M-estimation in Complex Modeling

Nabil Rachdi (PhD Student)

nabil.rachdi@eads.net

▶ Colloque Jeunes Probabilistes et Statisticiens (3 → 7 of May – Mont Dore) ◀

Advisors:

Jean-Claude Fort (Paris V), Thierry Klein (Toulouse III)
Fabien Mangeant (EADS IW), Régis Lebrun (EADS IW).

Outline

1. Motivations and notations
2. Model Selection
3. Risk excess bound
4. Theorical Results
5. Examples

General framework

Variable of Interest: Y, density f, probability measure $\mathbb{Q}$ unknown

Feature: Density distribution, mean, threshold probability, quantile etc...

Experimental data: $\mathcal{Y}_n^{exp} = Y_1^{exp}, ..., Y_n^{exp}$ (*a priori* training data) supposed i.i.d from $\mathbb{Q}$

- Link to history
- Arise from experiments, complex codes etc...
- Small number
- Difficult to obtain

Simulation model: $h \in \mathcal{H} : (\mathbf{x}, \boldsymbol{\theta}) \in \mathcal{X} \times \Theta \mapsto y = h(\mathbf{x}, \boldsymbol{\theta})$
For a configuration $\mathbf{x}_0$, the model gives $y_0 = h(\mathbf{x}, \boldsymbol{\theta}_0) \Rightarrow$ Deterministic !!

General framework

Variable of Interest: Y, density f, probability measure $\mathbb{Q}$ unknown

Feature: Density distribution, mean, threshold probability, quantile etc...

Experimental data: $\mathcal{Y}_n^{exp} = Y_1^{exp}, ..., Y_n^{exp}$ (*a priori* training data) supposed i.i.d from $\mathbb{Q}$

- Link to history
- Arise from experiments, complex codes etc...
- Small number
- Difficult to obtain

Simulated data: Consider that $\mathcal{X} \leftrightarrow (\mathcal{X}, \mathcal{B}, \mathbb{P}_X)$ i.e **x** (deterministic) $\leftrightarrow$ **X** (random).
For $m \geq 0$, let $\mathcal{Y}_m^{sim} = Y_1^{sim}, ..., Y_m^{sim}$ (depends on the model h and the parameter $\boldsymbol{\theta}$)

- $h \in \mathcal{H}$ (set of models), $\boldsymbol{\theta} \in \Theta$ (set of parameters)
- $Y_i^{sim} = h(X_i, \boldsymbol{\theta}), \quad i = 1, ..., m, \ X_i$ i.i.d random variables (distribution $\mathbb{P}_X$).
- $X \sqcup Y^{exp}$

General framework

Variable of Interest: Y, density f, probability measure $\mathbb{Q}$ unknown

Feature: Density distribution, mean, threshold probability, quantile etc...

Experimental data: $\mathcal{Y}_n^{exp} = Y_1^{exp}, ..., Y_n^{exp}$ (*a priori* training data) supposed i.i.d from $\mathbb{Q}$

- Link to history
- Arise from experiments, complex codes etc...
- Small number
- Difficult to obtain

Simulated data: Consider that $\mathcal{X} \leftrightarrow (\mathcal{X}, \mathcal{B}, \mathbb{P}_X)$ i.e **x** (deterministic) $\leftrightarrow$ **X** (random).
For $m \geq 0$, let $\mathcal{Y}_m^{sim} = Y_1^{sim}, ..., Y_m^{sim}$ (depends on the model h and the parameter $\boldsymbol{\theta}$)

- $h \in \mathcal{H}$ (set of models), $\boldsymbol{\theta} \in \Theta$ (set of parameters)
- $Y_i^{sim} = h(X_i, \boldsymbol{\theta}), \quad i = 1, ..., m, \ X_i$ i.i.d random variables (distribution $\mathbb{P}_X$).
- $X \sqcup Y^{exp}$

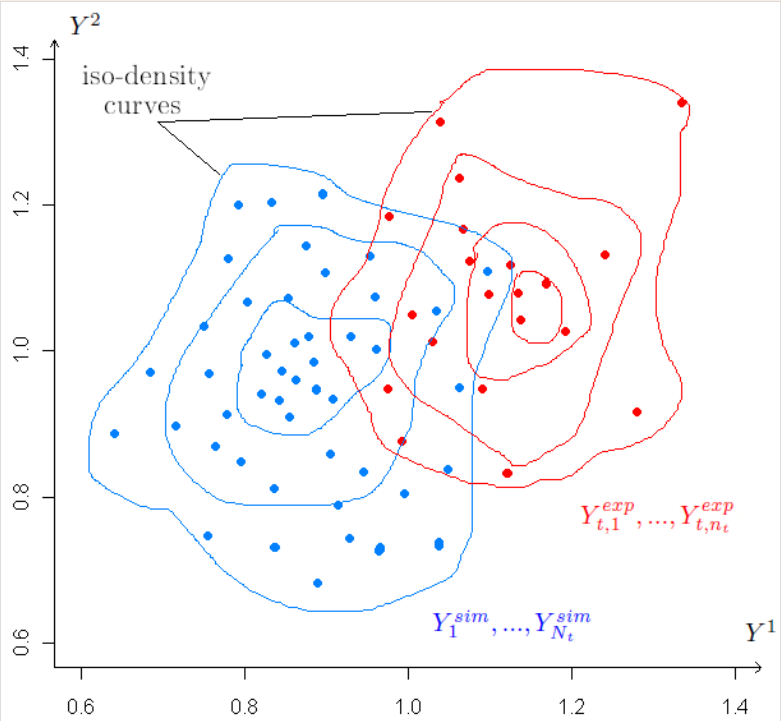
Goal: Use Simulated data to improve feature estimation of Y:

$\Rightarrow$ 1. Estimation procedure: choice of the model h, and parameter $\boldsymbol{\theta}$

$\Rightarrow$ 2. Study of a feature based on $(\widehat{Y}_1^{sim}, ..., \widehat{Y}_m^{sim})$ (*a posteriori* training data)
to compare with the feature based on $(Y_1^{exp}, ..., Y_n^{exp})$

Simulated data estimation

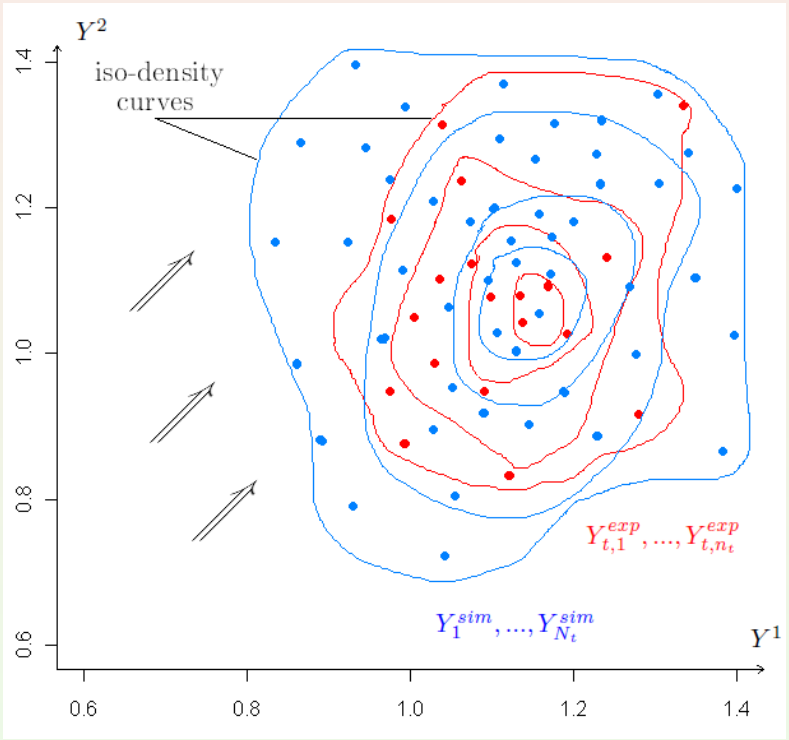
- Illustration of Experimental & Simulated Data: Example of a 2D-Performance $Y = (Y^1, Y^2)$



- Choice of $h \in \mathcal{H}$ and $\boldsymbol{\theta} \in \Theta$? $\Rightarrow$ driven by the feature considered

Simulated data estimation

- Illustration of Experimental & Simulated Data: Example of a 2D-Performance $Y = (Y^1, Y^2)$



- Choice of $h \in \mathcal{H}$ and $\boldsymbol{\theta} \in \Theta$? $\Rightarrow$ driven by the feature considered

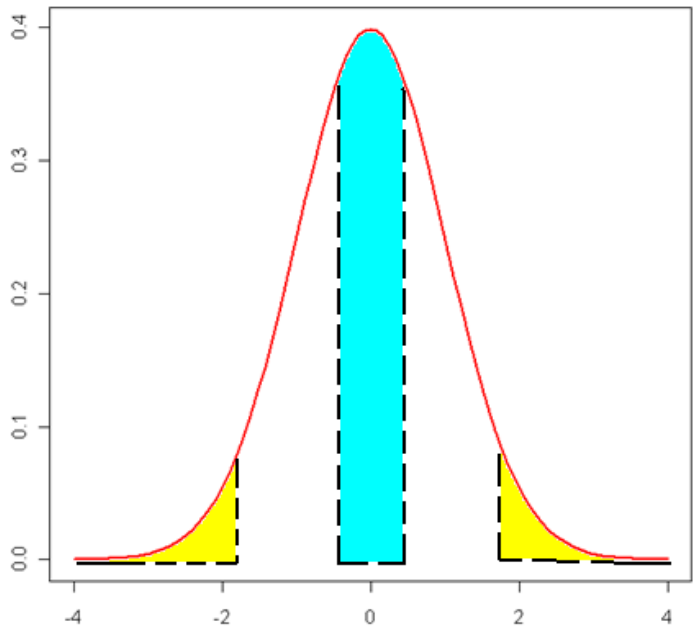
Definition: Feature

Let $\mathcal{P}$ be the set of all probability measures, $\mathbb{D}$ a metric space, we define a **feature** as a function

$$\begin{array}{rccc} \rho : & \mathcal{P} & \longrightarrow & \mathbb{D} \\ & \mu & \longmapsto & \rho(\mu). \end{array}$$

Some feature:

- $\rho(\mu) = \mathbb{E}_{W\sim\mu}(W)$ (mean), q_W^α (α-quantile) $\Rightarrow \mathbb{D} = \mathbb{R}$
- $\rho(\mu) = \mathbb{P}(W > s) \quad \Rightarrow \mathbb{D} = [0, 1]$
- $\rho(\mu) = f_\mu \quad \Rightarrow \mathbb{D} = \{\textit{set of distributions}\}$
- etc ...



