EADS Innovation Works

Statistical Learning and computer experiments

Palaiseau, January 2012

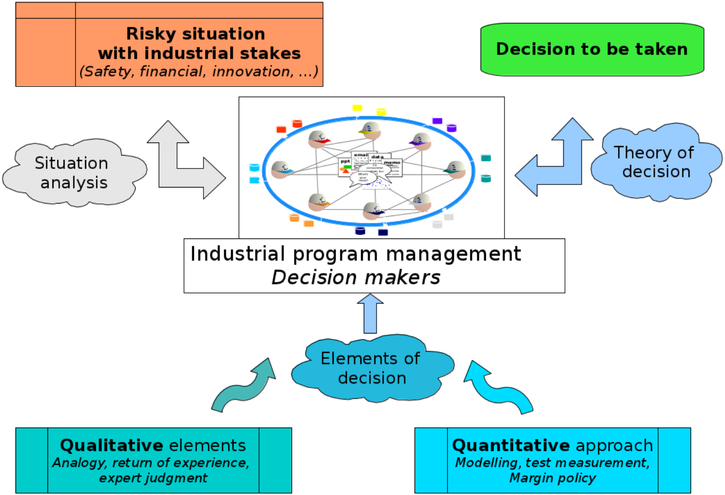
Fabien MANGEANT, Nabil RACHDI

January 9, 2012

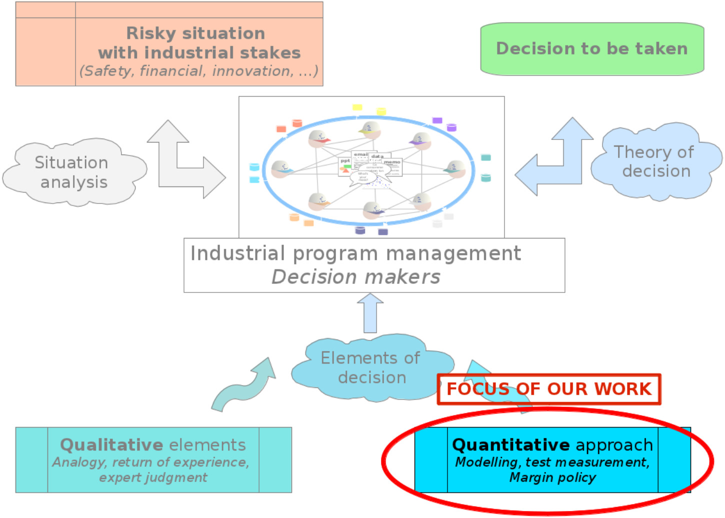
Outline

Outline

Uncertainty analysis in a decision process



Uncertainty analysis in a decision process



Engineering activities during the life cycle of an aircraft

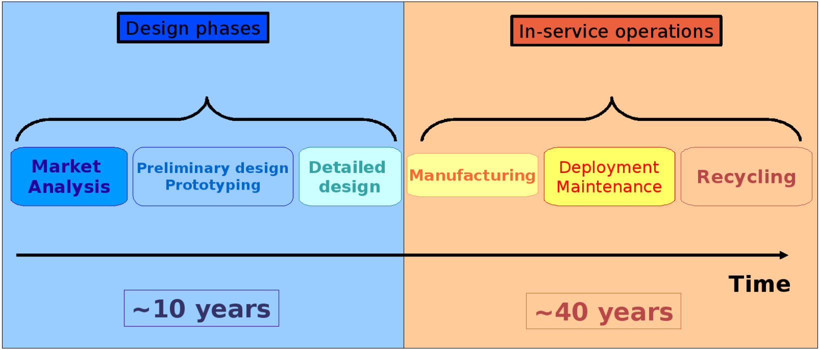


Figure: Life-cycle of a product/service/utility

Scope of the presentation

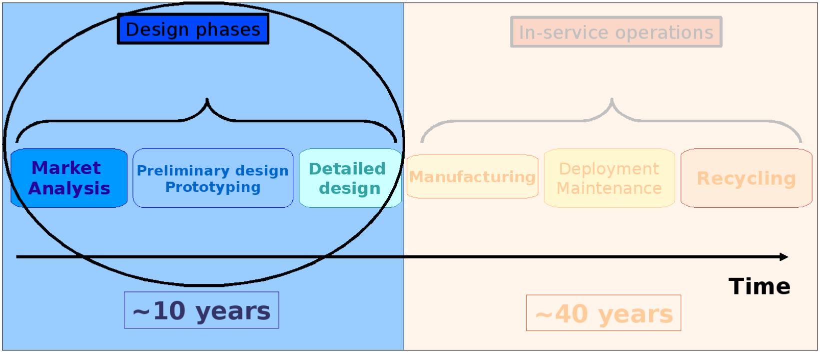


Figure: Design phases

What is our technical objective?

