Utilisation de données simulées pour l'estimation de certaines caractéristiques d'une loi. Applications à l'aéronautique.

Thierry Klein

Institut de Mathématiques ,Toulouse

Travail en commun avec

Jean-Claude Fort (Université Paris 5) et Nabil Rachdi (E.A.D.S.).

Nabil Rachdi; Jean-Claude Fort; Thierry Klein
Stochastic Inverse Problem with Noisy Simulator. Application to aeronautical model
Annales de la faculté des sciences de Toulouse Sér. 6, 21 no. 3 (2012), p. 593-622,

Nabil Rachdi; Jean-Claude Fort; Thierry Klein
Risk bounds for new M-estimation problems.
Accepté à ESAIM PS
disponible sur HAL

Cadre informel

Partie II

L'objectif est l'étude (estimation) de certaines caractéristiques ($\mathbb{E}$, Var, densité, quantile ...) d'une variable aléatoire Y.
Pour cela on dispose de deux sources distinctes d'information

1. **Quelques données expérimentales $Y_1, \ldots, Y_n$ i.i.d de même loi que Y (n petit).**

   • Les Y_i représentent soit des vraies données provenant de mesures expérimentales, soit des données provenant d'un modèle simulant de façon précise la réalité.

   • Dans tous les cas ces données sont coûteuses, elles sont peu nombreuses.

Cadre informel

Partie II

L'objectif est l'étude (estimation) de certaines caractéristiques ($\mathbb{E}$, Var, densité, quantile ...) d'une variable aléatoire Y.
Pour cela on dispose de deux sources distinctes d'information

1. **Quelques données expérimentales $Y_1, \ldots, Y_n$ i.i.d de même loi que Y (n petit).**

   • Les Y_i représentent soit des vraies données provenant de mesures expérimentales, soit des données provenant d'un modèle simulant de façon précise la réalité.

   • Dans tous les cas ces données sont coûteuses, elles sont peu nombreuses.

2. **Possibilité de simuler des variables $Z_1, \ldots Z_m$ i.i.d dont la loi est "proche" de celle de Y (m grand).**

   • Les Z_i proviennent de modèles simplifiés, elles sont pas trop coûteuses, on peut en avoir plein.

Un exemple

Etude de la masse de Fuel consommée par un avion lors de la phase croisière pour un trajet fixé. Le but est à partir de données de masses de fuel consommées, **d'identifier la consommation spécifique (SFC) de la motorisation** en tenant compte de l'incertitude sur la vitesse V et sur la finesse F de l'avion.

Données observées

$$M_{fuel}^{*,1}, \ldots, M_{fuel}^{*,n}, \quad n = 32$$

Masses de Fuel [kg]							
7918	7671	7719	7839	7912	7963	7693	7815
7872	7679	8013	7935	7794	8045	7671	7985
7755	7658	7684	7658	7690	7700	7876	7769
8058	7710	7746	7698	7666	7749	7764	7667

Table: Données provenant d'un modèle aéronautique complexe

Un exemple

Etude de la masse de Fuel consommée par un avion lors de la phase croisière pour un trajet fixé. Le but est à partir de données de masses de fuel consommées, **d'identifier la consommation spécifique (SFC) de la motorisation** en tenant compte de l'incertitude sur la vitesse V et sur la finesse F de l'avion.
Données observées

$$M_{fuel}^{*,1}, \ldots, M_{fuel}^{*,n}, \quad n = 32$$

Masses de Fuel [kg]							
7918	7671	7719	7839	7912	7963	7693	7815
7872	7679	8013	7935	7794	8045	7671	7985
7755	7658	7684	7658	7690	7700	7876	7769
8058	7710	7746	7698	7666	7749	7764	7667

Table: Données provenant d'un modèle aéronautique complexe

On dispose en fait de $n_{\max} = 200$ données

$$\left(M_{fuel}^{*,i}, SFC^{*,i}\right)_{i=1\ldots n_{\max}}$$

mais on en utilise que 32 valeurs de $M_{fuel}^{*,i}$ et aucunes valeurs de $SFC^{*,i}$, on utilisera tout le monde plus tard....
On suppose que les observations $M_{fuel}^{*,i}$ ont une distribution inconnue Q de densité f de support $\mathcal{I} := [M_{\inf} = 7600, M_{\sup} = 8100]$

Modèle aéronautique simplifié

$$M_{fuel} = \left(M_{empty} + M_{pload}\right)\left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F}\,10^{-3}} - 1\right)$$

Modèle aéronautique simplifié

$$M_{fuel} = \left(M_{empty} + M_{pload}\right)\left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F}\, 10^{-3}} - 1\right)$$

Dans cette formule, on trouve 7 variables d'entrées

1. 4 quantités fixes M_{empty}, M_{pload}, g et Ra.
2. Deux quantités non contrôlées, modélisées par des variables aléatoires V et F.

