Estimation paramétrique robuste et optimale

Mathieu Sart

Séminaire Stéphanois de Mathématiques Accessibles

15 octobre 2014

Rappels de probabilités élémentaires

- On considère une expérience dont on ne peut savoir, avec certitude le résultat. Une telle expérience sera appelée expérience aléatoire.
- On peut modéliser l'ensemble des résultats que cette expérience peut avoir par un ensemble Ω.
  - Exemple : si on mesure la durée de vie d'une ampoule, $\Omega = (0, +\infty)$, la durée de vie de deux ampoules $\Omega = (0, +\infty)^2$ ...
- On va considérer certaines parties A de Ω et associer à chacune d'entre elles un nombre $\mathbb{P}(A) \in [0, 1]$ qui va représenter la probabilité que l'événement A se produise.
  - Exemple : si on mesure la durée de vie de deux ampoules, $A = [1, +\infty) \times [2, +\infty)$ ...

Rappels de probabilités élémentaires

- On considère une expérience dont on ne peut savoir, avec certitude le résultat. Une telle expérience sera appelée expérience aléatoire.
- On peut modéliser l'ensemble des résultats que cette expérience peut avoir par un ensemble Ω.
  - Exemple : si on mesure la durée de vie d'une ampoule, $\Omega = (0, +\infty)$, la durée de vie de deux ampoules $\Omega = (0, +\infty)^2$ ...
- On va considérer certaines parties A de Ω et associer à chacune d'entre elles un nombre $\mathbb{P}(A) \in [0, 1]$ qui va représenter la probabilité que l'événement A se produise.
  - Exemple : si on mesure la durée de vie de deux ampoules, $A = [1, +\infty) \times [2, +\infty)$ ...
- De manière plus précise :
  - On munit Ω d'une tribu $\mathcal{F}$ ainsi que d'une mesure de probabilité $\mathbb{P} : \mathcal{F} \to [0, 1]$.
  - Tout ensemble $A \in \mathcal{F}$ sera appelé événement et $\mathbb{P}(A)$ représentera la probabilité que l'événement A se produise.

Rappels de probabilités élémentaires

- On appelle variable aléatoire réelle toute application (mesurable) X de Ω dans $\mathbb{R}$.

Rappels de probabilités élémentaires

- On appelle variable aléatoire réelle toute application (mesurable) X de Ω dans $\mathbb{R}$.
  - Dans l'exemple où l'on mesure la durée de vie de deux ampoules, $\Omega = (0, +\infty)^2$ et $X : \Omega \to \mathbb{R}$ définie par $X(w_1, w_2) = \min\{w_1, w_2\}$

Rappels de probabilités élémentaires

- On appelle variable aléatoire réelle toute application (mesurable) X de Ω dans $\mathbb{R}$.
  - Dans l'exemple où l'on mesure la durée de vie de deux ampoules, $\Omega = (0, +\infty)^2$ et $X : \Omega \to \mathbb{R}$ définie par $X(w_1, w_2) = \min\{w_1, w_2\}$
- On dit que X est une variable continue, s'il existe une fonction (mesurable) positive s telle que $\int_{\mathbb{R}} s(x)\,\mathrm{d}x = 1$ et telle que pour tout intervalle A,

$$\mathbb{P}(X \in A) = \int_A s(x)\,\mathrm{d}x$$

Dans toute la suite, les variables aléatoires seront supposées continues.

Cadre statistique

- Considérons maintenant n variables aléatoires continues indépendantes , $X_1, \ldots, X_n$ admettant une densité s. Cela signifie que pour tout intervalles $A_1, \ldots, A_n$,

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i \in A_i\right) = \prod_{i=1}^{n} \int_{A_i} s(x)\,\mathrm{d}x$$

Cadre statistique

- Considérons maintenant n variables aléatoires continues indépendantes , $X_1, \ldots, X_n$ admettant une densité s. Cela signifie que pour tout intervalles $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) = \prod_{i=1}^{n} \int_{A_i} s(x)\,\mathrm{d}x$$

- **On ne connaît pas** entièrement s **mais on observe** les variables aléatoires X_i
- But : essayer de déterminer ce que vaut s.

Cadre statistique

- Considérons maintenant n variables aléatoires continues indépendantes , $X_1, \ldots, X_n$ admettant une densité s. Cela signifie que pour tout intervalles $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) = \prod_{i=1}^{n} \int_{A_i} s(x)\,\mathrm{d}x$$

- **On ne connaît pas** entièrement s **mais on observe** les variables aléatoires X_i
- But : essayer de déterminer ce que vaut s.
- Pour cela, on va construire des estimateurs $\hat{s}$ de s, c'est à dire des fonctions $\hat{s}$ de la forme $\hat{s} = \psi(X_1, \ldots, X_n)$ qui sont "le plus proche possible" de s.

