

Inférence statistique

Cours 3 : Maximum de Vraisemblance

Michal W. Urdanivia^{*}

^{*}UGA, Faculté d'Économie, GAEL,
e-mail : michal.wong-urdanivia@univ-grenoble-alpes.fr

14 septembre 2022

Plan

1. Vraisemblance
2. Maximum de vraisemblance
3. Interprétation du principe du MV

Plan

1. Vraisemblance

2. Maximum de vraisemblance

3. Interprétation du principe du MV

1. Vraisemblance

Vraisemblance, cas discret

- Soit un modèle statistique $(\mathcal{E}, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$ associé à un échantillon de v.a. i.i.d, X_1, X_2, \dots, X_n . Supposons que \mathcal{E} est discret (i.e., fini, ou dénombrable).
- **Définition** : la **vraisemblance** du modèle est l'application \mathcal{L}_n (ou seulement \mathcal{L}) définie par :

$$\begin{aligned}\mathcal{L}_n : \mathcal{E}^n \times \Theta &\rightarrow \mathbb{R} \\ (x_1, x_2, \dots, x_n, \theta) &\mapsto P_\theta(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n).\end{aligned}$$

Vraisemblance, cas discret

● Exemples :

1. Tirages de Bernoulli : si $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Ber}(p)$ pour un $p \in (0, 1)$:
 - $\mathcal{E} = \{0, 1\}$,
 - $\Theta = (0, 1)$,
 - $\forall (x_1, \dots, x_n) \in \{0, 1\}^n, \forall p \in (0, 1)$:

$$\begin{aligned}\mathcal{L}(x_1, \dots, x_n, p) &= \prod_{i=1}^n P_p(X_i = x_i) \\ &= \prod_{i=1}^n p^{x_i} (1-p)^{(1-x_i)} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{(n - \sum_{i=1}^n x_i)}.\end{aligned}$$

Vraisemblance, cas discret

2. Modèle de Poisson : si $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{Pois}(\lambda)$, pour un $\lambda > 0$:

- $\mathcal{E} = \mathbb{N}$,
- $\Theta = (0, \infty)$,
- $\forall (x_1, x_2, \dots, x_n) \in \mathbb{N}^n, \forall \lambda > 0$:

$$\begin{aligned}\mathcal{L}(x_1, x_2, \dots, x_n, \lambda) &= \prod_{i=1}^n P_\lambda(X_i = x_i) \\ &= \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! \dots x_n!}.\end{aligned}$$

Vraisemblance, cas continu

- Soit un modèle statistique $(\mathcal{E}, \mathcal{F}, (P)_{\theta \in \Theta})$ associé à un échantillon de v.a. i.i.d, X_1, X_2, \dots, X_n . Supposons que tous les P_θ ont une densité f_θ par rapport à la mesure de Lebesgue.
- **Définition** : la vraisemblance du modèle est l'application \mathcal{L} définie par :

$$\begin{aligned}\mathcal{L} : \mathcal{E}^n \times \Theta &\rightarrow \mathbb{R} \\ (x_1, x_2, \dots, x_n, \theta) &\mapsto \prod_{i=1}^n f_\theta(x_i).\end{aligned}$$

Vraisemblance, cas continu

● Exemples :

1. **Modèle Gaussien** : Si $X_1, X_2, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, pour un $\mu \in \mathbb{R}$, et $\sigma^2 > 0$:

- $\mathcal{E} = \mathbb{R}$,
- $\Theta = \mathbb{R} \times \mathbb{R}_+^*$,
- $\forall (x_1, x_2, \dots, x_n) \in \mathbb{R}^n, \forall (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$:

$$\mathcal{L}(x_1, x_2, \dots, x_n, \mu, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right).$$

Plan

1. Vraisemblance

2. Maximum de vraisemblance

3. Interprétation du principe du MV

2. Maximum de vraisemblance

Estimateur du maximum de vraisemblance

- Soit X_1, X_2, \dots, X_n un échantillon i.i.d. associé au modèle statistique $(\mathcal{E}, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$ et soit \mathcal{L} la vraisemblance du modèle.
- **Définition** : l'estimateur du maximum de vraisemblance est défini par,

$$\hat{\theta}_n^{MV} = \arg \max_{\theta \in \Theta} \mathcal{L}(X_1, X_2, \dots, X_n, \theta),$$

dès lors qu'il existe.

- **Remarque(log-vraisemblance)** : en pratique, on utilise le fait que,

$$\hat{\theta}_n^{MV} = \arg \max_{\theta \in \Theta} \log \mathcal{L}(X_1, X_2, \dots, X_n, \theta).$$

Estimateur du maximum de vraisemblance

● Exemples

- Tirages de Bernoulli : $\hat{p}_n^{MV} = \bar{X}_n$.
- Modèle de Poisson : $\hat{\lambda}_n^{MV} = \bar{X}_n$.
- Modèle Gaussien : $(\hat{\mu}_n, \hat{\sigma}_n^2) = (\bar{X}_n, \hat{S}_n)$.

