

12 Estimation et borne de Cramer-Rao

Il est courant, en environnement stochastique, de vouloir estimer la valeur d'un paramètre θ figurant dans la loi de probabilité d'un phénomène aléatoire à l'aide d'une fonction \mathbf{T} dépendant de tirages effectués sur le phénomène étudié. On se pose alors :

- (a) d'une part la question de savoir si la fonction \mathbf{T} que l'on utilise pour estimer le paramètre θ est ou non **biaisé**, ce qui revient à comparer la valeur du paramètre θ à l'espérance $\mathbb{E}[\mathbf{T}]$ de l'estimateur,
- (b) d'autre part la question de savoir si la fonction \mathbf{T} constitue un estimateur **efficace**, c'est à dire de dispersion $\mathbb{V}[\mathbf{T}]$ aussi petite que possible.

L'intérêt de l'inégalité de Cramer-Rao est qu'elle fournit, pour une large classe d'estimateurs non biaisés, une borne inférieure pour la variance $\mathbb{V}[\mathbf{T}]$ et donc une mesure de l'efficacité de \mathbf{T} . On consultera la référence [26] pour une description complète de la problématique de l'estimation.

12.1 Cas scalaire

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, et soit \mathbf{w} une variable aléatoire définie sur Ω à valeurs dans \mathbb{R} , dont la loi admet une densité p qui dépend d'un paramètre $\theta \in \Theta \subset \mathbb{R}$. On note E_θ le support de la densité p pour la valeur θ :

$$E_\theta = \{w \in \mathbb{R}, p(w, \theta) > 0\} .$$

Définition 3. On appelle *quantité d'information de Fisher* fournie par la variable aléatoire \mathbf{w} sur le paramètre θ , notée $\mathcal{I}_{\mathbf{w}}(\theta)$, la quantité (si elle existe) :

$$\mathcal{I}_{\mathbf{w}}(\theta) = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ln p(\mathbf{w}, \theta) \right)^2 \right] = \int_{E_\theta} \left(\frac{\partial}{\partial \theta} \ln p(w, \theta) \right)^2 p(w, \theta) dw . \quad (109)$$

Proposition 8. Si le support E_θ ne dépend pas de θ , pour p suffisamment régulière, on a :

$$\mathcal{I}_{\mathbf{w}}(\theta) = \mathbb{V} \left[\frac{\partial}{\partial \theta} \ln p(\mathbf{w}, \theta) \right] = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \ln p(\mathbf{w}, \theta) \right] . \quad (110)$$

Preuve. Par définition, on a :

$$\int_{E_\theta} p(w, \theta) dw = 1 \quad \forall \theta \in \Theta .$$

- (a) Si la dérivation sous le signe somme est licite, on peut écrire :

$$0 = \frac{d}{d\theta} \int_{E_\theta} p(w, \theta) dw = \int_{E_\theta} \frac{\partial}{\partial \theta} p(w, \theta) dw = \int_{E_\theta} \frac{\partial}{\partial \theta} \ln p(w, \theta) p(w, \theta) dw = \mathbb{E} \left[\frac{\partial}{\partial \theta} \ln p(\mathbf{w}, \theta) \right] .$$

La variable aléatoire $\frac{\partial}{\partial \theta} \ln p(\mathbf{w}, \theta)$ est donc de moyenne nulle, d'où la première égalité.

- (b) Partant de :

$$\int_{E_\theta} \frac{\partial}{\partial \theta} \ln p(w, \theta) p(w, \theta) dw = 0 ,$$

et dérivant une seconde fois sous le signe somme, on obtient :

$$0 = \frac{d}{d\theta} \int_{E_\theta} \frac{\partial}{\partial \theta} \ln p(w, \theta) p(w, \theta) dw = \int_{E_\theta} \frac{\partial^2}{\partial \theta^2} \ln p(w, \theta) p(w, \theta) dw + \int_{E_\theta} \left(\frac{\partial}{\partial \theta} \ln p(w, \theta) \right)^2 p(w, \theta) dw ,$$

ce qui fournit la deuxième égalité et termine la preuve de la proposition. □

On déduit de cette première proposition le résultat (évident) suivant, qui constitue un début de justification au nom de “quantité d’information” donné à $\mathcal{I}_{\mathbf{w}}(\theta)$.