Exemple en statistique paramétrique

- On laisse allumer n ampoules, et on note X_i la durée de vie de l'ampoule numéro i.
- Chaque X_i est supposée admettre la *même* densité s.

Exemple en statistique paramétrique

- On laisse allumer n ampoules, et on note X_i la durée de vie de l'ampoule numéro i.
- Chaque X_i est supposée admettre la *même* densité s.
- Si l'on suppose que le taux de défaillance d'une ampoule ne dépend pas du temps, on peut montrer que chaque X_i admet la densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

pour un certain $\lambda > 0$.

- Estimer s revient alors à estimer λ.

Exemple en statistique paramétrique

- On observe n variables $X_1, \ldots, X_n$ i.i.d, de densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

et on veut estimer λ.

Exemple en statistique paramétrique

- On observe n variables $X_1, \ldots, X_n$ i.i.d, de densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

  et on veut estimer λ.
- Méthode des moments : on calcule $\mathbb{E}(X)$. Si $\mathbb{E}(X)$ dépend de λ, on peut écrire $\mathbb{E}(X) = f(\lambda)$ donc $\lambda = f^{-1}(\mathbb{E}(X))$ si f est inversible.

Exemple en statistique paramétrique

- On observe n variables $X_1, \ldots, X_n$ i.i.d, de densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

et on veut estimer λ.

- Méthode des moments : on calcule $\mathbb{E}(X)$. Si $\mathbb{E}(X)$ dépend de λ, on peut écrire $\mathbb{E}(X) = f(\lambda)$ donc $\lambda = f^{-1}(\mathbb{E}(X))$ si f est inversible. Par la loi des grands nombres, $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i \simeq \mathbb{E}(X)$, et il semble donc raisonnable d'estimer λ par

$$\hat{\lambda} = f^{-1}(\bar{X}_n).$$

Exemple en statistique paramétrique

- On observe n variables $X_1, \ldots, X_n$ i.i.d, de densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

et on veut estimer λ.

- Méthode des moments : dans cet exemple

$$\mathbb{E}(X) = \int_{\mathbb{R}} x\, s(x)\, \mathrm{d}x = 1/\lambda.$$

Par la loi des grands nombres,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \simeq \mathbb{E}(X).$$

Par conséquent, on peut estimer λ par $\hat{\lambda} = 1/\bar{X}_n$.

Exemple en statistique paramétrique

- On observe n variables $X_1, \ldots, X_n$ i.i.d, de densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

et on veut estimer λ.

- Méthode du maximum de vraisemblance : le vecteur aléatoire $Z = (X_1, \ldots, X_n)$ admet la densité

$$L_\lambda(x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \mathbb{1}_{[0,+\infty)}(x_i)$$

Exemple en statistique paramétrique

- On observe n variables $X_1, \ldots, X_n$ i.i.d, de densité

$$s(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$$

et on veut estimer λ.

- Méthode du maximum de vraisemblance : le vecteur aléatoire $Z = (X_1, \ldots, X_n)$ admet la densité

$$L_\lambda(x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \mathbb{1}_{[0,+\infty)}(x_i)$$

On définit l'estimateur $\hat{\lambda}$ comme étant n'importe quel maximiseur de la fonction

$$\lambda \mapsto L_\lambda(X_1, \ldots, X_n)$$

Cadre statistique

- Lorsque l'on observe n variables $X_1, \ldots, X_n$ i.i.d admettant une densité s que l'on souhaite estimer, il faut tout d'abord modéliser l'information que l'on a sur s.

Cadre statistique

- Lorsque l'on observe n variables $X_1, \ldots, X_n$ i.i.d admettant une densité s que l'on souhaite estimer, il faut tout d'abord modéliser l'information que l'on a sur s.
  - Dans l'exemple précédent, on supposait que s appartenait au modèle $S = \{f_\lambda,\ \lambda > 0\}$ où $f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$

Cadre statistique

- Lorsque l'on observe n variables $X_1, \ldots, X_n$ i.i.d admettant une densité s que l'on souhaite estimer, il faut tout d'abord modéliser l'information que l'on a sur s.
  - Dans l'exemple précédent, on supposait que s appartenait au modèle $S = \{f_\lambda,\ \lambda > 0\}$ où $f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$
- Pour cela, on considère un ensemble de densité S et on suppose que $s \in S$ ou s est "proche" de S. On aimerait avoir une procédure statistique qui fournit de "bons" estimateur $\hat{s}$ de s.