Choice of $(h, \boldsymbol{\theta})$ for a Feature ρ

In our framework:

- the model

$$\mathbb{P}_X \xrightarrow{(h,\boldsymbol{\theta})} \mathbb{P}_{h,\boldsymbol{\theta}} \xrightarrow{\text{feature } \rho} \rho(\mathbb{P}_{h,\boldsymbol{\theta}}) := \rho_h(\boldsymbol{\theta})$$

  Important Remark: the feature $\rho_h(\boldsymbol{\theta})$ can be unreachable $\rightarrow$ we say that h is complex

- denote the "true" feature by $\rho(\mathbb{Q}) := \rho^*$
- For fix $h \in \mathcal{H}$, choice of $\boldsymbol{\theta}$

$$\boldsymbol{\theta}_0 = \underset{\Theta}{\text{Argmin}}\, \mathcal{D}(\rho_h(\boldsymbol{\theta}), \rho^*)$$

  with $\mathcal{D}$ *a measure of distance* $\mathbb{D} \times \mathbb{D}$.

- We use the form:

$$M(h, \boldsymbol{\theta}) = \int \gamma_{h,\boldsymbol{\theta}}(y)\, \mathbb{Q}(dy)$$

  where the function $\gamma_{h,\boldsymbol{\theta}}$ is called contrast of $(h, \boldsymbol{\theta})$.
  We have

$$\gamma_{h,\boldsymbol{\theta}} = \Psi\left(\rho_h(\boldsymbol{\theta})\right)$$

  with Ψ some operator on $\mathbb{D}$.

- **Example of contrasts:**

  If the feature is the density
  - $\gamma_{h,\boldsymbol{\theta}}(y) = -\ln\left(\rho_h(\boldsymbol{\theta})\right)(y) \quad \Rightarrow \quad M(h,\boldsymbol{\theta}) = K(\rho^*, \rho_h(\boldsymbol{\theta}))$
  - $\gamma_{h,\boldsymbol{\theta}}(y) = ||\rho_h(\boldsymbol{\theta})||_2^2 - 2\rho_h(\boldsymbol{\theta})(y) \quad \Rightarrow \quad M(h,\boldsymbol{\theta}) = ||\rho^* - \rho_h(\boldsymbol{\theta})||_2^2.$
  - etc...

  The same for mean, quantile, threshold probability etc...

- **Criterion to minimize:**

$$M(h,\boldsymbol{\theta}) = \int \gamma_{h,\boldsymbol{\theta}}(y)\, \mathbb{Q}(dy)$$

- Assume that $(h^*, \boldsymbol{\theta}^*)$ is the unique minimum of $M(\cdot,\cdot)$

- **Difficulties:**
  - The measure $\mathbb{Q}$ is unknown (Classical !)
  - For complex models h, the feature $\rho_h(\boldsymbol{\theta})$ is unreachable so the contrast too $(\gamma_{h,\boldsymbol{\theta}} = \Psi\left(\rho_h(\boldsymbol{\theta})\right))$.

$$M(h,\boldsymbol{\theta}) = \int \gamma_{h,\boldsymbol{\theta}}(y)\,\mathbb{Q}(dy)$$

- **Alternative: Use of Experimental & Simulated data**
  - Replace $\mathbb{Q}$ by its empirical version $\mathbb{Q}_n$ (depends on n-Experimental data $\mathcal{Y}_n^{exp}$)
    We have
    $$M_n(h,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \gamma_{h,\boldsymbol{\theta}}(Y_i^{exp})$$
  - Since $\gamma_{h,\boldsymbol{\theta}}^m = \Psi\left(\rho_h^m(\boldsymbol{\theta})\right)$, replace the feature $\rho_h(\boldsymbol{\theta})$ by an estimator $\rho_h^m(\boldsymbol{\theta})$ (depends on m-Simulated data $\mathcal{Y}_m^{sim}$).

▷ Pakes and Pollard (1989) studied estimators minimizing an approximation $\widetilde{M}_n(h,\boldsymbol{\theta}) \sim M_n(h,\boldsymbol{\theta})$ (Optimization Estimators).

▷ They prove $\sqrt{n}$-consistency and give limit theorems

▷ The conditions are given on the (approximated) criterion $\widetilde{M}_n(h,\boldsymbol{\theta})$

▶ We want conditions on the simulation (quantification) and on the feature (regularity)