Entrée	valeur ou valeur nominale	unité
M_{empty}	42600	kg
M_{pload}	19900	kg
g	9.8	m/s^2
Ra	3000	km
V_{nom}	231	m/s
F_{nom}	19	–

Les v.a. V et F sont modélisées par $V = V_{nom} + \epsilon_V$, $F = F_{nom} + \epsilon_F$

variable	Distribution	Paramètre
V	Uniform	$[V_{min} = 226, V_{max} = 234]$
F	Beta	$(7, 2, [F_{min} = 18.7, F_{max} = 19.05])$

3. La quantité qui nous intéresse SFC.

A propos de la quantité d'intérêt

$$M_{fuel} = (M_{empty} + M_{pload}) \left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F} 10^{-3}} - 1 \right)$$ comparaison

- SFC est une mesure de la qualité du moteur qui possède une composante aléatoire, elle sera modélisée par une v.a à support compact

$$SFC = \mu_{SFC} + \epsilon$$

Dans un premier temps de grands experts en aéronautique nous proposent de prendre $\epsilon = \sigma_{SFC} \epsilon_{SFC}$ avec $\epsilon_{SFC} \sim \mathcal{N}_{[-3,3]}(0,1)$ avec $\mu_{SFC} \in [15, 20]$ et $\sigma_{SFC} \in [s, 1]$ (s petit).

- Notre problème est maintenant de caractériser μ_{SFC} et σ_{SFC}.

A propos de la quantité d'intérêt

$$M_{fuel} = (M_{empty} + M_{pload}) \left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F} 10^{-3}} - 1 \right)$$ comparaison

- SFC est une mesure de la qualité du moteur qui possède une composante aléatoire, elle sera modélisée par une v.a à support compact

$$SFC = \mu_{SFC} + \epsilon$$

Dans un premier temps de grands experts en aéronautique nous proposent de prendre $\epsilon = \sigma_{SFC} \epsilon_{SFC}$ avec $\epsilon_{SFC} \sim \mathcal{N}_{[-3,3]}(0,1)$ avec $\mu_{SFC} \in [15, 20]$ et $\sigma_{SFC} \in [s, 1]$ (s petit).

- Notre problème est maintenant de caractériser μ_{SFC} et σ_{SFC}.
- Ecriture du modèle statistique
  On a un vecteur aléatoire $\epsilon = (\epsilon_{SFC}, \epsilon_V, \epsilon_F)^T$, un vecteur de paramètre $\theta = (\mu_{SFC}, \sigma_{SFC})^T$.
  Une famille paramétrée par $\theta \in \Theta$

$$M_{fuel} = h(\epsilon, \theta) = (M_{empty} + M_{pload}) \left(exp \left(\frac{(\mu_{SFC} + \sigma_{SFC}\, \epsilon_{SFC}) \cdot g \cdot Ra}{(V_{nom} + \epsilon_V) \cdot (F_{nom} + \epsilon_F)} \cdot 10^{-3} \right) - 1 \right)$$

avec

$$\Theta = [15, 20] \times [s, 1]$$

Dans la suite on notera Q_θ la loi de $M_{fuel} = h(\epsilon, \theta)$
L'objectif est maintenant d'estimer $\theta \in \Theta$ à partir $M_{fuel}^{*,1}, ..., M_{fuel}^{*,n}$.

Ça n'a pas l'air bien rigoureux tout ça!!!!!

- Nos vraies données $M_{fuel}^{*,1}, \ldots, M_{fuel}^{*,n}$, $n = 32$ forment un n-échantillon de distribution inconnue Q.

  Ici, la densité f de Q est inconnue, on ne connaît donc pas de relation reliant θ aux $M_{fuel}^{*,i}$.

On remplace la loi Q par la loi Q_θ loi de $M_{fuel} = h(\epsilon, \theta)$ (on considère que notre modèle approché n'est pas trop mauvais). On note ρ_θ la densité correspondante. On décide maintenant d'estimer θ par Maximum de vraisemblance

$$\hat{\theta} = Argmin_{\theta \in \Theta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\rho_\theta(M_{fuel}^{*,i})) \right\}$$

▸ Mais que fait-il?