Cadre statistique

- Lorsque l'on observe n variables $X_1, \ldots, X_n$ i.i.d admettant une densité s que l'on souhaite estimer, il faut tout d'abord modéliser l'information que l'on a sur s.
  - Dans l'exemple précédent, on supposait que s appartenait au modèle $S = \{f_\lambda,\ \lambda > 0\}$ où $f_\lambda(x) = \lambda e^{-\lambda x} \mathbb{1}_{[0,+\infty)}(x)$
- Pour cela, on considère un ensemble de densité S et on suppose que $s \in S$ ou s est "proche" de S. On aimerait avoir une procédure statistique qui fournit de "bons" estimateur $\hat{s}$ de s.
- Par souci de simplicité, dans toute la suite de l'exposé l'ensemble S sera un modèle paramétrique : $S = \{f_\theta,\ \theta \in \Theta\}$ où $\Theta \subset \mathbb{R}$.

La méthode des moments et du max de vraisemblance

- Sous des hypothèses :
  - de régularité sur le modèle S
  - et qu'il existe θ_0 tel que $s = f_{\theta_0}$

  l'estimateur du maximum de vraisemblance possède des propriétés d'optimalité (efficacité).

La méthode des moments et du max de vraisemblance

- Sous des hypothèses :
  - de régularité sur le modèle S
  - et qu'il existe θ_0 tel que $s = f_{\theta_0}$

  l'estimateur du maximum de vraisemblance possède des propriétés d'optimalité (efficacité).
- Les estimateurs donnés par la méthode des moments peuvent ne pas être optimaux : si par exemple $f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}$, alors l'estimateur de θ par méthode des moments est $\hat{\theta} = 2\bar{X}_n$.

La méthode des moments et du max de vraisemblance

- Sous des hypothèses :
  - de régularité sur le modèle S
  - et qu'il existe θ_0 tel que $s = f_{\theta_0}$

  l'estimateur du maximum de vraisemblance possède des propriétés d'optimalité (efficacité).
- Les estimateurs donnés par la méthode des moments peuvent ne pas être optimaux : si par exemple $f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}$, alors l'estimateur de θ par méthode des moments est $\hat{\theta} = 2\bar{X}_n$. Pourtant :

$$|\hat{\theta} - \theta| \simeq 1/\sqrt{n}$$

alors qu'il existe des estimateurs $\tilde{\theta}$ tels que

$$|\tilde{\theta} - \theta| \simeq 1/n$$

La méthode des moments et du max de vraisemblance

Ces deux méthodes ne sont pas universelles : il existe des modèles paramétriques "simples" pour lesquels ces méthodes ne fonctionnent pas

- Par exemple, si
$$f_\theta(x) = \frac{1}{\pi\left((x-\theta)^2+1\right)},$$
les X_i n'admettent aucun moment, et par conséquent la méthode des moments ne fournit aucun estimateur.
- Par exemple, si
$$f_\theta(x) = \begin{cases} \frac{1}{4\sqrt{|x-\theta|}}\mathbb{1}_{[-1,1]}(x-\theta) & \text{pour tout } x \in \mathbb{R}\setminus\{\theta\}, \\ 0 & \text{pour } x = \theta. \end{cases}$$
l'estimateur du maximum de vraisemblance n'existe pas.

La méthode des moments et du max de vraisemblance

Ces deux méthodes ne fournissent pas des estimateurs robustes : s'il s'avère que $s \notin S$, les estimateurs peuvent être très mauvais même si s "est très proche" de S.

La méthode des moments et du max de vraisemblance

Ces deux méthodes ne fournissent pas des estimateurs robustes : s'il s'avère que $s \notin S$, les estimateurs peuvent être très mauvais même si s "est très proche" de S. Exemple :

- Si $f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}$, l'estimateur du maximum de vraisemblance de s est $\hat{s} = f_{\hat{\theta}_{\text{emv}}}$ avec $\hat{\theta}_{\text{emv}} = \max_{1\leq i\leq n} X_i$.
- Si $s = (1-p)\mathbb{1}_{[0,1]} + \frac{p}{2}\mathbb{1}_{[0,2]}$ avec p très petit, $s \simeq \mathbb{1}_{[0,1]}$ et un bon estimateur $\hat{s}$ devrait être proche de s, au moins lorsque p est assez petit et n assez grand.

La méthode des moments et du max de vraisemblance

Ces deux méthodes ne fournissent pas des estimateurs robustes : s'il s'avère que $s \notin S$, les estimateurs peuvent être très mauvais même si s "est très proche" de S. Exemple :

- Si $f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}$, l'estimateur du maximum de vraisemblance de s est $\hat{s} = f_{\hat{\theta}_{\text{emv}}}$ avec $\hat{\theta}_{\text{emv}} = \max_{1\leq i\leq n} X_i$.
- Si $s = (1-p)\mathbb{1}_{[0,1]} + \frac{p}{2}\mathbb{1}_{[0,2]}$ avec p très petit, $s \simeq \mathbb{1}_{[0,1]}$ et un bon estimateur $\hat{s}$ devrait être proche de s, au moins lorsque p est assez petit et n assez grand.
- Pourtant, quelque soit $p > 0$, l'estimateur du maximum de vraisemblance $\hat{s} = f_{\hat{\theta}_{\text{emv}}}$ converge vers $\frac{1}{2}\mathbb{1}_{[0,2]}$.