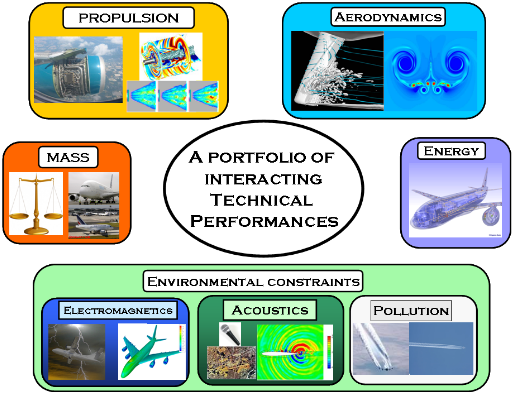


Figure: Portfolio of technical performances

Performances

Aerodynamic: Drag,

Mass: Maximum Weight,

Acoustics: Perceived Noise Level,

Energy: Maximum Electric Power,

Propulsion: Specific Fuel Consumption...

$$\Downarrow$$

$$\mathbf{y}^* = (y_1^*, \ldots, y_Q^*)$$

A naive presentation of the engineering challenge

Description of the situation

- A target $\mathcal{T}$ is given to the variable $\mathbf{y}^*$. This target can evolve during the time of the design.
- These performances are uncontrolled for many reasons (lack of knowledge, variability, approximation, dependency, ...).
- The amount of available information $\mathcal{I}$ for each variable y_i^* evolves during the time of the design (either over the knowledge of the input variables, parameters, mesurements, availability of numerical models).
- At a given time of the design, these technical performances must be estimated with a level of confidence.

A naive presentation of the engineering challenge

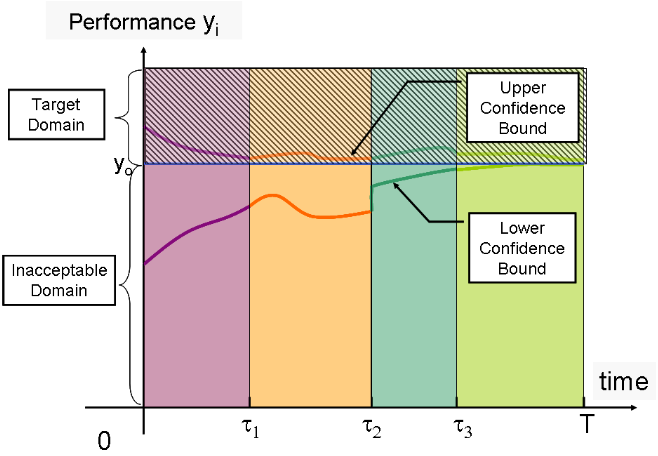


Figure: Evolution of a performance during the design phase

A naive presentation of the engineering challenge

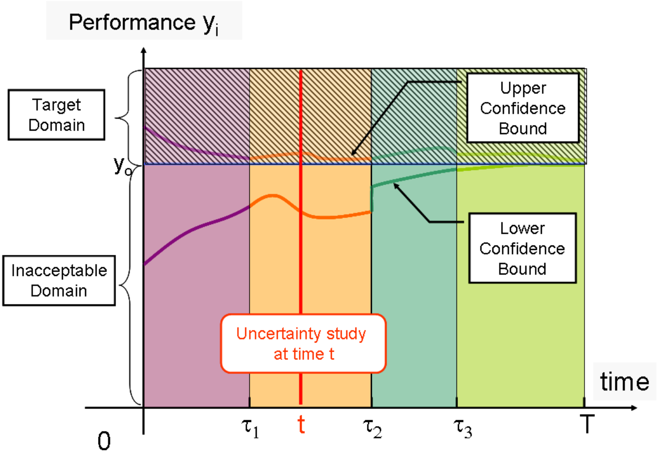


Figure: An uncertainty study at a given time of the design

A naive presentation of the engineering challenge

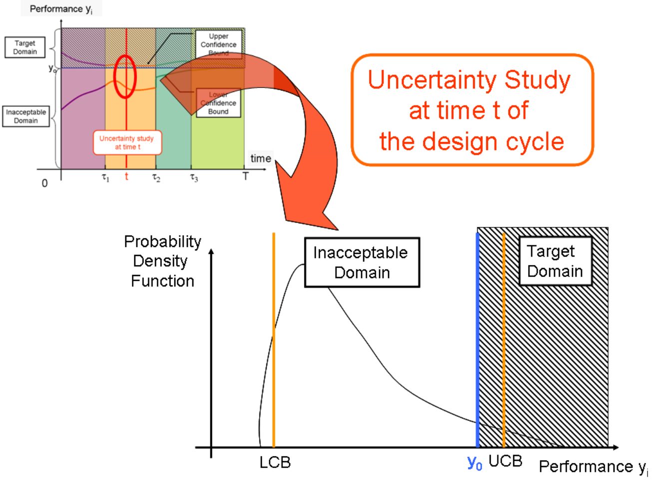


Figure:

A naive presentation of the engineering challenge

Objectives in a mathematical framework

In a probabilistic framework, two main goals can be identified:

1. To control the stochastic behaviour of the performances $\mathbf{y}^*$ to reach the initial or adapted target $\mathcal{T}$.
2. To estimate on-demand some measures of risks during the time of the design.