- **Criterion to minimize in practice:**

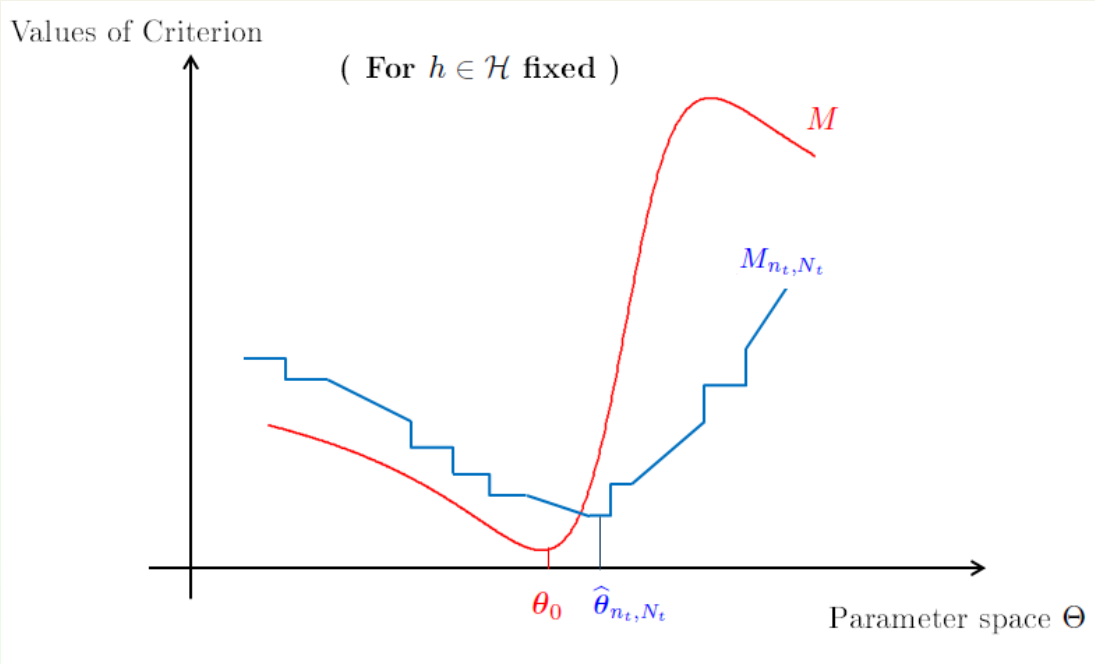
$$M_{n,m}(h,\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \gamma^{m}_{h,\boldsymbol{\theta}}(Y_i^{exp})$$

- **Estimator of** $\theta_0(h) = \text{Argmin}_{\boldsymbol{\theta}\in\Theta}\ M(h,\boldsymbol{\theta})$

$$\widehat{\boldsymbol{\theta}}_{n,m}(h) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}}\ M_{n,m}(h,\boldsymbol{\theta})\,.$$

First question: Consistency

$$\widehat{\boldsymbol{\theta}}_{n,m}(h) \underset{\substack{n\to+\infty\\ m\to+\infty}}{\longrightarrow} \boldsymbol{\theta}_0(h)\,?$$



Source of errors

Proposition: Model Error Upper Bound

We prove

$$\underbrace{M(h,\widehat{\boldsymbol{\theta}}_{n,m}(h)) - M(h^*,\boldsymbol{\theta}^*)}_{\text{risk excess of } (h,\,\widehat{\boldsymbol{\theta}}_{n,m}(h))} \lesssim \frac{1}{\sqrt{n}}||\,\mathbb{G}_n\gamma^m_{h,.}\,||_\Theta + ||\mathcal{E}^m_h||_\Theta + \Delta_h$$

- **Variance terms (random terms):**
  - $\frac{1}{\sqrt{n}}||\,\mathbb{G}_n\gamma^m_{h,.}\,||_\Theta = \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\gamma^m_{h,\boldsymbol{\theta}}(Y_i^{exp}) - \mathbb{E}_Y(\gamma^m_{h,\boldsymbol{\theta}}(Y))\right)\right|$ (deviation)

    $\Rightarrow$ *Estimation Error of Statistical Data* $\quad(\mathbb{G}_n := \sqrt{n}(\mathbb{Q}_n - \mathbb{Q}))$
  - $||\mathcal{E}^m_h||_\Theta = \sup_{\boldsymbol{\theta}\in\Theta}||\,\gamma^m_{h,\boldsymbol{\theta}} - \gamma_{h,\boldsymbol{\theta}}\,||_{1,\mathbb{Q}}$, with $||g||_{1,\mathbb{Q}} = \int_{\mathbb{R}}|g(y)|\,\mathbb{Q}(dy)$

    $\Rightarrow$ *Simulation Error*
- **Bias term :**
  - $\Delta_h = M(h,\boldsymbol{\theta}_0(h)) - M(h^*,\boldsymbol{\theta}^*)$

    $\Rightarrow$ *Approximation Error of the model h*

Link with classical methods

Classical Inequality

Recall $M(h,\theta) = \int \gamma_{h,\theta}(y)\,\mathbb{Q}(dy)$, we have

$$M(h,\widehat{\theta}_n(h)) - M(h^*,\theta^*) \leq \frac{1}{\sqrt{n}} || \,\mathbb{G}_n\gamma_{h,.}\,||_\Theta + \Delta_h$$

- **For no complex models:** (Statistical models, simplified models etc...)

  $\Rightarrow$ the feature $\rho_h(\theta)$ is reachable $\Rightarrow$ Simulation is useless

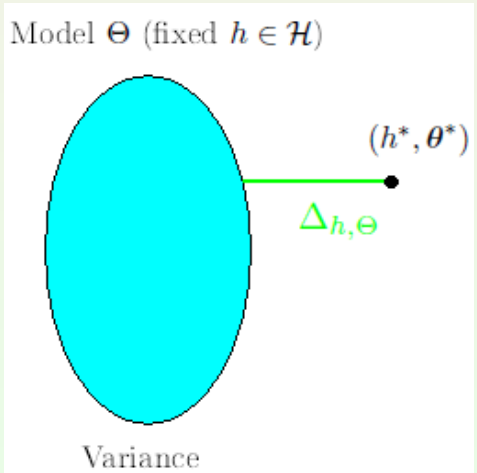
Advantages

- Well studied
- Maximum Likelihood, Regression etc...