Objectifs

On aimerait avoir une procédure d'estimation plus ou moins *universelle* qui réalise les objectifs suivants :

Objectifs

On aimerait avoir une procédure d'estimation plus ou moins *universelle* qui réalise les objectifs suivants :

- Lorsque le modèle est vrai, i.e, $s = f_{\theta_0}$ l' estimateur est de la forme $\hat{s} = f_{\hat{\theta}}$ et $\hat{\theta}$ converge vers le vrai paramètre θ_0 à la vitesse optimale.

Objectifs

On aimerait avoir une procédure d'estimation plus ou moins *universelle* qui réalise les objectifs suivants :

- Lorsque le modèle est vrai, i.e, $s = f_{\theta_0}$ l' estimateur est de la forme $\hat{s} = f_{\hat{\theta}}$ et $\hat{\theta}$ converge vers le vrai paramètre θ_0 à la vitesse optimale.
- L'estimateur est robuste par rapport à des erreurs de modèles, c'est à dire qu'il se comporte encore correctement lorsque $s \notin S$ mais lui est proche.

Objectifs

On aimerait avoir une procédure d'estimation plus ou moins *universelle* qui réalise les objectifs suivants :

- Lorsque le modèle est vrai, i.e, $s = f_{\theta_0}$ l' estimateur est de la forme $\hat{s} = f_{\hat{\theta}}$ et $\hat{\theta}$ converge vers le vrai paramètre θ_0 à la vitesse optimale.
- L'estimateur est robuste par rapport à des erreurs de modèles, c'est à dire qu'il se comporte encore correctement lorsque $s \notin S$ mais lui est proche.

Remarque : ces objectifs sont difficiles à atteindre lorsque le modèle n'est pas régulier comme dans l'exemple où

$$f_\theta(x) = \begin{cases} \frac{1}{4\sqrt{|x-\theta|}} \mathbb{1}_{[-1,1]}(x-\theta) & \text{pour tout } x \in \mathbb{R} \setminus \{\theta\}, \\ 0 & \text{pour } x = \theta. \end{cases}$$

Objectifs

On aimerait avoir une procédure d'estimation plus ou moins *universelle* qui réalise les objectifs suivants :

- Lorsque le modèle est vrai, i.e, $s = f_{\theta_0}$ l' estimateur est de la forme $\hat{s} = f_{\hat{\theta}}$ et $\hat{\theta}$ converge vers le vrai paramètre θ_0 à la vitesse optimale.
- L'estimateur est robuste par rapport à des erreurs de modèles, c'est à dire qu'il se comporte encore correctement lorsque $s \notin S$ mais lui est proche.

Dans toute la suite, on utilisera la distance de Hellinger h définie pour toutes densités f et f' par

$$h^2(f, f') = \frac{1}{2} \int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{f'(x)} \right)^2 \, \mathrm{d}x$$

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Alors,

$$h^2(s, f_{\theta'}) + T_E(f_\theta, f_{\theta'}) = h^2(s, f_\theta) + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} \left(\sqrt{s} - \sqrt{\frac{f_\theta + f_{\theta'}}{2}}\right)^2 \mathrm{d}x$$

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Alors,

$$h^2(s, f_{\theta'}) + T_E(f_\theta, f_{\theta'}) = h^2(s, f_\theta) + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} \left(\sqrt{s} - \sqrt{\frac{f_\theta + f_{\theta'}}{2}}\right)^2 \mathrm{d}x$$

On peut montrer que

$$\frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} \left(\sqrt{s} - \sqrt{\frac{f_\theta + f_{\theta'}}{2}}\right)^2 \mathrm{d}x \leq \frac{1}{\sqrt{2}} (h^2(s, f_\theta) + h^2(s, f_{\theta'}))$$

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Alors,

$$(1 - 1/\sqrt{2})h^2(s, f_{\theta'}) + T_E(f_\theta, f_{\theta'}) \leq (1 + 1/\sqrt{2})h^2(s, f_\theta)$$

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Alors,

$$(1 - 1/\sqrt{2}) h^2(s, f_{\theta'}) + T_E(f_\theta, f_{\theta'}) \leq (1 + 1/\sqrt{2}) h^2(s, f_\theta)$$

En particulier :

- Si $T_E(f_\theta, f_{\theta'}) \geq 0$, alors $h^2(s, f_{\theta'}) \leq \frac{\sqrt{2}+1}{\sqrt{2}-1} h^2(s, f_\theta)$
- Si $T_E(f_\theta, f_{\theta'}) \leq 0$, alors $h^2(s, f_\theta) \leq \frac{\sqrt{2}+1}{\sqrt{2}-1} h^2(s, f_{\theta'})$

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Regarder le signe de $T_E(f, f')$ permet de sélectionner une fonction $\hat{s} \in \{f_\theta, f_{\theta'}\}$ telle que

$$h^2(s, \hat{s}) \leq C \inf_{\theta \in \Theta} h^2(s, f_\theta)$$

où $C = (\sqrt{2}+1)/(\sqrt{2}-1)$.