A naive presentation of the engineering challenge

Objectives in a mathematical framework

In a probabilistic framework, two main goals can be identified:

1. To control the stochastic behaviour of the performances $\mathbf{y}^*$ to reach the initial or adapted target $\mathcal{T}$.
2. **To estimate on-demand one or several measures of risks during the time of the design. –> This is a new discipline for engineers and where we focus our current efforts!**

What kind of information do we manipulate?

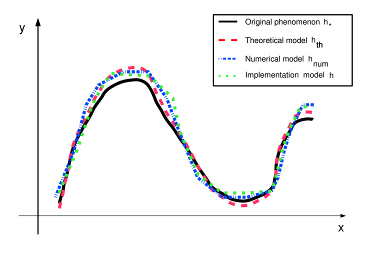
Elements of information

- A **reference database** $(\mathbf{Y}_1^*, \cdots, \mathbf{Y}_n^*)$ that is enriched during the design cycle.
- A **panoply of numerical models** $\mathcal{H} = \{h_1, \cdots, h_D\}$ that is enriched during the design cycle.
- A **quantification of the uncertainties attached to the inputs** of the numerical models represented **by a statistical law** $\mathbb{P}_{\mathbf{X}}$ that is enriched during the design cycle
- A **definition of the target** $\mathcal{T}$ and its associated level of confidence α to be reached that is enriched during the design cycle.
- A **global computational budget** $\mathcal{B}$ that can be allocated at different times of the design cycle.

What kind of information do we manipulate?

What does "model" uncertainty recover in our context?

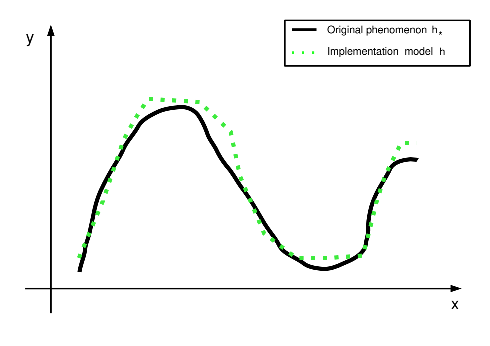
- **Reference model** h^*: Usually not accessible, expression of a natural or a complex technical object.
- **Theoretical model** h_{th}: Scientific expert activity (theoretical solution of a PDE system, ...), corresponding to the level of understanding and simplification of the problem.
- **Numerical model** h_{num}: Numerical solution of the theoretical model (effects of meshing, choice of a numerical scheme)
- **Implementation model** h: Software implementation of the model on a given hardware architecture (computer accuracy, choice of coding rules, ..).



What kind of information do we manipulate?

What does “model” uncertainty recover in our context?

h_i is a numerical representation of the phenomenon, and is represented by a function (also called “model“) belonging to $\mathcal{F}(\mathcal{X}_i \times \Theta_i, \mathcal{Y})$.



What kind of information do we manipulate?

Properties of a numerical model h

- **Dimension**: h is classically a real function belonging to $\mathcal{F}(\mathbb{R}^P \times \mathbb{R}^T, \mathbb{R}^Q)$. Even if the dimension of **x** can be large, most of the engineering problems we are focused on only contain $P \leq 50$ and $Q \leq 5$.
- **Computational budget**: A single computation of h can be very expensive. The computational budget $\lfloor$ will be represented by the number m of runs affordable to solve the problem.
- **Black box/white box**: h is either a black box (*the inner operations of the model are not accessible*), a grey box (*part of the inner operations is accessible*) or a white box (*all the operations of the model are accessible*).
- **Mathematical properties**: the basic mathematical properties (regularity, monotony, linearity or non linearity towards certain parameters) may be unknown to the engineer.
- **Domain of validity**: h should be delivered with its domain of validity $\mathcal{V}^{[\epsilon]} \subseteq \mathbb{R}^P \times \mathbb{R}^T$.

What kind of information do we manipulate?