Ça n'a pas l'air bien rigoureux tout ça!!!!!

- Nos vraies données $M_{fuel}^{*,1}, \ldots, M_{fuel}^{*,n}$, $n = 32$ forment un n-échantillon de distribution inconnue Q.

  Ici, la densité f de Q est inconnue, on ne connaît donc pas de relation reliant θ aux $M_{fuel}^{*,i}$.

  On remplace la loi Q par la loi Q_θ loi de $M_{fuel} = h(\epsilon, \theta)$ (on considère que notre modèle approché n'est pas trop mauvais). On note ρ_θ la densité correspondante. On décide maintenant d'estimer θ par Maximum de vraisemblance

$$\hat{\theta} = Argmin_{\theta \in \Theta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\rho_\theta(M_{fuel}^{*,i})) \right\}$$

  ▸ Mais que fait-il?

- Malheureusement pour nous, vu la fonction $h(\epsilon, \theta) = (M_{empty} + M_{pload}) \left(exp \left(\frac{(\mu_{SFC} + \sigma_{SFC}\, \epsilon_{SFC}) \cdot g \cdot Ra}{(V_{nom} + \epsilon_V) \cdot (F_{nom} + \epsilon_F)} \cdot 10^{-3} \right) - 1 \right)$ on n'a pas d'expression analytique pour ρ_θ.

Ça n'a pas l'air bien rigoureux tout ça!!!!!

$$\hat{\theta} = Argmin_{\theta \in \Theta} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log(\rho_\theta(M_{fuel}^{*,i})) \right\}$$

- C'est pas grave, car on peut simuler (presque gratuitement) un m-échantillon (m grand 10^4) de loi Q_θ et on estime la densité avec notre méthode préférée (par exemple à l'aide d'un noyau)

$$\widehat{\theta} = Argmin_{\theta \in \Theta} \left\{ -\sum_{i=1}^{n} \log \left(\frac{1}{m} \sum_{j=1}^{m} K_{b_\theta^m} \left(h(\epsilon_j, \theta) - M_{fuel}^{*,i} \right) \right) \right\} .$$

- On optimise en θ à l'aide d'une méthode Quasi-Newton, et on obtient les résultats suivants

$$\widehat{\theta} = (\widehat{\boldsymbol{\mu}}_{SFC} = 17.397, \widehat{\boldsymbol{\sigma}}_{SFC} = 0.201).$$

Recyclons les données oubliées

A un moment vous avez pu lire:

On dispose en fait de $n_{\max} = 200$ *données* $\left(M_{fuel}^{*,i}, SFC^{*,i}\right)_{i=1\dots n_{\max}}$ *mais on en utilise que* 32 *valeurs de* $M_{fuel}^{*,i}$ *et aucunes valeurs de* $SFC^{*,i}$*, on utilisera tout le monde plus tard....*

A fin de se persuader que notre procédure est raisonnable, on estime maintenant θ uniquement à l'aide des n_{max} données réelles, et on va comparer cette estimation à l'estimation précédente. On trouve alors

	Echantillon de référence	SFC estimé	Erreur relative
Moyenne	**17.49**	**17.397**	**0.5 %**
Ecart type	**0.57**	**0.201**	**60.6 %**

Recyclons les données oubliées

A un moment vous avez pu lire:

On dispose en fait de $n_{\max} = 200$ *données* $\left(M_{fuel}^{*,i}, SFC^{*,i}\right)_{i=1\dots n_{\max}}$ *mais on en utilise que* 32 *valeurs de* $M_{fuel}^{*,i}$ *et aucunes valeurs de* $SFC^{*,i}$*, on utilisera tout le monde plus tard....*

A fin de se persuader que notre procédure est raisonnable, on estime maintenant θ uniquement à l'aide des n_{max} données réelles, et on va comparer cette estimation à l'estimation précédente. On trouve alors

	Echantillon de référence	SFC estimé	Erreur relative
Moyenne	**17.49**	**17.397**	**0.5 %**
Ecart type	**0.57**	**0.201**	**60.6 %**

Commentaires

1. *Notre procédure semble plutôt bien estimer la moyenne et plutôt mal estimer l'écart type.*
2. *Erreur statistique* $n = 32$ *c'est petit (mais on fait avec ce que l'on a sous la main).*
3. *Erreur de modélisation*
   1. *Erreur de modèle physique* voir le modèle*: approximation de notre modèle n'est pas parfaite*
   2. *Erreur de modélisation pour la loi de SFC (nous allons essayer de faire mieux).*