Comparaison de deux densités

Le modèle le plus simple est celui où $S = \{f_\theta, f_{\theta'}\}$ auquel cas il faut comparer $h^2(s, f_\theta)$ à $h^2(s, f_{\theta'})$.
Posons

$$T_E(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta}\right) \mathrm{d}x + \frac{1}{\sqrt{2}} \int_{\mathbb{R}} \frac{\sqrt{f_{\theta'}} - \sqrt{f_\theta}}{\sqrt{f_\theta + f_{\theta'}}} s \, \mathrm{d}x.$$

Regarder le signe de $T_E(f, f')$ permet de sélectionner une fonction $\hat{s} \in \{f_\theta, f_{\theta'}\}$ telle que

$$h^2(s, \hat{s}) \leq C \inf_{\theta \in \Theta} h^2(s, f_\theta)$$

où $C = (\sqrt{2}+1)/(\sqrt{2}-1)$.

Introduction du test

On définit donc

$$T(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta} \right) \mathrm{d}x + \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} \frac{\sqrt{f_{\theta'}(X_i)} - \sqrt{f_\theta(X_i)}}{\sqrt{f_\theta(X_i) + f_{\theta'}(X_i)}}$$

Introduction du test

On définit donc

$$T(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta} \right) \mathrm{d}x + \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} \frac{\sqrt{f_{\theta'}(X_i)} - \sqrt{f_\theta(X_i)}}{\sqrt{f_\theta(X_i) + f_{\theta'}(X_i)}}$$

Alors, on peut montrer que $T(f_\theta, f_{\theta'})$ et $T_E(f_\theta, f_{\theta'}) = \mathbb{E}\left[T(f_\theta, f_{\theta'})\right]$ sont proches avec grande probabilité.

Introduction du test

On définit donc

$$T(f_\theta, f_{\theta'}) = \frac{1}{2\sqrt{2}} \int_{\mathbb{R}} \sqrt{f_\theta + f_{\theta'}} \left(\sqrt{f_{\theta'}} - \sqrt{f_\theta} \right) \mathrm{d}x + \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} \frac{\sqrt{f_{\theta'}(X_i)} - \sqrt{f_\theta(X_i)}}{\sqrt{f_\theta(X_i) + f_{\theta'}(X_i)}}$$

Alors, on peut montrer que $T(f_\theta, f_{\theta'})$ et $T_E(f_\theta, f_{\theta'}) = \mathbb{E}\left[T(f_\theta, f_{\theta'})\right]$ sont proches avec grande probabilité.

Regarder le signe de $T(f_\theta, f_{\theta'})$ permet alors de sélectionner un estimateur $\hat{s} = f_{\hat{\theta}} \in \{f_\theta, f_{\theta'}\}$ tel que pour tout $\zeta > 0$,

$$\mathbb{P}\left[Ch^2(s, \hat{s}) \leq \inf_{\theta \in \{\theta, \theta'\}} h^2(s, f_\theta) + \zeta \right] \geq 1 - e^{-n\zeta}$$

où $C > 0$ est une constante numérique.

Modèle paramétrique

- On considère désormais des modèles paramétriques $S = \{f_\theta, \theta \in \Theta\}$ indexés par un intervalle fini $\Theta = [m, M]$ de $\mathbb{R}$.

Modèle paramétrique

- On considère désormais des modèles paramétriques $S = \{f_\theta, \theta \in \Theta\}$ indexés par un intervalle fini $\Theta = [m, M]$ de $\mathbb{R}$.
- Pour construire un bon estimateur $\hat{s} \in S$, on aimerait comparer un grand nombre de densités f_θ et $f_{\theta'}$.

Modèle paramétrique

- On considère désormais des modèles paramétriques $S = \{f_\theta, \theta \in \Theta\}$ indexés par un intervalle fini $\Theta = [m, M]$ de $\mathbb{R}$.
- Pour construire un bon estimateur $\hat{s} \in S$, on aimerait comparer un grand nombre de densités f_θ et $f_{\theta'}$.
- On commet à chaque fois une erreur lorsque l'on estime $T_E(f_\theta, f_{\theta'})$ par $T(f_\theta, f_{\theta'})$.

