X)^T] = J_\phi(\alpha, \beta) = \begin{pmatrix} 1 & 0 \end{pmatrix}$, ici $\phi(\alpha, \beta) = \alpha$. La borne de Cramer-Rao donne qu'alors, pour de tels T et pour tout α, β ,

$$\text{Var}_{\alpha, \beta}(T) \geq J_\phi(\alpha, \beta) I(\alpha, \beta)^{-1} J_\phi(\alpha, \beta)^T$$

. On a, sous les conditions d'identifiabilité,

$$I(\alpha, \beta)^{-1} = \frac{1}{n \sum_{i=1}^n t_i^2 - \left(\sum_{i=1}^n t_i \right)^2} \begin{pmatrix} \sum_{i=1}^n t_i^2 & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & n \end{pmatrix}.$$

ce qui donne

$$\text{Var}_{\alpha,\beta}(T) \geq \frac{\sum_{i=1}^n t_i^2}{n \sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2}.$$

Cette borne est d'autant plus grande que les t_i sont grands et proches les uns des autres: c'est logique, l'estimation de α est plus difficile car la partie en $t_i\beta$ prend plus de signal.

4. Supposons que l'on connaît β . Reprendre les calculs pour trouver $\mathcal{I}(\alpha)$ et donner dans ce cas une borne inférieure sur la variance de tout estimateur non biaisé de α . Solution. Dans ce cas $\mathcal{I}(\alpha) = n$ et la borne est unidimensionnelle, elle s'écrit

$$\text{Var}_{\alpha,\beta}(T) \geq \frac{1}{n}.$$

On peut remarquer que l'estimateur de la moyenne empirique, recentré par les $t_i\beta$, est alors efficace.

5. Comment les bornes inférieures des questions 3 et 4 se comparent-elles ? Interpréter. Solution. On a par comparaison immédiate que $\frac{\sum_{i=1}^n t_i^2}{n \sum_{i=1}^n t_i^2 - (\sum_{i=1}^n t_i)^2} \geq \frac{1}{n}$. En effet, on s'attend à bien mieux pouvoir estimer α si l'information sur β est donnée. Cela est compatible avec les bornes obtenues.

Exercice 2.2 (Deux estimateurs dans le modèle uniforme). On considère le modèle dans lequel X_1, \dots, X_n sont des variables i.i.d. suivant une loi uniforme sur $[0, \theta]$, où $\theta > 0$ est un paramètre inconnu.

1. Trouver un estimateur $\hat{\theta}_1$ de θ en utilisant la méthode des moments appliquée au premier moment. Solution. Puisque $\mathbb{E}_\theta[X_1] = \theta/2$, la méthode des moments donne l'estimateur $\hat{\theta}_1 = 2\bar{X}_n$.
2. Déterminer la vraisemblance dans ce modèle, et montrer qu'il existe un unique estimateur du maximum de vraisemblance $\hat{\theta}_2$. Donner son expression. Solution. La fonction de vraisemblance s'écrit

$$L(\theta, x) = \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \mathbb{1}_{0 \leq \min x_i} \mathbb{1}_{\max x_i \leq \theta} \stackrel{\mathbb{P}_\theta - p.s.}{=} \begin{cases} 0 & \text{si } \theta < \max x_i \\ \frac{1}{\theta^n} & \text{si } \theta \geq \max x_i \end{cases},$$

ce qui admet un unique maximum global en $\theta = \max_i x_i$. Ainsi, l'estimateur du maximum de vraisemblance est unique et s'écrit $\hat{\theta}_2 = \max_i X_i$.

3. Calculer le biais, la variance et le risque quadratique de $\hat{\theta}_1$ et de $\hat{\theta}_2$. On calculera la densité de $\hat{\theta}_2$. Solution. Pour $\hat{\theta}_1$, on a $\mathbb{E}_\theta[\hat{\theta}_1] = \theta$ pour tout $\theta > 0$, donc l'estimateur est non biaisé. Sa variance est

$$\frac{4}{n} \text{Var}_\theta(X_1) = \frac{\theta^2}{3n},$$

d'où

$$R_\theta(\hat{\theta}_1) = 0 + \frac{\theta^2}{3n} = \frac{\theta^2}{3n}.$$

Pour $\hat{\theta}_2$, on commence par calculer sa densité g_θ . Par un calcul classique via la fonction de répartition, on obtient :