What is a panoply $\mathcal{H}$ of models? $\mathcal{H} = \{h_1, \cdots, h_D\}, h_i \in \mathcal{F}(\mathcal{X}_i \times \Theta_i, \mathcal{Y})$

Example:

- **Model** h_1: Linear regression based on a database $\mathcal{D}_1$
- **Model** h_2: Neural network based on a database $\mathcal{D}_2$
- **Model** h_3: Linear PDE model based on a simplified geometry $\mathcal{G}_S$ and solved by numerical method $\mathcal{M}_1$
- **Model** h_4: Linear PDE model based on a simplified geometry $\mathcal{G}_S$ and solved by numerical method $\mathcal{M}_2$
- **Model** h_5: Linear PDE model based on a complex geometry $\mathcal{G}_C$ and solved by numerical method $\mathcal{M}_1$
- **Model** h_6: Non linear PDE model based on a simplified geometry $\mathcal{G}_S$ and solved by numerical method $\mathcal{M}_1$
- ...
- **Model** h_D: Non linear PDE model based on a complex geometry $\mathcal{G}_C$ and solved by numerical method $\mathcal{M}_3$

Outline

Notations

- **Variable of interest:** $\mathbf{Y}^*$ with values $\mathbf{y}^* \in \mathbb{R}^{\mathbf{Q}}$ and unknown statistical law $\mathbb{Q}$
- **Reference database:** $((\mathbf{X}_1^*, \mathbf{Y}_1^*), \cdots, (\mathbf{X}_n^*, \mathbf{Y}^*))$ or $(\mathbf{Y}_1^*, \cdots, \mathbf{Y}_n^*)$ when the $\mathbf{X}_i^*$'s are not observed
- **Model h:** $h \in \mathcal{H} = \{h_1, \cdots, h_D\}, \quad h : (\mathbf{x}, \theta) \in \mathcal{X} \times \mathbf{\Theta} \mapsto \mathbf{y} = h(\mathbf{x}, \theta) \in \mathcal{Y}$
- **Computational budget** $\mathcal{B}$: m simulations $(\mathbf{X}_k, h(\mathbf{X}_k, \theta))_{k=1,\cdots,m}$ with $\mathbf{X}_i$ iid following $\mathbb{P}_{\mathbf{X}}$.
- **Features of interest:** $(\rho_j(\mathbb{Q}))_{j\in\mathcal{J}}, \rho_j(\mathbb{Q}) \in \mathbb{F}_j$. Also abusively noted $\rho(\mathbf{Y}_j^*)$.

Definitions

Contrast

Definition: A contrast function is defined by:

$$\begin{aligned} \Psi : \quad \mathbb{F} \times \mathcal{Y} &\longrightarrow \mathbb{R} \\ (\rho, y) &\mapsto \Psi(\rho, y) \end{aligned}$$

Examples

- $\mathbb{F} = \mathbb{R}$:
  - **Mean-squared contrast**: $\Psi(\rho, y) = (y - \rho)^2$
- $\mathbb{F} = \{\text{Set of density function}\}$:
  - **Log-contrast**: $\Psi(\rho, y) = -\log(\rho(y))$
  - **L_2-contrast**: $\Psi(\rho, y) = \|\rho\|_2^2 - 2\rho(y)$

Definitions

Risk function

Definition: Given $(\Psi, \mathbb{F}, \mathbb{Q})$, the risk function $\mathcal{R}_\Psi$ is a real function defined as:

$$\forall \rho \in \mathbb{F}, \quad \mathcal{R}_\Psi(\rho) := \int_{\mathcal{Y}} \Psi(\rho, v)\, \mathbb{Q}(dv) = \mathbb{E}_{V\sim\mathbb{Q}}\left[\Psi(\rho, V)\right]$$

Application to our problem

- $\rho = \rho_h(\theta)$
- $\mathcal{R}_\Psi(h,\theta) = \mathbb{E}_{\mathbf{Y}^*\sim\mathbb{Q}}\left[\Psi(\rho_h(\theta), \mathbf{Y}^*)\right]$
- Some classical risk functions:
  - The **mean-squared** contrast gives a distance between means: $\mathcal{R}_\Psi(h,\theta) = (\mathbb{E}\left[\mathbf{Y}^*\right] - \rho_h(\theta))^2 + \mathrm{Var}\left[\mathbf{Y}^*\right]$
  - The **log-contrast** gives the Kullbach-Leiber divergence between pdfs: $R_\Psi(h,\theta) = KL(f_{\mathbf{Y}^*}, \rho_h(\theta)) - \mathbb{E}\left[\log(\mathbf{Y}^*)\right]$, where $KL(g_1, g_2) = \int \log(\frac{g_1}{g_2}(y))\, g_1(y)\, dy$

Pb 1: Mono feature estimation by a single model approach

Mathematical goal

Let $\mathbb{Q}$ be the unknown probability measure associated to the real random variable $\mathbf{Y}^*$ defined over $(\mathbb{R}^Q, \mathcal{B}(R^Q), \mathbb{Q})$. Our main goal is to predict one feature $\rho(\mathbb{Q})$ of the distribution $\mathbb{Q}$.

