Estimation ponctuelle: famille exponentielle

Statistique mathématique
M2 santé publique, université Paris-Sud

10 novembre 2017

Zoom sur les familles paramétriques

Rappelons le point de départ :

- Collection de v.a (un vecteur aléatoire) $X = (X_1, \ldots, X_n)$
- $X \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ une famille paramétrique de paramètre $\theta \in \Theta \subseteq \mathbb{R}^d$

Zoom sur les familles paramétriques

Rappelons le point de départ :

- Collection de v.a (un vecteur aléatoire) $X = (X_1, \ldots, X_n)$
- $X \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ une famille paramétrique de paramètre $\theta \in \Theta \subseteq \mathbb{R}^d$

Le problème de l'estimation ponctuelle

- Supposons que F est complètement définie par son paramètre θ qui inconnu
- Soit $(x_1, \ldots, x_n)$ des réalisations de $X \sim F_\theta$
- Estimer la valeur de θ qui a *généré* les réalisations $(x_1, \ldots, x_n)$

Zoom sur les familles paramétriques

- Le jeu de données est l'unique information en notre possession (autre que la connaissance de la famille $\mathcal{F}$).
- Tout ce qu'on peut "faire" pour apprendre sur θ n'est rien d'autre qu'une fonction du jeu de données $g(x_1, \ldots, x_n)$.
- Jusqu'ici, on s'est intéressé au jeu de données : approximation de la loi + perte d'information ....
- Que peut-on dire à propos de $\mathcal{F}$?

Zoom sur les familles paramétriques

Nous décrivons $\mathcal{F}$ par la paramétrisation $\theta \in \Theta \mapsto F_\theta$:

Définition (paramétrisation)

Soit Θ un ensemble, $\mathcal{F}$ une famille (une collection) de lois et $g : \Theta \to \mathcal{F}$ une application surjective. Le couple (Θ, g) est appelé *paramétrisation* de $\mathcal{F}$.

$\to$ assigner un label $\theta \in \Theta$ pour chaque élément de $\mathcal{F}$.

Définition (Modèle paramétrique)

Un *modèle paramétrique* avec un espace paramétrique $\Theta \subseteq \mathbb{R}^d$ est une famille de lois de probabilité dont la paramétrisation est donnée par Θ, $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$.

Question

- Nous avons vu certains exemples de lois avec des propriétés individuelles ....
- Existe-t-il une **famille générale** de lois qui contient un certain nombre de lois classiques comme cas particuliers et dont les **propriétés peuvent être étudiées** ?

La famille exponentielle

Définition (famille exponentielle)

Soit $X = (X_1, \ldots, X_n)$ de loi jointe F_θ de paramètre $\theta \in \Theta \subseteq \mathbb{R}^p$. On dit que la famille de lois $\{F_\theta : \theta \in \Theta\}$ est une famille exponentielle à k paramètres si la densité jointe (ou la fonction de masse jointe) de $X = (X_1, \ldots, X_n)$ est de la forme

$$f(x;\theta) = \exp\left\{\sum_{i=1}^{k} c_i(\theta)\, T_i(x) - d(\theta) + S(x)\right\}, \quad x \in \mathcal{X}, \theta \in \Theta,$$

où $\text{supp}\{f(\cdot;\theta)\} = \mathcal{X}$ est indépendant de θ.

- Bien qu'ils coincident souvent ... k n'est pas nécessairement égal à p.
- On peut définir une nouvelle paramétrisation (re-paramétrisation) $\phi_i = c_i(\theta)$, c'est ce qu'on va appeler *paramètre naturel*.