Information de Fisher(Définition)

- Définissons la log-vraisemblance d'une observation par :

$$\ell(\theta) = \log \mathcal{L}(X, \theta), \quad \theta \in \Theta.$$

- Supposons que ℓ est p.s. doublement différentiable.
- Sous certaines conditions de régularité, l'**information de Fisher** du modèle statistique est définie par :

$$\mathcal{I}(\theta) = \mathbb{V}_{\theta}(\nabla_{\theta} \ell(\theta)) = -\mathbb{E}_{\theta} \left(\frac{\partial^2 \ell}{\partial \theta \partial \theta'}(\theta) \right).$$

Estimateur du MV : théorème

- Soit $\theta^* \in \Theta$ (le **vrai paramètre**). Supposons les conditions suivantes vérifiées :
 1. Le modèle est identifié.
 2. Pour tout $\theta \in \Theta$, le support de P_θ ne dépend pas de θ .
 3. θ^* n'est pas sur la limite de Θ .
 4. $\mathcal{I}(\theta)$ est inversible dans un voisinage de θ^* .
 5. Quelques conditions techniques additionnelles.
- Alors, $\hat{\theta}_n^{MV}$ vérifie :
 - $\hat{\theta}_n^{MV} \xrightarrow{P} \theta^*$, par rapport à P_{θ^*} ;
 - $\sqrt{n}(\hat{\theta}_n^{MV} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1})$, par rapport à P_{θ^*} .

Plan

1. Vraisemblance

2. Maximum de vraisemblance

3. Interprétation du principe du MV

3. Interprétation du principe du MV

Distance en variation totale

- On considère le modèle statistique $(\mathcal{E}, \mathcal{F}, (P_\theta)_{\theta \in \Theta})$ associé à l'échantillon de v.a. i.i.d X_1, X_2, \dots, X_n .
- Le vrai paramètre du modèle est $\theta^* \in \Theta$, i.e., $X \sim P_{\theta^*}$.
- **Objectif du statisticien** : étant donné X_1, X_2, \dots, X_n trouver/construire un estimateur $\hat{\theta} := \hat{\theta}(X_1, X_2, \dots, X_n)$ tel que $P_{\hat{\theta}}$ soit proche de P_{θ^*} pour la vraie valeur du paramètre θ^* .
- Ceci signifie que l'on souhaite que $|P_{\hat{\theta}}(A) - P_{\theta^*}(A)|$ soit petit pour tout $A \subset \mathcal{F}$

Distance en variation totale

- **Définition** : la **distance en variation totale(DVT)** entre deux mesures de probabilité P_θ et $P_{\theta'}$ est définie par,

$$d_{vt}(P_\theta, P_{\theta'}) = \max_{A \in \mathcal{F}} |P_\theta(A) - P_{\theta'}(A)|.$$

Distance en variation totale

- Supposons que \mathcal{E} soit discret (i.e., fini ou dénombrable). Ceci inclut les v.a., suivant des lois de Bernoulli, binomiales, de Poisson,...
- Dans ce cas, X a une **fonction de masse** :

$$P_{\theta}(X = x) =: p_{\theta}(x) \quad \text{pour tout } x \in \mathcal{E}, \quad p_{\theta}(x) > 0, \quad \sum_{x \in \mathcal{E}} p_{\theta}(x) = 1.$$

- Alors la DVT entre P_{θ} et $P_{\theta'}$ est simplement une fonction des fonctions de masse p_{θ} et $p_{\theta'}$:

$$d_{vt}(P_{\theta}, P_{\theta'}) = \frac{1}{2} \sum_{x \in \mathcal{E}} |p_{\theta}(x) - p_{\theta'}(x)|.$$

Distance en variation totale

- Supposons que \mathcal{E} est continu, par exemple pour de v.a. gaussiennes ou exponentielles, ...
- Supposons aussi que la densité de X pour tout $A \in \mathcal{F}$ est donnée par,

$$P_{\theta}(X \in A) = \int_A f_{\theta}(x)dx, \quad f_{\theta}(x) \geq 0, \quad \int_{\mathcal{F}} f_{\theta}(x)dx = 1.$$

- La DVT entre P_{θ} et $P_{\theta'}$ est simplement fonction des densités f_{θ} et $f_{\theta'}$:

$$d_{vt}(P_{\theta}, P_{\theta'}) = \frac{1}{2} \int_{x \in \mathcal{E}} |f_{\theta}(x) - f_{\theta'}(x)|.$$

Propriétés de la DVT

- i) $d_{vt}(P_\theta, P_{\theta'}) = d_{vt}(P_{\theta'}, P_\theta)$ (symétrie).
 - ii) $d_{vt}(P_\theta, P_{\theta'}) \geq 0$.
 - iii) $d_{vt}(P_\theta, P_{\theta'}) = 0$ ssi $P_\theta = P_{\theta'}$
 - iv) $d_{vt}(P_\theta, P_{\theta'}) \leq d_{vt}(P_\theta, P_{\theta''}) + d_{vt}(P_{\theta''}, P_{\theta'})$ (inégalité triangulaire).
- La DTV est ainsi une **distance** entre lois de probabilité.