General description of the statistical problem

We want to develop robust estimation procedures of the feature ρ based upon the availability of a reference database $(\mathbf{Y}_1^*, \cdots, \mathbf{Y}_n^*)$, a numerical model $h(\mathbf{x}, \theta)$, with $\mathbf{X}$ following $\mathbb{P}_{\mathbf{X}}$ and a computational budget $\mathcal{B}$ that can be spent either m times all at once or in an adptative way.

Pb 1: Mono feature estimation by a single model approach

Examples of probabilistic measures of risk $\rho(\mathbf{Y}^*)$

$$
\begin{aligned}
\text{Mean:} \quad & \rho_\mu(\mathbf{Y}^*) = \mathbb{E}\left[\mathbf{Y}^*\right] && \in \mathbb{F} = \mathbb{R} \\
\text{Variance:} \quad & \rho_\sigma(\mathbf{Y}^*) = \mathrm{Var}\left[\mathbf{Y}^*\right] && \in \mathbb{F} = \mathbb{R}_+ \\
\text{Quantile:} \quad & \rho_q(\mathbf{Y}^*) = q_r(\mathbf{Y}^*) && \in \mathbb{F} = \mathbb{R}_+ \\
\text{Probability:} \quad & \rho_p(\mathbf{Y}^*) = \mathbb{P}\left(\mathbf{Y}^* \in \mathcal{D}_P\right) && \in \mathbb{F} = [0,1] \\
\text{CDF:} \quad & \rho_{cdf}(\mathbf{Y}^*) = \mathbb{P}\left(\mathbf{Y}^* \leq \mathbf{y}^*\right) && \in \mathbb{F} = \mathcal{F}_{cdf}(\mathbb{R}^Q, [0,1]) \\
\text{PDF:} \quad & \rho_{pdf}(\mathbf{Y}^*) = f_{\mathbf{Y}^*}(\mathbf{y}^*) && \in \mathbb{F} = \mathcal{F}_{pdf}(\mathbb{R}^Q, \mathbb{R}_+)
\end{aligned}
$$

Pb1: Example of density prediction

Suppose that $(\mathbf{X}_1^*, \mathbf{Y}_1^*), ..., (\mathbf{X}_n^*, \mathbf{Y}_n^*)$ are available.

- **Calibration of θ by mean-Squares minimization**

$$\widehat{\theta}_{MS} = \underset{\theta \in \Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i^* - h(\mathbf{X}_i^*, \theta))^2$$

Pb1: Example of density prediction

Suppose that $(\mathbf{X}_1^*, \mathbf{Y}_1^*), ..., (\mathbf{X}_n^*, \mathbf{Y}_n^*)$ are available.

- **Calibration of θ by mean-Squares minimization**

$$\widehat{\theta}_{MS} = \underset{\theta \in \Theta}{\text{Argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i^* - h(\mathbf{X}_i^*, \theta))^2$$

- **Prediction of ρ**
  Compute the probability density of $h(\mathbf{X}, \widehat{\theta}_{MS})$ under $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}$

$$\boxed{\rightarrow \widehat{f}_{MS}}$$

EADS

Pb1: Example of density prediction

Other M-estimators...

- **Kullback-Leibler minimization** $KL(f_1, f_2) = \int_{\mathcal{Y}} \log(\frac{f_1}{f_2})\, f_1$
  - f = density of $\mathbf{Y}^*$, f_θ = density of $h(\mathbf{X}, \theta)$
  - **Goal**: Find θ that minimizes $KL(f, f_\theta)$.
- **Two difficulties**
  - f is unknown $\rightarrow$ replaced by $f^n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Y_i}$
  - f_θ untractable $\rightarrow$ replaced by a **simulation density** (Kernel, projection, etc...) $\left(f_\theta^m = \frac{1}{m}\sum_{j=1}^{m} K_{b_m}(\cdot - h(\mathbf{X}_j, \theta)), \quad \mathbf{X}_j \underset{i.i.d}{\sim} \mathbb{P}_\mathbf{X} \right)$
- **M-estimator**

$$\widehat{\theta}_{KL} = \underset{\theta\in\Theta}{\text{Argmin}}\, KL(f^n, f_\theta^m) = \underset{\theta\in\Theta}{\text{Argmin}} -\frac{1}{n}\sum_{i=1}^{n} \log(f_\theta^m)(\mathbf{Y}_i^*)$$

- **Prediction**

  Compute the probability density of $h(\mathbf{X}, \widehat{\theta}_{KL})$ under $\mathbf{X} \sim \mathbb{P}_\mathbf{X}$

$$\boxed{\rightarrow \widehat{f}_{KL}}$$

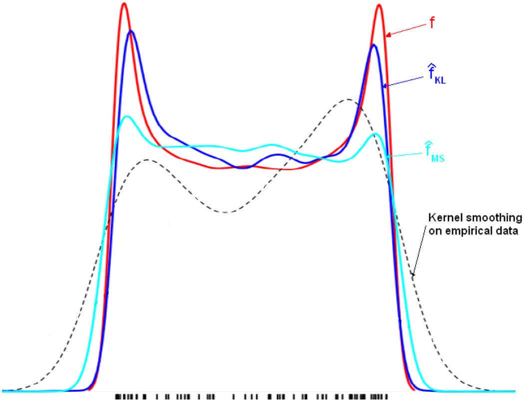
Question ?