Motivation et idée de la famille exponentielle

Considérons le problème variationnel suivant :

Maximum d'entropie sous contraintes

Déterminer une fonction de densité f de support $\mathcal{X}$ d'entropie maximale

$$H(f) = -\int_{\mathcal{X}} f(x) \log f(x) dx$$

sous les contraintes linéaires

$$\int_{\mathcal{X}} T_i(x) f(x) dx = \alpha_i, \quad i = 1, \ldots, k$$

Motivation et idée de la famille exponentielle

Considérons le problème variationnel suivant :

Maximum d'entropie sous contraintes

Déterminer une fonction de densité f de support $\mathcal{X}$ d'entropie maximale

$$H(f) = -\int_{\mathcal{X}} f(x) \log f(x) dx$$

sous les contraintes linéaires

$$\int_{\mathcal{X}} T_i(x) f(x) dx = \alpha_i, \quad i = 1, \ldots, k$$

Principe : comment choisir un modèle probabiliste pour une situation donnée ?
Approche par maximum d'entropie :

- Pour une situation donnée, choisir la loi de probabilité qui donne le maximum d'incertitude en respectant un certain nombre de contraintes.

Principe du maximum d'entropie

Proposition

Lorsque la solution du problème de maximisation sous contraintes existe, elle est unique et est de la forme *famille exponentielle*.

Preuve

La preuve fait appel aux multiplicateurs de Lagrange.

Quelques exemples

1. $X \sim \text{Binomial}(n, \theta)$.
2. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \Gamma(\alpha, \lambda)$.
3. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \theta^2)$.
4. $X \sim \mathcal{U}[0, \theta]$.

Exemple 1

Exemple : loi Binomiale

Soit $X \sim \text{Binomial}(n, \theta)$ avec n connu.

Exemple 1

Exemple : loi Binomiale

Soit $X \sim \text{Binomial}(n, \theta)$ avec n connu.

$$f(x;\theta) = C_n^x \theta^x (1-\theta)^{n-x}$$

$$= \exp\left[\underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{c(\theta)} \underbrace{x}_{T(x)} + \underbrace{n \ln(1-\theta)}_{d(\theta)} + \underbrace{\ln C_n^x}_{S(x)}\right]$$

Exemple 2

Exemple : loi Gamma

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \Gamma(\alpha, \lambda)$. Ici $\theta = (\alpha, \lambda)$, on sait que

$$f(x; \alpha, \lambda) = \prod_{i=1}^{n} \frac{\lambda^\alpha x_i^{\alpha-1} \exp\left(-\lambda x_i\right)}{\Gamma(\alpha)}$$

$$= \exp\left[\underbrace{(\alpha - 1)}_{c_1(\theta)} \underbrace{\sum_{i=1}^{n} \log x_i}_{T_1(x)} \underbrace{-\lambda}_{c_2(\theta)} \underbrace{\sum_{i=1}^{n} x_i}_{T_2(x)} + \underbrace{n\alpha \log \lambda - n \log \Gamma(\alpha)}_{d(\theta)}\right]$$

Exemple 2

Exemple : loi Gamma

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \Gamma(\alpha, \lambda)$. Ici $\theta = (\alpha, \lambda)$, on sait que

$$f(x; \alpha, \lambda) = \prod_{i=1}^{n} \frac{\lambda^\alpha x_i^{\alpha-1} \exp\left(-\lambda x_i\right)}{\Gamma(\alpha)}$$

$$= \exp\left[\underbrace{(\alpha - 1)}_{c_1(\theta)} \underbrace{\sum_{i=1}^{n} \log x_i}_{T_1(x)} \underbrace{-\lambda}_{c_2(\theta)} \underbrace{\sum_{i=1}^{n} x_i}_{T_2(x)} + \underbrace{n\alpha \log \lambda - n \log \Gamma(\alpha)}_{d(\theta)}\right]$$

Exemple 3

Exemple : loi normale

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \theta^2)$. On a

$$f(x;\theta) = \prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} \exp\left[-\frac{1}{2\theta^2}(x_i-\theta)^2\right]$$

$$= \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{n}{2}\big((1+2\log\theta)+\log(2\pi)\big)\right].$$

Notons que qu'ici $k = 2$ alors que l'espace paramétrique est de dimension 1.