Proposition 9. *Soit $(\mathbf{w}_1, \dots, \mathbf{w}_n)$ un n -échantillon de la variable aléatoire \mathbf{w} (v.a. indépendantes et identiquement distribuées). Si le support E_θ ne dépend pas du paramètre θ , on a :*

$$\mathcal{I}_{\mathbf{w}_1, \dots, \mathbf{w}_n}(\theta) = n\mathcal{I}_{\mathbf{w}}(\theta) . \quad (111)$$

On s’intéresse alors au problème d’estimer le paramètre θ à partir d’observations faites sur la variable aléatoire \mathbf{w} . Le théorème suivant montre que l’information de Fisher permet de mesurer la qualité de tels estimateurs.

Théorème 14. *Soit \mathbf{w} une variable aléatoire définie sur Ω à valeurs dans \mathbb{R} , dont la densité p dépend d’un paramètre $\theta \in \Theta \subset \mathbb{R}$. On suppose que le support E_θ de la densité p ne dépend pas de θ . On considère un n -échantillon $(\mathbf{w}_1, \dots, \mathbf{w}_n)$ de la variable aléatoire \mathbf{w} et on se donne un estimateur $\mathbf{T} = \varphi(\mathbf{w}_1, \dots, \mathbf{w}_n)$ du paramètre θ basé sur ce n -échantillon. Alors, sous les conditions d’existence et de régularité adéquates, on dispose de l’inégalité suivante, dite de Cramer-Rao :*

$$\mathbb{V}[\mathbf{T}] \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}[\mathbf{T}]\right)^2}{n\mathcal{I}_{\mathbf{w}}(\theta)} . \quad (112)$$

Preuve. On note \mathbf{v} le n -échantillon (w_1, \dots, w_n) dont la densité se met sous la forme :

$$L(\mathbf{v}, \theta) = \prod_{i=1}^n p(w_i, \theta) ,$$

avec $\mathbf{v} = (w_1, \dots, w_n) \in \mathbb{R}^n$.

(a) Comme on suppose que la dérivation sous le signe somme est légitime, on a :

$$\int_{E_\theta^n} L(\mathbf{v}, \theta) d\mathbf{v} = 1 \implies 0 = \int_{E_\theta^n} \frac{\partial}{\partial \theta} L(\mathbf{v}, \theta) d\mathbf{v} = \int_{E_\theta^n} \frac{\partial}{\partial \theta} (\ln L(\mathbf{v}, \theta)) L(\mathbf{v}, \theta) d\mathbf{v} = \mathbb{E} \left[\frac{\partial}{\partial \theta} \ln L(\mathbf{v}, \theta) \right] ,$$

ce qui prouve que la variable aléatoire $\frac{\partial}{\partial \theta} \ln L(\mathbf{v}, \theta)$ est de moyenne nulle.

(b) Par un raisonnement de même type, on a que :

$$\frac{d}{d\theta} \mathbb{E}[\mathbf{T}] = \int_{E_\theta^n} \varphi(\mathbf{v}) \frac{\partial}{\partial \theta} L(\mathbf{v}, \theta) d\mathbf{v} = \mathbb{E} \left[\mathbf{T} \frac{\partial}{\partial \theta} \ln L(\mathbf{v}, \theta) \right] .$$

Il résulte de (a) et (b) que :

$$\frac{d}{d\theta} \mathbb{E}[\mathbf{T}] = \mathbb{E} \left[(\mathbf{T} - \mathbb{E}[\mathbf{T}]) \frac{\partial}{\partial \theta} \ln L(\mathbf{v}, \theta) \right] ,$$

et on obtient la relation (112) par application de l’inégalité de Schwartz. □

Ce théorème permet d’assurer que la variance d’un estimateur reste toujours supérieure à une borne ne dépendant que de la variable aléatoire de départ, pourvu que l’espérance de l’estimateur soit connue. L’exemple standard d’application est celui où l’on dispose d’un estimateur **sans biais** \mathbf{T} d’une fonction donnée h dépendant de θ . Alors,

$$\mathbb{E}[\mathbf{T}] = h(\theta) \implies \frac{d}{d\theta} \mathbb{E}[\mathbf{T}] = h'(\theta) ,$$

et l’inégalité de Cramer-Rao devient :

$$\mathbb{V}[\mathbf{T}] \geq \frac{(h'(\theta))^2}{n\mathcal{I}_{\mathbf{w}}(\theta)} . \quad (113)$$

Remarque 33. On notera que la fonction φ ne doit pas dépendre du paramètre θ . Ainsi, l'estimateur constant $\varphi(w_1, \dots, w_n) = \theta$ est de variance nulle, mais le théorème de Cramer-Rao ne s'applique pas. . . \square

Définition 4. On dit qu'un estimateur \mathbf{T} est **efficace** lorsque sa variance atteint la borne de Cramer-Rao.