What is the "best" estimator of f ,

$\widehat{f}_{MS}$ or $\widehat{f}_{KL}$?

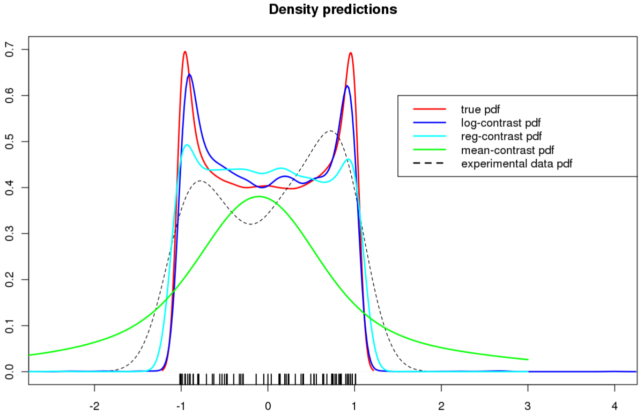
Pb 1: Toy application

- $\mathbf{Y}^* = \sin(\mathbf{X}^*) + 0.01\,\varepsilon, \quad \mathbf{X}^* \perp \varepsilon \sim \mathcal{N}(0,1)$
- $h(\mathbf{X},\theta) = \theta_1 + \theta_2\,X + \theta_3\,X^3,\ X \sim \mathbb{P}^{\mathbf{x}} = \mathcal{N}(0,1)$
- $n = 50$ and $m = 10^3$



Pb 1: Toy application

- $\mathbf{Y}^* = \sin(\mathbf{X}^*) + 0.01\,\varepsilon, \quad \mathbf{X}^* \perp \varepsilon \sim \mathcal{N}(0,1)$
- $h(\mathbf{X}, \theta) = \theta_1 + \theta_2\, X + \theta_3\, X^3,\ X \sim \mathbb{P}^{\mathsf{x}} = \mathcal{N}(0,1)$
- $n = 50$ and $m = 10^4$



Pb1: Theoretical results from N. Rachdi PhD Thesis

Theorem: Oracle Inequality (N. Rachdi *et al* 2010)

Under some conditions on the contrast Ψ and under tightness conditions, for all $\varepsilon > 0$, with high probability it holds:

$$0 \leq \mathcal{R}_\Psi(h, \widehat{\theta}) - \inf_{\theta \in \Theta} \left(\mathcal{R}_\Psi(h, \theta)\right) \leq \frac{K^{\epsilon}_{(\widetilde{\rho}, \Psi)}}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} \left(K^{\epsilon}_{(\widetilde{\rho}, h)} + B_m\right)\right)$$

where $K^{\epsilon}_{(\widetilde{\rho}, \Psi)}$, $K^{\epsilon}_{(\widetilde{\rho}, h)}$ some *concentration constants* and B_m a bias factor

Pb1: Theoretical results from N. Rachdi PhD Thesis

Theorem: Oracle Inequality (N. Rachdi *et al* 2010)

Under some conditions on the contrast Ψ and under tightness conditions, for all $\varepsilon > 0$, with high probability it holds:

$$0 \leq \mathcal{R}_\Psi(h, \widehat{\theta}) - \inf_{\theta \in \Theta} \left(\mathcal{R}_\Psi(h, \theta)\right) \leq \frac{K^{\epsilon}_{(\widetilde{\rho}, \Psi)}}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}} \left(K^{\epsilon}_{(\widetilde{\rho}, h)} + B_m\right)\right)$$

where $K^{\epsilon}_{(\widetilde{\rho}, \Psi)}$, $K^{\epsilon}_{(\widetilde{\rho}, h)}$ some *concentration constants* and B_m a bias factor

- Nonasymptotic result, i.e valid for all $n, m \geq 1$
- $\inf_{\theta \in \Theta} \left(\mathcal{R}_\Psi(h, \theta)\right) =$ the minimal *risk* we can achieve for Ψ
  $=$ Modeling error (mesh size ..., model complexity)
- $\frac{K^{\epsilon}_{(\widetilde{\rho}, \Psi)}}{\sqrt{n}} \left(1 + \sqrt{\frac{n}{m}}(K^{\epsilon}_{(\widetilde{\rho}, h)} + B_m)\right) =$ Statistical error linked to model complexity and size of the databases