Drawback

- Δ_h can be large (due to simplification of h)

  Trade-off Bias-Variance



OUR GOAL: Extend the theory to physical models

Theorical result

- **Recall:**

  model h + parameter $\boldsymbol{\theta} \Rightarrow$ Feature $\rho_h(\boldsymbol{\theta}) \Rightarrow$ Contrast $\gamma_{h,\boldsymbol{\theta}}$

  In our framework:

  model h + parameter $\boldsymbol{\theta} \Rightarrow$ Feature $\rho_h^m(\boldsymbol{\theta}) \Rightarrow$ Contrast $\gamma_{h,\boldsymbol{\theta}}^m$

- **Inequality:**

$$M(h, \widehat{\boldsymbol{\theta}}_{n,m}(h)) - M(h^*, \boldsymbol{\theta}^*) \lesssim \frac{1}{\sqrt{n}} || \mathbb{G}_n \gamma_{h,.}^m ||_\Theta + ||\mathcal{E}_h^m||_\Theta + \Delta_h$$

- **Key point:** Let the class of functions

$$\Gamma_h^m := \{\gamma_{h,\boldsymbol{\theta}}^m, \boldsymbol{\theta} \in \Theta\} = \{\Psi(\rho_h^m(\boldsymbol{\theta})), \boldsymbol{\theta} \in \Theta\}$$

  - the class is random (simulated data)
  - the class changes with m

- **Important Remark:** $|| \mathbb{G}_n \gamma_{h,.}^m ||_\Theta = || \mathbb{G}_n ||_{\Gamma_h^m}$

  Main difficulty : Prove the tightness of the random variable sequence $\left(|| \mathbb{G}_n ||_{\Gamma_h^m} \right)_{n,m}$

Definition: $L_2(\mathbb{Q})$ Bracketing numbers and Entropy with bracketing

▷ Let $\mathcal{G}$ be a class of functions. Given two functions l, u, the bracket $[l, u]$ is the set of all functions g with $l \leq g \leq u$. An ε-bracket is a bracket $[l, u]$ with $||u - l||_{2,\mathbb{Q}} < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{G}, L_2(\mathbb{Q}))$ is the minimum number of ϵ-brackets needed to cover the class of functions $\mathcal{G}$.

▷ The entropy with bracketing is the logarithm of the bracketing number.

Definition: $L_2(\mathbb{Q})$ Bracketing integral

The bracketing integral is defined as

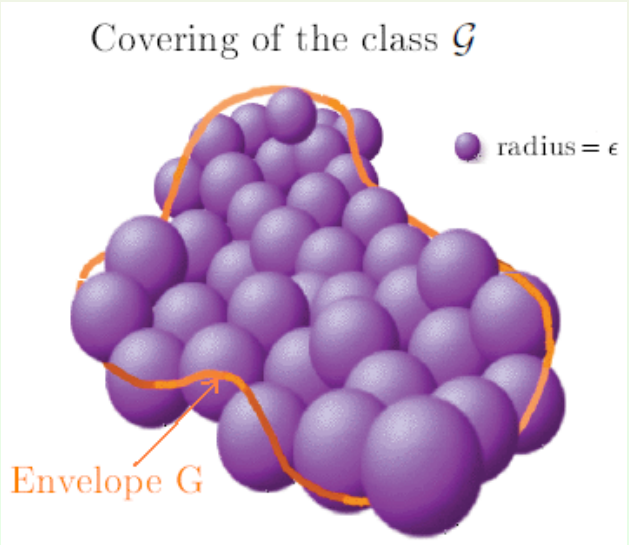
$$J_{[]}(\delta, \mathcal{G}, L_2(\mathbb{Q})) := \int_0^{\delta} \sqrt{\log N_{[]}(\varepsilon, \mathcal{G}, L_2(\mathbb{Q}))} d\varepsilon \,.$$

References:

v d Vaart & Wellner (1996) *Weak Convergence and Empirical Processes*

v d Vaart (1998) *Asymptotics Statistics*

S van de Geer (2000) *Empirical Processes in M-estimation*



Particular classes

- Recall that $\mathbb{G}_n := \sqrt{n}(\mathbb{Q}_n - \mathbb{Q})$

Definition: Glivenko-Cantelli classes

A class of function $\mathcal{G}$ is $(\mathbb{Q})$-Glivenko-Cantelli if

$$\frac{1}{\sqrt{n}} || \mathbb{G}_n ||_{\mathcal{G}} \xrightarrow[p.s]{} 0 \,.$$

Definition: Donsker classes

A class of function $\mathcal{G}$ is $(\mathbb{Q})$-Donsker if

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \quad in \quad l^{\infty}(\mathcal{G})$$

Theorems

1. If for all $\varepsilon > 0$, $N_{[]}(\varepsilon, \mathcal{G}, L_1(\mathbb{Q})) < \infty$ then $\mathcal{G}$ is Glivenko-Cantelli.
2. If $J_{[]}(1, \mathcal{G}, L_2(\mathbb{Q})) < \infty$ then $\mathcal{G}$ is Donsker.