A propos de la modélisation de la loi de SFC

Vous avez aussi pu lire

Dans un premier temps de grands experts en aéronautique nous proposent de prendre $\epsilon = \sigma_{SFC}\epsilon_{SFC}$ *avec* $\epsilon_{SFC} \sim \mathcal{N}_{[-3,3]}(0,1)$ *avec* $\mu_{SFC} \in [15,20]$ *et* $\sigma_{SFC} \in [s,1]$ *(s petit).*

A propos de la modélisation de la loi de SFC

Vous avez aussi pu lire

Dans un premier temps de grands experts en aéronautique nous proposent de prendre $\epsilon = \sigma_{SFC}\epsilon_{SFC}$ *avec* $\epsilon_{SFC} \sim \mathcal{N}_{[-3,3]}(0,1)$ *avec* $\mu_{SFC} \in [15,20]$ *et* $\sigma_{SFC} \in [s,1]$ *(s petit).*

Grands experts en aéronautique = N. Rachdi, J.C Fort, T. Klein.
Que conseils les vrais experts en aéronautique ?
Des ingénieurs Airbus spécialistes des moteurs considèrent que la loi de SFC est correctement modélisée par une v.a de type exponentielle: c'est à dire de la forme

$$\left\{ p(u;\theta) = \theta_2\, e^{-\theta_2(u-\theta_1)}\, \mathbf{1}_{[\theta_1,+\infty[}\,, \quad \theta = (\theta_1,\theta_2) \in \mathbb{R}_+ \times \mathbb{R}_+^* \right\}.$$

A propos de la modélisation de la loi de SFC

Vous avez aussi pu lire

Dans un premier temps de grands experts en aéronautique nous proposent de prendre $\epsilon = \sigma_{SFC}\epsilon_{SFC}$ *avec* $\epsilon_{SFC} \sim \mathcal{N}_{[-3,3]}(0,1)$ *avec* $\mu_{SFC} \in [15,20]$ *et* $\sigma_{SFC} \in [s,1]$ *(s petit).*

Grands experts en aéronautique = N. Rachdi, J.C Fort, T. Klein.
Que conseils les vrais experts en aéronautique ?
Des ingénieurs Airbus spécialistes des moteurs considèrent que la loi de SFC est correctement modélisée par une v.a de type exponentielle: c'est à dire de la forme

$$\left\{ p(u;\theta) = \theta_2\, e^{-\theta_2(u-\theta_1)}\, \mathbf{1}_{[\theta_1,+\infty[}\,, \quad \theta = (\theta_1,\theta_2) \in \mathbb{R}_+ \times \mathbb{R}_+^* \right\}.$$

On obtient avec ce modèle pour SFC les résultats suivants

	Echantillon de référence	SFC^{exp} $(n=32)$	Erreur relative
Mean	**17.49**	**17.52**	**0.17 %**
Stand. dev.	**0.57**	**0.29**	**49.12 %**

A comparer avec

	Echantillon de référence	SFC estimé	Erreur relative
Moyenne	**17.49**	**17.397**	**0.5 %**
Ecart type	**0.57**	**0.201**	**60.6 %**

Revenons au début de l'exposé ▸ Retour au début

Revenons au début de l'exposé Retour au début

1. Les Y_i proviennent de données expérimentales couteuses (soit en euros, soit en temps de calculs)
2. Les Z_k proviennent de modèles approchant la réalité. Ces modèles sont de la forme

$$Z = h(X, \theta)$$

Dans l'exemple on avait

$$M_{fuel} = h(\epsilon, \theta) = (M_{empty} + M_{pload}) \left(exp \left(\frac{(\mu_{SFC} + \sigma_{SFC}\, \epsilon_{SFC}) \cdot g \cdot Ra}{(V_{nom} + \epsilon_V) \cdot (F_{nom} + \epsilon_F)} \cdot 10^{-3} \right) - 1 \right)$$

X représente les variables d'entrées de notre modèle et $\theta \in \Theta$ sont des paramètres à estimer.

Remarque (Importante)

Les variables X intervenant dans le modèle approché ne sont pas les mêmes que les conditions initiales des vraies expériences. Ainsi, on ne suppose pas que les vraies données Y sont de la forme $Y = f(X)$.