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\, b,\, c_m$ ($c_m \underset{m}{\rightarrow} 0$) such that

- **Step 2:** two empirical processes suprema

- **Step 3:** union bound + tightness

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\, b,\, c_m$ ($c_m \underset{m}{\to} 0$) such that

$$\mathcal{R}_\Psi(\widehat{\theta}_\Psi) \leq \inf_{\theta\in\Theta} \left(\mathcal{R}_\Psi(\theta)\right) + \frac{a}{\sqrt{n}}\, ||\mathbb{G}_n||_{\mathcal{W}_{(\kappa,\Psi)}} + \frac{b}{\sqrt{m}}\, \|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\kappa,\boldsymbol{h})}} + c_m\,.$$

- **Step 2:** two empirical processes suprema
- **Step 3:** union bound + tightness

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\, b,\, c_m$ ($c_m \underset{m}{\rightarrow} 0$) such that

$$\mathcal{R}_\Psi(\widehat{\theta}_\Psi) \leq \inf_{\theta \in \Theta} \left(\mathcal{R}_\Psi(\theta)\right) + \frac{a}{\sqrt{n}} \|\mathbb{G}_{\boldsymbol{n}}\|_{\mathcal{W}_{(\kappa,\Psi)}} + \frac{b}{\sqrt{m}} \|\mathbb{K}^{\mathbf{x}}_{\boldsymbol{m}}\|_{\mathcal{P}_{(\kappa,\boldsymbol{h})}} + c_m \,.$$

- **Step 2:** two empirical processes suprema
- **Step 3:** union bound + tightness

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\, b,\, c_m$ ($c_m \underset{m}{\rightarrow} 0$) such that

$$\mathcal{R}_\Psi(\widehat{\theta}_\Psi) \leq \inf_{\theta\in\Theta} \left(\mathcal{R}_\Psi(\theta)\right) + \frac{a}{\sqrt{n}}\, ||\mathbb{G}_n||_{\mathcal{W}_{(\kappa,\Psi)}} + \frac{b}{\sqrt{m}}\, \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\kappa,\boldsymbol{h})}} + c_m\,.$$

- **Step 2:** two empirical processes suprema
- **Step 3:** union bound + tightness

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\ b,\ c_m\ (c_m \underset{m}{\to} 0)$ such that

$$\mathcal{R}_\Psi(\widehat{\theta}_\Psi) \le \inf_{\theta\in\Theta} \left(\mathcal{R}_\Psi(\theta)\right) + \frac{a}{\sqrt{n}}\, ||\mathbb{G}_n||_{\mathcal{W}_{(\kappa,\Psi)}} + \frac{b}{\sqrt{m}}\, \|\mathbb{K}^{\mathbf{x}}_m\|_{\mathcal{P}_{(\kappa,\boldsymbol{h})}} + c_m\,.$$

- **Step 2:** two empirical processes suprema
  - $\mathbb{G}_n(\cdot) := \sqrt{n}(Q_n - Q)(\cdot)$; $\mathcal{W}_{(\kappa,\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\kappa(\lambda)\,,\, y)\,,\, \lambda \in \mathcal{Y}\}$
  - $\mathbb{K}^{\mathbf{x}}_m(\cdot) = \sqrt{m}(P^{\mathbf{x}}_m - P^{\mathbf{x}})(\cdot)$; $\mathcal{P}_{(\kappa,\boldsymbol{h})} = \{\mathbf{x} \in \mathcal{X} \mapsto \kappa(h(\mathbf{x},\theta))(\lambda)\,,\, (\theta,\lambda) \in \Theta \times \mathcal{Y}\}$
- **Step 3:** union bound + tightness

EADS

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\, b,\, c_m\ (c_m \underset{m}{\to} 0)$ such that

$$\mathcal{R}_\Psi(\widehat{\theta}_\Psi) \leq \inf_{\theta\in\Theta} \left(\mathcal{R}_\Psi(\theta)\right) + \frac{a}{\sqrt{n}} \|\mathbb{G}_n\|_{\mathcal{W}_{(\kappa,\Psi)}} + \frac{b}{\sqrt{m}} \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\kappa,\boldsymbol{h})}} + c_m\,.$$

- **Step 2:** two empirical processes suprema
  - $\mathbb{G}_n(\cdot) := \sqrt{n}(Q_n - Q)(\cdot);\ \mathcal{W}_{(\kappa,\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\kappa(\lambda)\,,\, y)\,,\, \lambda \in \mathcal{Y}\}$
  - $\mathbb{K}_m^{\mathbf{x}}(\cdot) = \sqrt{m}(P_m^{\mathbf{x}} - P^{\mathbf{x}})(\cdot);\ \mathcal{P}_{(\kappa,\boldsymbol{h})} = \{\mathbf{x} \in \mathcal{X} \mapsto \kappa(h(\mathbf{x},\theta))(\lambda)\,,\, (\theta,\lambda) \in \Theta \times \mathcal{Y}\}$
- **Step 3:** union bound + tightness
  - tightness $\dashrightarrow$ "complexity" of classes of functions $\mathcal{W}_{(\kappa,\Psi)}$, $\mathcal{P}_{(\kappa,h)}$