Stratégie d'estimation

- Construire un estimateur $\hat{d}_{vt}(P_\theta, P_{\theta^*})$ pour tout $\theta \in \Theta$.
- Obtenir alors $\hat{\theta}$ qui minimise la fonction $\theta \mapsto \hat{d}_{vt}(P_\theta, P_{\theta^*})$.
- **Problème** : construire $\hat{d}_{vt}(P_\theta, P_{\theta^*})$ n'est pas évident !

Divergence de Kullback-Leibler

- Nous pouvons essayer de remplacer la DVT par une autre distance entre mesures de probabilité.
- Une, communément retenue car pratique est la **divergence de Kullback-Leibler**
- **Définition** : la divergence de Kullback-Leibler(KL) entre deux mesures de probabilité P_θ et $P_{\theta'}$ est définie par,

$$KL(P_\theta, P_{\theta'}) = \begin{cases} \sum_{x \in \mathcal{E}} p_\theta(x) \log \left(\frac{p_\theta(x)}{p_{\theta'}(x)} \right) & \text{lorsque } \mathcal{E} \text{ est discret} \\ \int_{\mathcal{E}} f_\theta(x) \log \left(\frac{f_\theta(x)}{f_{\theta'}(x)} \right) dx & \text{lorsque } \mathcal{E} \text{ est continu.} \end{cases}$$

Propriétés de la divergence de KL

- i) $KL(P_\theta, P_{\theta'}) \neq KL(P_{\theta'}, P_\theta)$, en général.
 - ii) $KL(P_\theta, P_{\theta'}) \geq 0$.
 - iii) $KL(P_\theta, P_{\theta'}) = 0$ ssi $P_\theta = P_{\theta'}$.
 - iv) $KL(P_\theta, P_{\theta'}) \not\leq KL(P_\theta, P_{\theta''}) + KL(P_{\theta''}, P_{\theta'})$ en général.
- Ce n'est donc pas une distance, d'où le nom de divergence.
 - Et l'asymétrie sera la clé pour pouvoir l'estimer !

MV et KL

- On considère le cas où \mathcal{E} est discret.
- La divergence entre P_{θ^*} et P_{θ} est,

$$\begin{aligned}\text{KL}(P_{\theta^*}, P_{\theta}) &= E_{\theta^*} \left[\log \left(\frac{p_{\theta^*}(X)}{p_{\theta}(X)} \right) \right] \\ &= E_{\theta^*} [\log p_{\theta^*}(X)] - E_{\theta^*} [\log p_{\theta}(X)]\end{aligned}$$

MV et KL

- Ainsi, la fonction $\theta \mapsto \text{KL}(P_{\theta^*}, P_{\theta})$ est de la forme :

$$\text{constante} - E_{\theta^*} [\log p_{\theta}(X)]$$

.

- Un estimateur consistant de $E_{\theta^*} [\log p_{\theta}(X)]$ est $\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i)$ (LGN).

MV et KL

- Et un estimateur de la divergence sera,

$$\hat{KL}(P_{\theta^*}, P_{\theta}) = \text{constante} - \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i).$$

MV et KL

- Si nous revenons au problème de la minimisation de la distance entre les mesures P_{θ^*} et P_{θ} pour définir l'estimateur $\hat{\theta}$, alors d'après ce qui précède il s'agit de résoudre,

$$\min_{\theta \in \Theta} \hat{KL}(P_{\theta^*}, P_{\theta}) = \min_{\theta \in \Theta} \left[\text{constante} - \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \right]$$

MV et KL

- Et nous avons,

$$\begin{aligned}\min_{\theta \in \Theta} \left[\text{constante} - \frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \right] &\Leftrightarrow \min_{\theta \in \Theta} \left[-\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \right] \\ &\Leftrightarrow \max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i) \right] \\ &\Leftrightarrow \max_{\theta \in \Theta} \left[\sum_{i=1}^n \log p_{\theta}(X_i) \right] \\ &\Leftrightarrow \max_{\theta \in \Theta} \left[\log \left(\prod_{i=1}^n p_{\theta}(X_i) \right) \right] \\ &\Leftrightarrow \max_{\theta \in \Theta} \left[\prod_{i=1}^n p_{\theta}(X_i) \right]\end{aligned}$$

MV et KL

- Qui est donc le principe du MV !!