Objectif et cadre formel

- Objectif: Construire un simulateur aléatoire (de la réalité) $h(X, \hat{\theta})$ qui prédit le mieux possible une caractéristique donnée de la distribution des données $Y_1, \ldots, Y_n$.
- Formalisation mathématique:
  1. Les données expérimentales $Y_1, \ldots, Y_n$ sont de loi inconnue $\mathcal{Q}$ à valeurs dans $\mathcal{Y}$ inclus dans un compact et admettant une densité f.
  2. Le modèle réduit h
  $$\begin{array}{rccc} h: & \mathcal{X} \times \Theta & \to & \mathcal{Y} \\ & (x, \theta) & \mapsto & h(x, \theta) \end{array}$$
  3. Les v.a. d'entrée $X \in \mathcal{X}$ ont une distribution inconnue par contre on sait obtenir via simulation un échantillon de taille $m \gg n$. Ce qui permet de construire un m-échantillon de sorties simulées
  $$h(X_1, \theta), \ldots, h(X_m, \theta).$$

Objectif et cadre formel

- Une quantité d'intérêt pour une loi μ est une certaine fonctionnelle de μ notée $\rho(\mu)$ (par exemple la moyenne, la variance de la loi, sa fonction de répartition ...). On notera $\rho_h(\theta)$ la quantité d'intérêt de la loi de $h(X, \theta)$ Détail quantité d'intérêt.
  Pour fixer les idées, supposons que la quantité d'intérêt soit la moyenne alors

$$\rho_h(\theta) = \mathbb{E}_X\left(h(X, \theta)\right)$$

- Contraste et fonction de risque

$$\psi : \rho \mapsto \psi(\rho, \cdot),\ \mathcal{R}_\psi(h, \theta) = \mathbb{E}_Y \psi(\rho_\theta(h), Y)$$

  Dans l'exemple précédent $\psi(\rho, y) = (y - \rho)^2$ et $\mathcal{R}_\psi(\rho) = \mathrm{Var}(Y) + (\mathbb{E}(Y) - \rho_h(\theta))^2$.

Objectif et cadre formel

Une fois fixé la quantité d'intérêt et le contraste, le véritable objectif est maintenant d'estimer

$$\theta^\star \in Argmin_{\theta\in\Theta}\mathcal{R}_\psi(h,\theta).$$

La fonction $\mathcal{R}_\psi(h,\theta)$ n'est pas calculable car

1. La loi $\mathcal{Q}$ des Y est inconnue
2. La loi de $h(X,\theta)$ n'est connue qu'à travers ses réalisations

On va donc faire deux approximations successives

1. On approche le risque par sa version empirique en Y

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\rho_h(\theta),Y_i\right)$$

2. On approche $\rho_h(\theta)$ à l'aide des m réalisations $h(X_1,\theta),\ldots,h(X_m,\theta)$.

$$\rho_h^m=\frac{1}{m}\sum_{j=1}^{m}\tilde{\rho}\left(h(X_j,\theta)\right)$$

Dans le cas de la moyenne $\tilde{\rho}\left(h(X_j,\theta)\right)=h(X_j,\theta)$.

Borne pour le risque

On estime alors θ par

$$\widehat{\theta} = \mathit{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} \Psi\left(\frac{1}{m}\sum_{j=1}^{m} \widetilde{\rho}(h(\mathcal{X}_j,\theta))\,,\ Y_i\right)\ .$$

Théorème

Sous H_1, H_2 et H_3 ▸ Les hypothèses *, alors $\forall\varepsilon>0$, Il existe K_1 et K_2 t.q. avec proba supérieure à $1-2\varepsilon$ on a*

$$\mathcal{R}_\Psi(h,\widehat{\theta}) \leq \inf_{\theta\in\Theta}\left(\mathcal{R}_\Psi(h,\theta)\right) + \frac{K_1}{\sqrt{n}}\left(1+\sqrt{\frac{n}{m}}(K_2+B_m)\right)$$

où les constantes K_1 et K_2 dépendent des deux processus empiriques de A_Ψ, M. B_m est un facteur de biais dépendant de $B_h(m)$.