Exemple. Soit \mathbf{w} une variable aléatoire réelle suivant une loi normale de moyenne μ et d'écart-type σ . On montre facilement que l'estimateur de Monte-Carlo basé sur le n -échantillon $(\mathbf{w}_1, \dots, \mathbf{w}_n)$:

$$\mathbf{T} = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i ,$$

est un estimateur *sans biais* et *efficace* de la moyenne μ .

12.2 Cas général

On considère maintenant un vecteur aléatoire \mathbf{w} défini sur Ω à valeurs dans \mathbb{R}^m dont la densité p , dépendant d'un paramètre $\theta \in \Theta \subset \mathbb{R}^d$, est telle que son support $E_\theta = \{u \in \mathbb{R}^m, p(u, \theta) > 0\}$ soit indépendant de θ . On définit alors la quantité d'information fournie par le vecteur aléatoire \mathbf{w} sur le paramètre θ comme étant la matrice de covariance $\mathcal{I}_{\mathbf{w}}(\theta)$ du vecteur aléatoire $\frac{\partial}{\partial \theta} \ln p(\mathbf{w}, \theta)$ à valeurs dans \mathbb{R}^d . On montre que, sous des hypothèses de régularité suffisantes, les éléments de cette matrice vérifient une propriété analogue à celle de la proposition 8, à savoir :

$$[\mathcal{I}_{\mathbf{w}}(\theta)]_{i,j} = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta_i} \ln p(\mathbf{w}, \theta) \right) \left(\frac{\partial}{\partial \theta_j} \ln p(\mathbf{w}, \theta) \right) \right] = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln p(\mathbf{w}, \theta) \right] , \quad (114)$$

et que la propriété d'additivité (proposition 9) reste vraie.

Théorème 15. Soit $(\mathbf{w}_1, \dots, \mathbf{w}_n)$ un n -échantillon de la variable \mathbf{w} , soit (h_1, \dots, h_s) s fonctions réelles dépendant de la variable θ et soit $(\mathbf{T}_1, \dots, \mathbf{T}_s)$ un estimateur sans biais de $(h_1(\theta), \dots, h_s(\theta))$ basé sur le n -échantillon $(\mathbf{w}_1, \dots, \mathbf{w}_n)$. On note :

- $\mathbb{V}[\mathbf{T}]$ la matrice de covariance du vecteur aléatoire $(\mathbf{T}_1, \dots, \mathbf{T}_s)$,
- $H'(\theta)$ la matrice des dérivées de la fonction $(h_1, \dots, h_s) : (H'(\theta))_{i,j} = \frac{\partial}{\partial \theta_j} h_i(\theta)$,

et on suppose que la matrice $\mathcal{I}_{\mathbf{w}}(\theta)$ est inversible. Alors, sous les conditions d'existence et de régularité adéquates, on a (au sens des matrices semi-définies positives) :

$$\mathbb{V}[\mathbf{T}] - \frac{1}{n} H'(\theta) (\mathcal{I}_{\mathbf{w}}(\theta))^{-1} (H'(\theta))^\top \geq 0 . \quad (115)$$

Preuve. Soit \mathbf{v} le n -échantillon $(\mathbf{w}_1, \dots, \mathbf{w}_n)$, de densité $L(\mathbf{v}, \theta) = \prod_{i=1}^n p(u_i, \theta)$, avec $\mathbf{v} = (u_1, \dots, u_n)$. On a montré que les variables aléatoires $\frac{\partial}{\partial \theta_j} \ln L(\mathbf{v}, \theta)$ sont de moyenne nulle et que l'on a :

$$\mathbb{E}[\mathbf{T}_i] = h_i(\theta) \implies \frac{\partial}{\partial \theta_j} h_i(\theta) = \mathbb{E} \left[\mathbf{T}_i \frac{\partial}{\partial \theta_j} \ln L(\mathbf{v}, \theta) \right] .$$

La matrice de covariance K du vecteur $(\mathbf{T}_1, \dots, \mathbf{T}_s, \frac{\partial}{\partial \theta_1} \ln L(\mathbf{v}, \theta), \dots, \frac{\partial}{\partial \theta_d} \ln L(\mathbf{v}, \theta))$ se met sous la forme :

$$K = \begin{pmatrix} \mathbb{V}[\mathbf{T}] & H'(\theta) \\ H'(\theta)^\top & n\mathcal{I}_{\mathbf{w}}(\theta) \end{pmatrix} .$$