Modèle paramétrique

- On considère désormais des modèles paramétriques $S = \{f_\theta,\ \theta \in \Theta\}$ indexés par un intervalle fini $\Theta = [m, M]$ de $\mathbb{R}$.
- Pour construire un bon estimateur $\hat{s} \in S$, on aimerait comparer un grand nombre de densités f_θ et $f_{\theta'}$.
- On commet à chaque fois une erreur lorsque l'on estime $T_E(f_\theta, f_{\theta'})$ par $T(f_\theta, f_{\theta'})$.
- Pour limiter ces erreurs, on discrétise Θ en un ensemble fini Θ_{dis}. On considère $\varepsilon > 0$ et on définit

$$\Theta_{\text{dis}} = \left\{ m + k\varepsilon,\ k \in \mathbb{N},\ k \leq (M-m)\varepsilon^{-1} \right\}$$

- Soit π une projection de Θ sur Θ_{dis} et

$$T(\theta, \theta') = T(f_{\pi(\theta)}, f_{\pi(\theta')})$$

Modèle paramétrique

- On a donc accès pour tout $\theta, \theta' \in \Theta = [m, M]$, à une fonction mesurable des observations $T(\theta, \theta')$ telle que :
  - Si $T(\theta, \theta') \geq 0$ alors, on a à peu près $h^2(s, f_\theta) > \kappa h^2(f_\theta, f_{\theta'})$ pour un $\kappa > 0$
  - Si $T(\theta, \theta') \leq 0$ alors, on a à peu près $h^2(s, f_{\theta'}) > \kappa h^2(f_\theta, f_{\theta'})$ pour un $\kappa > 0$

Modèle paramétrique

- On a donc accès pour tout $\theta, \theta' \in \Theta = [m, M]$, à une fonction mesurable des observations $T(\theta, \theta')$ telle que :
  - Si $T(\theta, \theta') \geq 0$ alors, on a à peu près $h^2(s, f_\theta) > \kappa h^2(f_\theta, f_{\theta'})$ pour un $\kappa > 0$
  - Si $T(\theta, \theta') \leq 0$ alors, on a à peu près $h^2(s, f_{\theta'}) > \kappa h^2(f_\theta, f_{\theta'})$ pour un $\kappa > 0$
- Soit pour tout $\theta \in \Theta$ et $r > 0$, $\mathcal{B}(\theta, r) = \{\theta' \in \Theta,\ h(f_\theta, f_{\theta'}) \leq r\}$

Modèle paramétrique

- On a donc accès pour tout $\theta, \theta' \in \Theta = [m, M]$, à une fonction mesurable des observations $T(\theta, \theta')$ telle que :
  - Si $T(\theta, \theta') \geq 0$ alors, on a à peu près $h^2(s, f_\theta) > \kappa h^2(f_\theta, f_{\theta'})$ pour un $\kappa > 0$
  - Si $T(\theta, \theta') \leq 0$ alors, on a à peu près $h^2(s, f_{\theta'}) > \kappa h^2(f_\theta, f_{\theta'})$ pour un $\kappa > 0$
- Soit pour tout $\theta \in \Theta$ et $r > 0$, $\mathcal{B}(\theta, r) = \{\theta' \in \Theta,\ h(f_\theta, f_{\theta'}) \leq r\}$
- Supposons qu'il existe θ_0 tel que $s = f_{\theta_0}$. Alors, on déduit des points ci dessus que, vraisemblablement :
  - Si $T(\theta, \theta') \geq 0$, $\theta_0 \in \Theta \setminus \mathcal{B}(\theta, \kappa^{1/2} h(f_\theta, f_{\theta'}))$
  - Si $T(\theta, \theta') \leq 0$, $\theta_0 \in \Theta \setminus \mathcal{B}(\theta', \kappa^{1/2} h(f_\theta, f_{\theta'}))$
- Cette idée permet de construit de manière itérative une suite décroissante d'intervalles.

Hypothèse sur le modèle

Hypothèse

Il existe des nombres strictement positifs α, $\underline{R}$, $\overline{R}$ *tels que pour tout* $\theta, \theta' \in [m, M]$,

$$\underline{R}|\theta - \theta'|^{\alpha} \leq h^2\left(f_\theta, f_{\theta'}\right) \leq \overline{R}|\theta - \theta'|^{\alpha}.$$

Hypothèse sur le modèle

Hypothèse

Il existe des nombres strictement positifs α, $\underline{R}$, $\overline{R}$ *tels que pour tout* $\theta, \theta' \in [m, M]$,

$$\underline{R}|\theta - \theta'|^\alpha \leq h^2\left(f_\theta, f_{\theta'}\right) \leq \overline{R}|\theta - \theta'|^\alpha.$$

Exemples :

- Modèles réguliers : l'hypothèse est vraie avec $\alpha = 2$.
- Modèle loi uniforme : $f_\theta = \theta^{-1}\mathbb{1}_{[0,\theta]}$, vraie avec $\alpha = 1$.
- Modèle

$$f_\theta(x) = \begin{cases} \frac{1}{4\sqrt{|x-\theta|}}\mathbb{1}_{[-1,1]}(x-\theta) & \text{pour tout } x \in \mathbb{R} \setminus \{\theta\}, \\ 0 & \text{pour } x = \theta. \end{cases}$$

L'hypothèse est vraie avec $\alpha = 1/2$.