Un exemple d'illustration

Prenons Y de la forme

$$Y \quad = \quad \sin(\xi) + 0.01\,\varepsilon,$$

ξ et ε sont deux gaussiennes centrées réduites indépendantes.
On se donne $Y_1, ..., Y_n$ i.i.d. On désire estimer la densité f de Y, à partir de $Y_1, ..., Y_n$ et du modèle h donné par

$$h(\mathcal{X}, \theta) \quad = \quad \theta_1 + \theta_2\,\mathcal{X} + \theta_3\,\mathcal{X}^3\,, \quad \mathcal{X} \sim \mathcal{N}(0,1),\ \theta = (\theta_1, \theta_2, \theta_3).$$

On va dans un premier temps estimer θ puis estimer f à l'aide de la densité de $h(X, \widehat{\theta})$.
On fait tourner notre procédure d'estimation de θ puis on estime la densité f, on obtient les résultats suivants:

	$KL(f, \widehat{f})$
$n = 50,\ m = 10^3$	$(4.9 \pm 2).10^{-2}$
$n = 500,\ m = 5.10^3$	$(2.3 \pm 0.4).10^{-2}$
$n = 1000,\ m = 10^4$	$(1.1 \pm 0.2).10^{-2}$

Merci pour votre attention

Un grand merci aux organisateurs

Les 3 hypothèses

1. H_1 = le contrôle du terme de biais ▸ Le biais
2. H_2 =contrôle des processus empiriques.
   En gros, deux type de contrôle: la tension du sup des processus empiriques sur certaines classes de fonctions. Le bon comportement du processus empirique des Y_i en fonction de celui des X_i
3. H_3 = la régularité Lipschitz du contraste Ψ.

▸ Retour au Théorème

Une quantité d'intérêt relative à une mesure μ est une fonctionnelle de μ à valeur dans $\mathcal{F}$, $\mathcal{F}$est par exemple $\mathbb{R}$ si la quantité est l'espérance, cela peut être un espace de fonction si la quantité d'intérêt est une densité...
Dans la suite, on munira $\mathcal{F}$ d'une norme.
Contraste et fonction de risque .
Un contraste est une application à valeurs dans $L_1(\mathbb{Q})$

$$\begin{array}{rcl} \Psi : \mathcal{F} & \longrightarrow & L_1(\mathbb{Q}) \\ \rho & \longmapsto & \Psi(\rho, \cdot) : y \in \mathcal{Y} \longmapsto \Psi(\rho, y)\,, \end{array} \tag{1}$$

tel que

$$\rho^* = \mathit{argmin}_{\rho \in \mathcal{F}} \mathbb{E}_Y \, \Psi\,(\rho, Y)$$

est *unique*.
On appelle **fonction de risque** l'application

$$\forall\, \rho \in \mathcal{F}\,, \quad \mathcal{R}_\Psi(\rho) := \mathbb{E}_Y \, \Psi\,(\rho, Y)\ .$$

▸ Retour cadre formel

$$\sigma_h^m(\theta) := \|\rho_h^m(\theta) - \rho_h(\theta)\|_{\mathcal{F}}\,.$$

Alors en utilisant inégalité triangulaire

$$\sigma_h^m(\theta) \quad = \quad \left\| \frac{1}{m} \sum_{j=1}^{m} \left[\widetilde{\rho}(h(\mathcal{X}_j, \theta)) - \mathbb{E}_{\mathcal{X}} \widetilde{\rho}(h(\mathcal{X}, \theta)) \right] \right\|_{\mathcal{F}} + B_h^m(\theta)$$

avec $B_h^m(\theta) := \|\mathbb{E}_{\mathcal{X}} \widetilde{\rho}(h(\mathcal{X}, \theta)) - \rho_h(\theta)\|_{\mathcal{F}}$ le terme de biais.

$$H_1 : \sup_{\theta} Bh^m(\theta) \leq B_h(m) < +\infty$$

▸ Retour Hypothèses

Où on parle de contraste pour la première fois

La densité des Y est une fonction f inconnue, lorsque l'on décide d'approcher f par une densité g, on peut quantifier la qualité de l'approximation à l'aide de la distance de Kullback modifiée

$$K(f,g) = \int \log(f/g) f dx - \int \log(x) f(x) dx.$$

On va réécrire autrement K

$$K(f,g) = \mathbb{E}_Y \left(-\log g(Y) \right)$$

La fonction $\psi : g \mapsto [y \mapsto -\log(g(y))]$ est appelée contraste.
La quantité $\mathbb{E}_Y\big(-\log\left(g(Y)\right)\big) := R_\psi(g)$ est appelée risque de g associé au contraste ψ. Bien entendu en pratique la quantité $R_\psi(g)$ est inaccessible, on se dépêche de la remplacée par son équivalent empirique soit

$$-\frac{1}{n}\sum_{i=1}^{n} \log g(Y_i).$$

Retour à l'estimation