Une matrice de covariance étant toujours semi-définie positive, on en déduit le résultat par la propriété :

$$z^\top K z \geq 0 ,$$

avec le choix :

$$z = \begin{pmatrix} t \\ -\frac{1}{n}(\mathcal{I}_w(\theta))^{-1}(H'(\theta))^\top t \end{pmatrix} .$$

□

Exemple. Soit w une variable aléatoire à valeurs dans \mathbb{R}^m suivant une loi normale de moyenne M et de matrice de covariance Γ . La densité de cette variable aléatoire est :

$$p(w, M) = \frac{1}{(2\pi)^{m/2} \sqrt{\det \Gamma}} \exp \left(-\frac{1}{2}(w - M)^\top \Gamma^{-1}(w - M) \right) .$$

Considérant la loi de la variable aléatoire w comme étant paramétrée par sa moyenne M , on montre facilement que la quantité d'information de Fisher fournie par w sur θ est :

$$\mathcal{I}_w(M) = \Gamma^{-1} .$$

On a alors que l'estimateur de Monte Carlo de la moyenne M est un estimateur efficace.

Remarque 34. **Le résultat du dernier exemple n'est pas vrai en toute généralité.** Si l'on considère une variable aléatoire w de moyenne M , de matrice de covariance Γ , *non gaussienne*, dont la densité $p(w, M)$ est paramétrée par sa moyenne, on a seulement l'inégalité :

$$\Gamma - (\mathcal{I}_w(M))^{-1} \geq 0 .$$

Quand l'inégalité est stricte, l'estimateur de Monte Carlo de la moyenne n'est pas efficace. □

12.3 Maximum de vraisemblance

On considère une variable aléatoire w dont la densité $p(w, \theta^\#)$ dépend d'un paramètre $\theta^\#$ que l'on cherche à déterminer. On suppose que le support de la fonction $p(\cdot, \theta)$ est indépendant de θ , et on se donne un n -échantillon (w_1, \dots, w_n) de w . On appelle *vraisemblance* de la réalisation (w_1, \dots, w_n) du n -échantillon la fonction L définie par :

$$L(w_1, \dots, w_n, \theta) = \prod_{i=1}^n p(w_i, \theta) .$$

La méthode du maximum de vraisemblance consiste à déterminer la valeur $\hat{\theta}_n$ qui maximise la vraisemblance L . Supposant que la fonction L est deux fois continûment différentiable en θ et que la matrice $-\nabla_{\theta, \theta}^2 \ln L$ (formée des dérivées secondes en θ du logarithme de L) est définie positive, on sait que $\hat{\theta}_n$ est solution de l'équation de la vraisemblance :

$$0 = \frac{\partial}{\partial \theta} \ln (L(w_1, \dots, w_n, \hat{\theta}_n)) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln (p(w_i, \hat{\theta}_n)) .$$

La valeur $\hat{\theta}_n$ est elle-même la réalisation d'une variable aléatoire $\hat{\theta}_n$, et on montre les résultats suivants (voir [26] pour plus de détails).

1. **Convergence presque sûre.** La suite $\{\widehat{\boldsymbol{\theta}}_n\}_{n \in \mathbb{N}}$ des solutions de l'équation de vraisemblance converge presque sûrement vers $\boldsymbol{\theta}^\sharp$, valeur du paramètre intervenant dans la loi de \boldsymbol{w} .
2. **Normalité asymptotique.** La suite $\{\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\sharp)\}_{n \in \mathbb{N}}$ converge en loi vers une loi normale centrée dont la matrice de covariance est l'inverse de la matrice d'information de Fisher :

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^\sharp) \xrightarrow{\mathcal{L}} \mathcal{N}(0, (\mathcal{I}_{\boldsymbol{w}}(\boldsymbol{\theta}^\sharp))^{-1}).$$

De ces deux résultats, on déduit que la solution $\widehat{\boldsymbol{\theta}}_n$ de l'équation de vraisemblance est asymptotiquement un estimateur sans biais efficace du paramètre $\boldsymbol{\theta}^\sharp$.

Remarque 35. Il existe un algorithme permettant de calculer de manière récursive la suite $\{\widehat{\boldsymbol{\theta}}_n\}_{n \in \mathbb{N}}$ des estimateurs du maximum de vraisemblance. Cet algorithme, basé sur le gradient stochastique, a été décrit au § 2.5.