- The class Γ_h^m is random and changes with m. We need more results.

Our Results

- Theorical study of the Γ_h^m-indexed empirical $\mathbb{G}_n = \sqrt{n}(\mathbb{Q}_n - \mathbb{Q})$ process

Consistency

We prove that $\widehat{\boldsymbol{\theta}}_{n,m} \underset{\substack{n\to+\infty \\ m\to+\infty}}{\longrightarrow} \boldsymbol{\theta}_0$, under some conditions in terms of :

- Model Complexity *(Bracketing Numbers)* (model h + feature regularity)
- Simulation Speed *(Size of Simulated Data set, m)*
- Control of Simulated contrasts *(Modified Lindeberg conditions)*

Limiting Distribution

1. It exists $r_{n,m} \geq 0$ such that $r_{n,m}(\widehat{\boldsymbol{\theta}}_{n,m} - \boldsymbol{\theta}_0) = \boldsymbol{O}_{\mathbb{P}}(1)$
2. Under some conditions on the model h and on the feature ρ, we can define $m_n := \inf\{m \geq 0,\ r_{n,m} \gtrsim n^{\frac{1}{2}-\varepsilon_n}\}$, where $\varepsilon_n > 0$ is a nonincreasing sequence.
3. Under some conditions,
$$\widehat{\boldsymbol{\theta}}_{n,m} - \boldsymbol{\theta}_0 \rightsquigarrow \mathcal{N}(0, \Sigma)$$

- In some applications, for $n = 30$ we have $m_n \sim 10000$

Example of the *Range* study

Phenomenon : $Y = Range$ (distance an aircraft can travel), feature = density distribution

- ***A priori* training data :** Experimental data, $n = 20$, $\mathcal{Y}_n^{exp} = Y_1^{exp}, ..., Y_n^{exp}$

  (obtained from complex model h^* supposed to be the "true")

- **Additional knowledge :** Simulated data, $m = 3000$, $\mathcal{Y}_m^{sim} = Y_1^{sim}, ..., Y_m^{sim}$ from

$$h(X, \theta) = \frac{F\,V}{C_s}\frac{1}{\theta_1}\log\left(\frac{1}{1-\theta_2}\right)$$

  - Uncertain Inputs $X = (F, V, C_s)^T$
  - Parameters $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta$

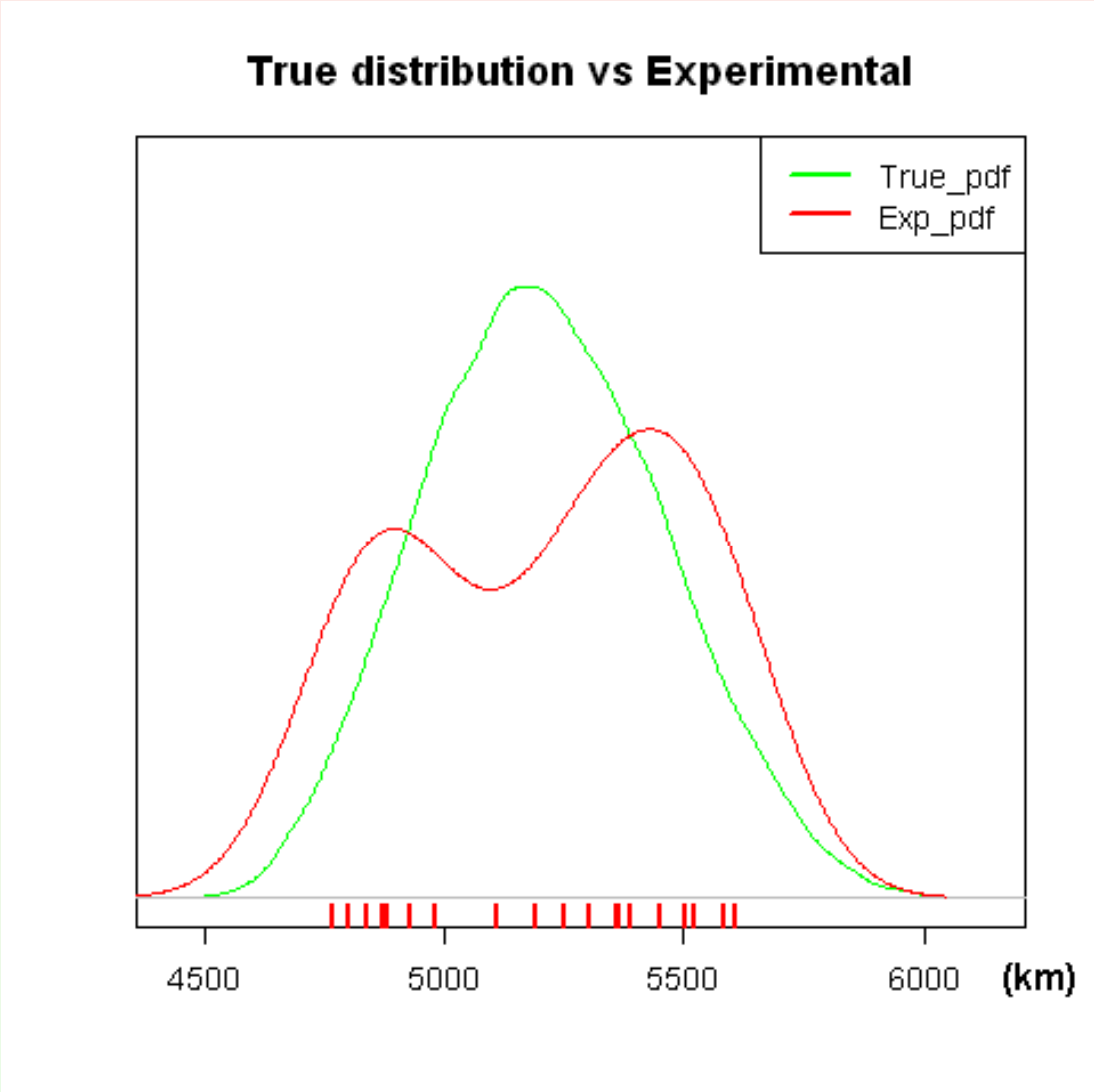
- **Choice of θ ?** $\quad \theta_0(h) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} \int_{\mathbb{R}} \gamma_{h,\boldsymbol{\theta}}(y)\, f(y)\, dy$