Une borne de risque non-asymptotique

Théorème

Supposons l'hypothèse précédente vérifiée. Alors pour tout $\xi > 0$, l'estimateur $\hat{s} = f_{\hat{\theta}}$ vérifie pour tout $\xi > 0$,

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\theta}}) \leq h^2(s, S) + \frac{D_S}{n} + \xi\right] \geq 1 - e^{-n\xi}$$

où D_S dépend des paramètres du modèle et du pas du réseau et où $C > 0$ dépend seulement de $\overline{R}/\underline{R}$. Sous une hypothèse faible sur S, C est numérique.

Une borne de risque non-asymptotique

Théorème

Supposons l'hypothèse précédente vérifiée. Alors pour tout $\xi > 0$, l'estimateur $\hat{s} = f_{\hat{\theta}}$ vérifie pour tout $\xi > 0$,

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\theta}}) \leq h^2(s, S) + \frac{D_S}{n} + \xi\right] \geq 1 - e^{-n\xi}$$

où D_S dépend des paramètres du modèle et du pas du réseau et où $C > 0$ dépend seulement de $\overline{R}/\underline{R}$. Sous une hypothèse faible sur S, C est numérique.

En particulier, si $s = f_{\theta_0}$ appartient à S, l'estimateur $\hat{\theta}$ converge p.s. vers θ_0 et il existe a, b tels que

$$\mathbb{P}\left[n^{1/\alpha}\left|\hat{\theta} - \theta_0\right| \geq \xi\right] \leq ae^{-b\xi^\alpha} \quad \text{pour tout } \xi > 0.$$

Une borne de risque non-asymptotique

Théorème

Supposons l'hypothèse précédente vérifiée. Alors pour tout $\xi > 0$, l'estimateur $\hat{s} = f_{\hat{\theta}}$ vérifie pour tout $\xi > 0$,

$$\mathbb{P}\left[Ch^2(s, f_{\hat{\theta}}) \leq h^2(s, S) + \frac{D_S}{n} + \xi\right] \geq 1 - e^{-n\xi}$$

où D_S dépend des paramètres du modèle et du pas du réseau et où $C > 0$ dépend seulement de $\overline{R}/\underline{R}$. Sous une hypothèse faible sur S, C est numérique.

En particulier, si $s = f_{\theta_0}$ appartient à S, l'estimateur $\hat{\theta}$ converge p.s. vers θ_0 et il existe a, b tels que

$$\mathbb{P}\left[n^{1/\alpha}\left|\hat{\theta} - \theta_0\right| \geq \xi\right] \leq ae^{-b\xi^\alpha} \quad \text{pour tout } \xi > 0.$$

Si, de plus, le modèle est suffisamment régulier, $\alpha = 2$ et donc

$$\mathbb{P}\left[\sqrt{n}\left|\hat{\theta} - \theta_0\right| \geq \xi\right] \leq ae^{-b\xi^2} \quad \text{pour tout } \xi > 0.$$

Connexion asymptotique avec le m.l.e

Théorème

Sous des hypothèses :

- *(trop) fortes de régularité sur le modèle paramétrique S*
- *que le modèle est vraie, i.e, $s = f_{\theta_0} \in S$*
- *que le pas du réseau ε_n, dépend de n et tend vers 0 de sorte que $\varepsilon_n = o(1/\sqrt{n})$ mais $|\log \varepsilon_n| = o(n)$.*
- *qu'il existe un estimateur du maximum de vraisemblance $\tilde{\theta}_{emv}$ qui converge en probabilité vers θ_0,*

Alors, l'estimateur $\hat{\theta}$ vérifie

$$\hat{\theta} = \tilde{\theta}_{emv} + \mathcal{O}_{\mathbb{P}}(\varepsilon_n)$$

En particulier il est efficace.

Remarque : lorsque le modèle est faux, i.e, $s \notin S$, l'estimateur possède néanmoins des propriétés de robustesses par rapport à la distance de Hellinger.