$$g_\theta(x) = \frac{nx^{n-1}}{\theta^n} \mathbb{1}_{[0,\theta]}(x).$$

Cela donne :

$$\mathbb{E}_\theta[\hat{\theta}_2] = \frac{1}{\theta^n} \int_0^\theta nx^n dx = \frac{n}{n+1} \theta, \quad \text{et} \quad \mathbb{E}_\theta[\hat{\theta}_2^2] = \frac{1}{\theta^n} \int_0^\theta nx^{n+2} dx = \frac{n}{n+2} \theta^2.$$

On en déduit :

$$b_{\theta}(\hat{\theta}_2) = -\frac{\theta}{n+1}, \quad \text{et} \quad \text{Var}_{\theta}(\hat{\theta}_2) = \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 = \frac{n}{(n+1)^2(n+2)}\theta^2,$$

ce qui donne finalement :

$$R_{\theta}(\hat{\theta}_2) = \left(\frac{\theta}{n+1}\right)^2 + \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{2}{(n+1)(n+2)}\theta^2.$$

4. Quel est le meilleur estimateur ? *Solution.* Il est facile de voir que $R_{\theta}(\hat{\theta}_2) \leq R_{\theta}(\hat{\theta}_1)$ dès que $n \geq 2$, donc $\hat{\theta}_2$ est meilleur que $\hat{\theta}_1$. En fait, il est même bien meilleur puisque lorsque $n \rightarrow \infty$, $R_{\theta}(\hat{\theta}_2) \sim \frac{2\theta^2}{n^2}$ alors que $R_{\theta}(\hat{\theta}_1) \sim \frac{\theta^2}{3n}$, ce qui est bien plus grand.

Exercice 2.3 (Estimation par maximum de vraisemblance dans le cas régulier). On considère un échantillon $X = (X_1, \dots, X_n)$ i.i.d. de loi η . Pour les modèles suivants, dont on admettra qu'ils sont réguliers, déterminer l'estimateur du maximum de vraisemblance, puis énoncer le théorème central limite associé.

1. η est la loi exponentielle de paramètre¹ $\lambda > 0$, et l'on veut estimer λ .
2. η est la loi de puissance à densité sur \mathbb{R} donnée par $x \mapsto \mathbf{1}_{x \geq 1} \frac{a-1}{x^a}$ avec $a > 1$ paramètre et l'on veut estimer a . *Solution.* On trouve $\hat{a} = 1 + \left(\frac{1}{n} \sum_{i=1}^n \log X_i\right)^{-1}$, en notant (ou pas) que $\log X_1 \sim \text{Exp}(a-1)$.

Exercice 2.4 (Information de Fisher via reparamétrisation). Soit $\{\mathbb{P}_{\theta}\}_{\theta \in \Theta}$ un modèle régulier unidimensionnel. Considérons une fonction différentiable bijective $h : \Lambda \rightarrow \Theta$. La famille $\{\mathbb{P}_{\theta}\}_{\theta \in \Theta}$ peut ainsi être reparamétrisée par $\{\mathbb{Q}_{\lambda}\}_{\lambda \in \Lambda}$. On note $\tilde{I}(\lambda)$ l'information de Fisher dans ce modèle reparamétrisé. Pour λ tel que $h(\lambda) = \theta$, donner $\tilde{I}(\lambda)$ en fonction de $I(\theta)$. Donner une interprétation physique de ce résultat. *Solution.* Notons que par définition \mathbb{Q}_{λ} est à densité $q_{\lambda} = p_{\theta}$. Ainsi

$$\begin{aligned} \tilde{I}(\lambda) &= \mathbb{E}_{\mathbb{Q}_{\lambda}} \left[\left(\frac{\partial}{\partial \lambda} \log q_{\lambda}(X) \right)^2 \right] = \mathbb{E}_{\mathbb{Q}_{\lambda}} \left[\left(\frac{\partial}{\partial \lambda} \log p_{h(\lambda)}(X) \right)^2 \right] \\ &= h'(\lambda)^2 \mathbb{E}_{\theta} \left[\left(\frac{\partial}{\partial \theta} \log p_{\theta}(X) \right)^2 \right] = h'(\lambda)^2 I(\theta). \end{aligned}$$