Exemple 3

Exemple : loi normale

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \theta^2)$. On a

$$f(x;\theta) = \prod_{i=1}^{n} \frac{1}{\theta\sqrt{2\pi}} \exp\left[-\frac{1}{2\theta^2}(x_i-\theta)^2\right]$$

$$= \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 + \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{n}{2}\left((1+2\log\theta)+\log(2\pi)\right)\right].$$

Notons que qu'ici $k = 2$ alors que l'espace paramétrique est de dimension 1.

Exemple : le modèle uniforme

Soit $X \sim \mathcal{U}[0,\theta]$. Nous avons $f(x;\theta) = \dfrac{\mathbb{I}\{\theta \in [0,\theta]\}}{\theta}$. La loi uniforme $\mathcal{U}[0,\theta]$ ne fait pas partie de la famille exponentielle car le support de sa densité dépend de son paramètre.

La famille exponentielle

Proposition

Supposons que $X = (X_1, \ldots, X_n)$ a une loi de la famille exponentielle avec un seul paramètre

$$f(x;\theta) = \exp\Big[c(\theta)\,T(x) - d(\theta) + S(x)\Big]$$

pour $x \in \mathcal{X}$ où

(a) L'espace des paramètre Θ est un ouvert

(b) $c(\theta)$ est une bijection sur Θ

(c) $c(\theta)$, $c^{-1}(\theta)$ et $d(\theta)$ sont deux fois différentiables sur Θ.

Alors

$$\mathbb{E}\Big[T(X)\Big] = \frac{d'(\theta)}{c'(\theta)} \quad \& \quad \mathrm{Var}\Big[T(X)\Big] = \frac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{\big[c'(\theta)\big]^3}$$

Preuve de la proposition

On définit $\phi = c(\theta)$ (ce qu'on va appeler *paramètre naturel*) de la famille exponentielle. On pose $d_0(\phi) = d\big(c^{-1}(\phi)\big)$, d_0 est bien définie car c est une bijection. La continuité de c nous assure que $\Phi = c\big(\Theta\big)$ est un ouvert. Prenons s suffisamment petit tel que $\phi + s \in \Phi$, on remarque que la fonction génératrice de T est donnée par

$$\begin{aligned}
\mathbb{E}\Big[e^{sT(X)}\Big] &= \int e^{sT(x)} e^{\phi T(x) - d_0(\phi) + S(x)} dx \\
&= e^{d_0(\phi+s) - d_0(\phi)} \underbrace{\int e^{(\phi+s)T(x) - d_0(\phi+s) + S(x)} dx}_{=1} \\
&= e^{d_0(\phi+s) - d_0(\phi)}.
\end{aligned}$$

On dérive par rapport s et on pose $s = 0$, on obtient

$$\mathbb{E}\big[T(X)\big] = d_0'(\phi) \quad \text{et} \quad \mathrm{Var}\big[T(X)\big] = d_0''(\phi).$$

Or $d_0''(\phi) = \dfrac{d'(\theta)}{c'(\theta)}$ et $d_0''(\phi) = \dfrac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{\big[c'(\theta)\big]^3}$.

Famille exponentielle et exhaustivité

Lemme

Supposons que $X = (X_1, \ldots, X_n)$ est issu d'une loi de la famille exponentielle à k paramètres de densité

$$f(x;\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)\, T_i(x) - d(\theta) + S(x)\right]$$

pour $x \in \mathcal{X}$. Alors, la statistique $\big(T_1(x), \ldots, T_k(x)\big)$ est exhaustive pour θ.

Famille exponentielle et exhaustivité

Lemme

Supposons que $X = (X_1, \ldots, X_n)$ est issu d'une loi de la famille exponentielle à k paramètres de densité

$$f(x;\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)\, T_i(x) - d(\theta) + S(x)\right]$$

pour $x \in \mathcal{X}$. Alors, la statistique $\big(T_1(x), \ldots, T_k(x)\big)$ est exhaustive pour θ.

Il suffit de poser $g\big(T(x);\theta\big) = \exp\{\sum_i T_i(x) c_i(\theta) + d(\theta)\}$ et $h(x) = e^{S(x)}\mathbb{I}\{x \in \mathcal{X}\}$, et appliquer le théorème de factorisation.