Simulations

Les simulations montrent que l'estimateur coincide presque avec l'estimateur du maximum de vraisemblance lorsque le réseau est assez petit, lorsque S est assez régulier et contient s :

		$n = 10$	$n = 25$	$n = 50$	$n = 75$	$n = 100$
Loi exponentielle	$\hat{q}_{0.99}$	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$
	$\hat{q}_{0.999}$	0.07	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$
	$\hat{q}_1$	1.9	0.3	0.06	0.005	$< \epsilon$
Loi Normale	$\hat{q}_1$	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$
Loi Cauchy	$\hat{q}_{0.999}$	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$	$< \epsilon$
	$\hat{q}_1$	1.5	0.1	$< \epsilon$	$< \epsilon$	$< \epsilon$

$\hat{q}_\alpha$ est le quantile d'ordre α de la variable aléatoire $|\hat{\theta} - \hat{\theta}_{\text{emv}}|$.
Etude numérique basée sur 10^6 simulations pour les exemples 1 et 2 et sur 10^4 pour l'exemple 3.
$\epsilon \simeq 10^{-6}$.

Simulations

On peut vérifier la robustesse de l'estimateur :

- Modèle : $S = \left\{\theta^{-1}\mathbb{1}_{[0,\theta]},\ \theta \in [0.01, 10]\right\}$
- Les données sont simulées selon la densité $s_p = (1-p)\mathbb{1}_{[0,1]} + p\frac{\mathbb{1}_{[0,2]}}{2}$

Simulations

On peut vérifier la robustesse de l'estimateur :

- Modèle : $S = \{\theta^{-1}\mathbb{1}_{[0,\theta]},\ \theta \in [0.01, 10]\}$
- Les données sont simulées selon la densité $s_p = (1-p)\mathbb{1}_{[0,1]} + p\frac{\mathbb{1}_{[0,2]}}{2}$
- Notation : $\widehat{R}_p(\tilde{\theta})$ estime $\mathbb{E}[h^2(s_p, f_{\tilde{\theta}})]$ en calculant pour chaque p, 5000 échantillons de taille $n = 100$.

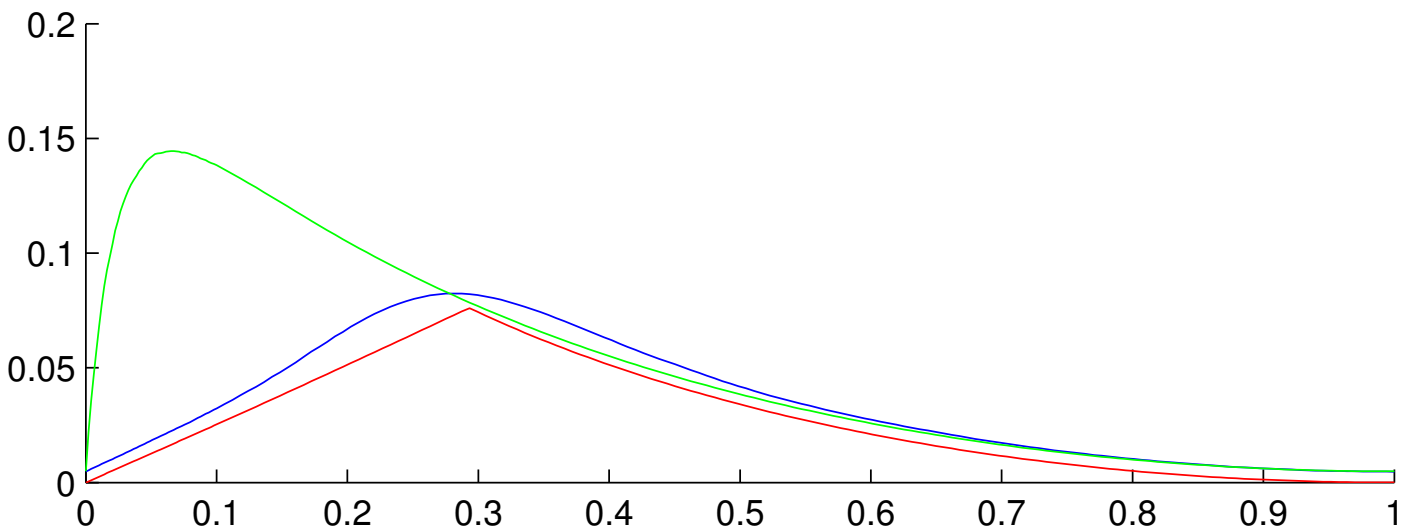

FIGURE: Rouge : $p \mapsto h^2(s_p, S)$. Bleu : $p \mapsto \widehat{R}_p(\hat{\theta})$. Vert : $p \mapsto \widehat{R}_p(\tilde{\theta}_{\text{emv}})$.

Références

- Pour la procédure :
  - Sart, Robust estimation on a parametric model via testing
- Pour la construction du test :
  - Baraud, Estimator selection with respect to Hellinger-type risks
- Pour plus d'informations sur l'estimation par tests :
  - Birgé, Model selection via testing : an alternative to (penalized) maximum likelihood estimators
  - Baraud, Birgé, Sart, A new method for estimation and model selection : ρ-estimation