Si nous reprenons l'interprétation de l'information de Fisher comme une vitesse au carré (c'est le carré de la vitesse locale de la log-vraisemblance), cette variation s'adapte à celle induite par la transformation h . Si h a de fortes variations autour de λ (est très irrégulière), alors la vitesse en λ s'en verra affectée, et augmentera par rapport à la vitesse mesurée en θ . L'estimation de λ sera alors plus facile que celle de θ (cf Cramer-Rao).

Exercice 2.5 (Un modèle autorégressif). Nous considérons un modèle avec (X_1, \dots, X_n) tels que $X_1 = \theta + Z_1$, et récursivement, $X_{j+1} = \rho(X_j - \theta) + \theta + Z_{j+1}$ avec les Z_j i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Nous supposons que $\rho \in [0, 1]$ et qu'il est connu. On admettra la régularité de ce modèle.

1. Calculer la matrice d'information de Fisher $I(\theta, \sigma^2)$ en fonction de θ, σ^2, ρ et n . *Solution.* Le modèle est markovien, et log-vraisemblance s'écrit :

$$\ell(\theta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{(X_1 - \theta)^2}{2\sigma^2} - \frac{1}{2\sigma^2} \sum_{j=1}^{n-1} (X_{j+1} - \theta - \rho(X_j - \theta))^2.$$

¹sa densité sur \mathbb{R}_+ est donnée par $x \mapsto \lambda e^{-\lambda x}$.

Son gradient vaut

$$\nabla_{\theta, \sigma^2} \ell(\theta, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} \left[(X_1 - \theta) + (\rho + 1) \sum_{j=1}^{n-1} (X_{j+1} - \theta - \rho(X_j - \theta)) \right] \\ \frac{1}{2\sigma^2} \left[-n + \frac{(X_1 - \theta)^2}{\sigma^2} + \frac{1}{\sigma^2} \sum_{j=1}^{n-1} (X_{j+1} - \theta - \rho(X_j - \theta))^2 \right] \end{pmatrix}.$$

Pour calculer la matrice de covariance associée, il peut être bon de le reformuler en fonction des Z_j :

$$\nabla_{\theta, \sigma^2} \ell(\theta, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} \left[Z_1 + (1 - \rho) \sum_{j=2}^n Z_j \right] \\ \frac{1}{2\sigma^2} \left[-n + \frac{1}{\sigma^2} \sum_{j=1}^n Z_j^2 \right] \end{pmatrix}.$$

En utilisant que les Z_j sont i.i.d., que moments impairs d'une gaussienne centrée sont nuls, et $\text{Var}(Z_1^2) = \sigma^4(3 - 1^2) = 2\sigma^4$, on a :

$$I(\theta, \sigma^2) = \begin{pmatrix} \frac{1+(n-1)(1-\rho)^2}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Notons qu'elle ne dépend pas de θ : l'estimation de θ est de la même difficulté partout.

- Donner une borne inférieure pour la variance d'un estimateur sans biais de θ . Commenter et interpréter les valeurs de cette borne lorsque $\rho = 0$, et lorsque $\rho = 1$. *Solution.* La borne de Cramér-Rao pour un estimateur sans biais $\hat{\theta}$ de θ est l'inverse de l'information de Fisher en θ , donc :

$$\text{Var}(\hat{\theta}) \geq [I(\theta, \sigma^2)^{-1}]_{1,1} = \frac{\sigma^2}{1 + (n-1)(1-\rho)^2}.$$

On retrouve d'ailleurs la borne classique de Cramer-Rao lorsque $\rho = 0$, les variables sont alors i.i.d. Lorsque $\rho = 1$, en reprenant l'énoncé on $X_{j+1} = X_j + Z_{j+1}$, le bruit ne fait que s'ajouter, et on ne peut pas estimer σ^2 mieux qu'avec X_1 , donc avec erreur σ^2 , qui ne décroît pas en n . De façon générale, cette borne croît avec ρ , on estime plus difficilement θ quand ρ augmente.