$$\gamma_{h,\boldsymbol{\theta}} = -\ln(f_{h,\boldsymbol{\theta}}) \quad f_{h,\boldsymbol{\theta}} \leftrightarrow f_{h,\boldsymbol{\theta}}^m\,(Kernel) \quad f \leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i^{exp}}$$

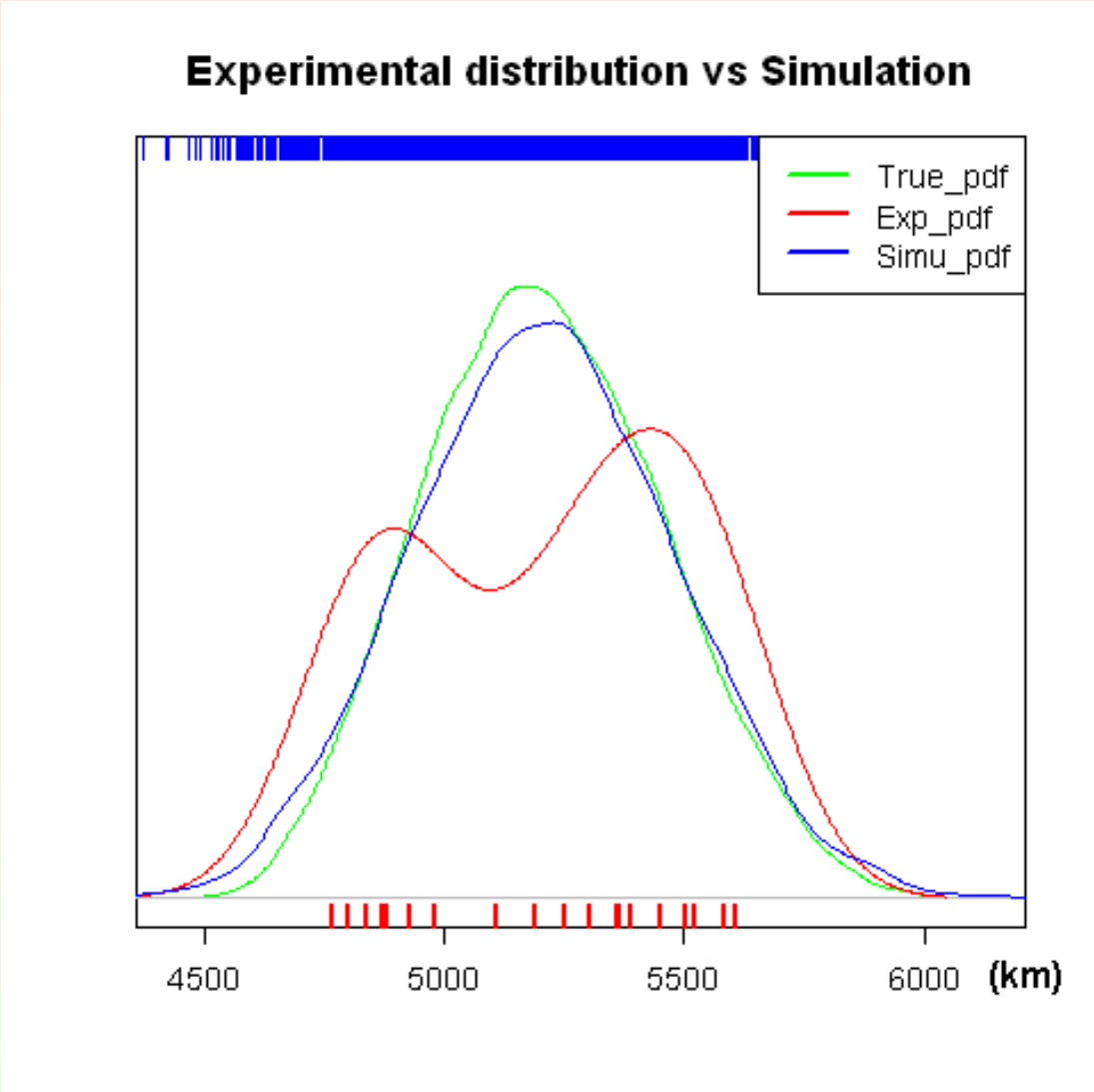
$$\widehat{\boldsymbol{\theta}}_{n,m}(h) = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmin}} -\frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{1}{m}\sum_{j=1}^{m} K_{h_m}(Y_i^{exp} - Y_j^{sim})\right)$$