Famille exponentielle et complétude

Théorème

Supposons que $X = (X_1, \ldots, X_n)$ est issu de la famille à k paramètres de densité

$$f(x;\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)\, T_i(x) - d(\theta) + S(x)\right]$$

pour $x \in \mathcal{X}$. On définit l'ensemble $C = \left\{\big(c_1(\theta), \ldots, c_k(\theta)\big) : \theta \in \Theta\right\}$. Si l'ensemble C contient un ouvert de la forme $]a_1, b_1[\times\cdots\times]a_k, b_k[$, alors la statistique $\big(T_1(x), \ldots, T_k(x)\big)$ est complète pour θ et donc exhaustive minimale.

Famille exponentielle et complétude

Théorème

Supposons que $X = (X_1, \ldots, X_n)$ est issu de la famille à k paramètres de densité

$$f(x;\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)\, T_i(x) - d(\theta) + S(x)\right]$$

pour $x \in \mathcal{X}$. On définit l'ensemble $C = \left\{\big(c_1(\theta), \ldots, c_k(\theta)\big) : \theta \in \Theta\right\}$. Si l'ensemble C contient un ouvert de la forme $]a_1, b_1[\times \cdots \times]a_k, b_k[$, alors la statistique $\big(T_1(x), \ldots, T_k(x)\big)$ est complète pour θ et donc exhaustive minimale.

- Ce résultat est essentiellement une conséquence de l'unicité de la fonction caractéristique.
- Intuitivement, ce résultat montre que la statistique exhaustive de dimension k de la famille exponentielle est complète. La dimension effective de l'espace paramétrique naturel est k.

Échantillon issu de la famille exponentielle

Soit $X_1, \ldots, X_n$ iid de loi issue de la famille exponentielle à k paramètres. Considérons la loi jointe de $X = (X_1, \ldots, X_n)$

$$f(x;\theta) = \prod_{j=1}^{n} \exp\left[\sum_{i=1}^{k} c_i(\theta)\, T_i(x_j) - d(\theta) + S(x_j)\right]$$

$$= \exp\left[\sum_{i=1}^{k} c_i(\theta)\tau_i(x) - nd(\theta) + \sum_{j=1}^{n} S(x_j)\right]$$

pour $\tau_i(x) = \sum_{j=1}^{n} T_i(x_j)$, la statistique dite *naturelle*, $i = 1, \ldots, k$.

- Notons que la statistique naturelle est de dimension k $\forall n$.
- Que peut-on dire de la loi de $\boldsymbol{\tau} = \big(\tau_1(X), \ldots, \tau_k(X)\big)$?

La statistique naturelle

Lemme 1

La loi jointe de $\boldsymbol{\tau} = \big(\tau_1(X), \ldots, \tau_k(X)\big)$ est de la famille exponentielle de paramètres naturels $c_1(\theta), \ldots, c_k(\theta)$.

La statistique naturelle

Lemme 1

La loi jointe de $\boldsymbol{\tau} = \big(\tau_1(X), \ldots, \tau_k(X)\big)$ est de la famille exponentielle de paramètres naturels $c_1(\theta), \ldots, c_k(\theta)$.

Lemme 2

Pour tout $A \subseteq \{1, \ldots, k\}$, la loi jointe de $\big\{\tau_i(X); i \in A\big\}$ conditionnellement à $\big\{\tau_i(X); i \in \bar{A}\big\}$ est de la famille exponentielle et ne dépend que de $\big\{c_i(\theta); i \in A\big\}$.

Statistique naturelle et exhaustivité

- On sait déjà que $\boldsymbol{\tau}$ est exhaustive pour $\phi = c(\theta)$.
- Les résultats précédents nous montre que chaque τ_i est exhaustive pour $\phi_i = c_i(\theta)$.
- En effet $\boldsymbol{\tau}_A$ est exhaustive pour ϕ_A pour tout $A \subseteq \{1, \ldots, k\}$
- Chaque statistique naturelle contient l'information pertinente pour le paramètre naturel.