- ***A posteriori* training data :** $(Y_1^{exp}, ..., Y_n^{exp}, \widehat{Y}_1^{sim}, ..., \widehat{Y}_m^{sim}) \longrightarrow n + m = 3020!$

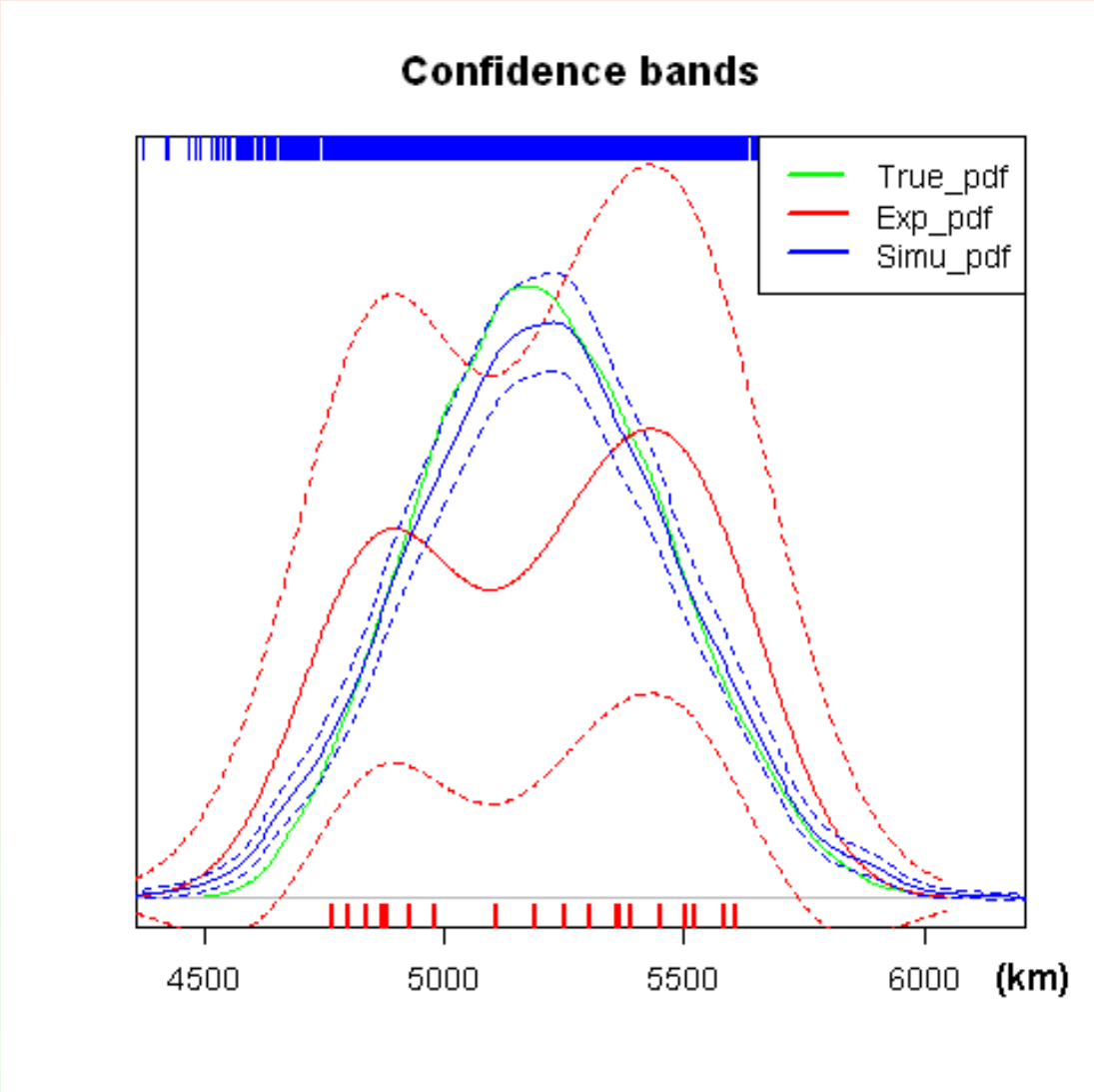
- Feature with *a priori* (Experimental) and *a posteriori* (Experimental + Simulated) training data



- Feature with *a priori* (Experimental) and *a posteriori* (Experimental + Simulated) training data



- Feature with *a priori* (Experimental) and *a posteriori* (Experimental + Simulated) training data



Work in progress

- **Non parametric estimation improvement:**
  - Let φ be a non parametric estimator of the feature ρ

$$\begin{array}{rcl} \varphi : \{\textit{set of samples}\} & \longrightarrow & \mathbb{D} \\ \Lambda & \longmapsto & \varphi(\Lambda). \end{array}$$

  - We note $\varphi_n = \varphi(\mathcal{Y}_n^{exp})$ and $\varphi_m = \varphi(\mathcal{Y}_m^{sim})$, and we consider the aggregate estimator

$$\varphi_{n,m} = \alpha \cdot \varphi_n + (1 - \alpha) \cdot \varphi_m$$

    with $0 < \alpha < 1$.
  - Estimation of $\alpha \longrightarrow$ Simulate or not to simulate ?
  - Asymptotic properties

Thank you for your attention !