Proof Ingredients

to identify empirical processes

- **Step 1:** def. $\widehat{\theta}_\Psi$ + def. θ_Ψ + assumptions
  we prove that $\exists\, a,\, b,\, c_m$ $(c_m \underset{m}{\to} 0)$ such that

$$\mathcal{R}_\Psi(\widehat{\theta}_\Psi) \leq \inf_{\theta\in\Theta} \left(\mathcal{R}_\Psi(\theta)\right) + \frac{a}{\sqrt{n}}\, \|\mathbb{G}_n\|_{\mathcal{W}_{(\kappa,\Psi)}} + \frac{b}{\sqrt{m}}\, \|\mathbb{K}_m^{\mathbf{x}}\|_{\mathcal{P}_{(\kappa,\boldsymbol{h})}} + c_m\,.$$

- **Step 2:** two empirical processes suprema
  - $\mathbb{G}_n(\cdot) := \sqrt{n}(Q_n - Q)(\cdot)$; $\mathcal{W}_{(\kappa,\Psi)} = \{y \in \mathcal{Y} \mapsto \Psi(\kappa(\lambda)\,,\, y)\,,\, \lambda \in \mathcal{Y}\}$
  - $\mathbb{K}_m^{\mathbf{x}}(\cdot) = \sqrt{m}(P_m^{\mathbf{x}} - P^{\mathbf{x}})(\cdot)$; $\mathcal{P}_{(\kappa,\boldsymbol{h})} = \{\mathbf{x} \in \mathcal{X} \mapsto \kappa(h(\mathbf{x},\theta))(\lambda)\,,\, (\theta,\lambda) \in \Theta \times \mathcal{Y}\}$
- **Step 3:** union bound + tightness
  - tightness $\dashrightarrow$ Bracketing entropy of the classes $\mathcal{W}_{(\kappa,\Psi)}$, $\mathcal{P}_{(\kappa,h)}$

EADS

Pb1: Theoretical results from N. Rachdi PhD Thesis

- **Compare** $\mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi^P})$ **and** $\mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi})$

  study the difference $\mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi^P}) - \mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi})$

- **By definition of** θ_{Ψ^P}**:** $\mathcal{R}_{\Psi^P}(\theta_{\Psi^P}) - \mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi}) \leq 0$ for all $\widehat{\theta}_{\Psi}$
- **Question :** $\mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi^P}) - \mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi}) \leq$ 0? *a.s? w.h.p?, in* L_1*? $\cdots$ difficult in general*?

Proposition: [Mean squares for mean prediction] (N. Rachdi, JC. Fort 2010)

- **Feature of interest:** $\rho^p = \mathbb{E}(Y) \dashrightarrow \Psi^p : (\rho, y) \mapsto (\rho - y)^2$
- **Model:** $h(\mathbf{X}, \theta) = \Phi(\mathbf{X}) \cdot \theta$, $\Phi = (\phi_1, ..., \phi_k)$ orho. w.r.t $P_{\mathbf{X}}$
- **Suppose:** $Y_i = \Phi(\mathbf{X}_i) \cdot \theta^* + \varepsilon_i$, $\mathbb{E}(\varepsilon_i) = 0$ i.i.d
- **Let 2 Ψ-estimators:** $\widehat{\theta}_{\Psi^P} = \text{Argmin}_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - \mathbb{E}\Phi(\mathbf{X}) \cdot \theta)^2$ and $\widehat{\theta}_{\Psi_{reg}} = \text{Argmin}_{\theta \in \Theta} \sum_{i=1}^{n} (Y_i - \Phi(\mathbf{X}_i) \cdot \theta)^2$
- **Result:**

$$\mathbb{E}_{(\mathbf{X}_i, Y_i)_{1..n}} \left(\mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi^P}) - \mathcal{R}_{\Psi^P}(\widehat{\theta}_{\Psi}) \right) \leq 0$$

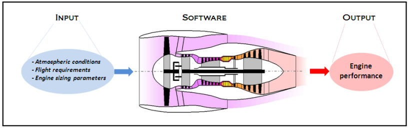
Inverse Problems Applications

N. Rachdi, JC. Fort & T. Klein *Stochastic Inverse Problem with Noisy Simulator* (2011) submitted

- **Fuel Mass data:**

Reference Fuel Masses [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

- **Model (noisy simulator):**



- **Goal:** Identify SFC (Specific Fuel Consumption)

Pb 2: Mono feature estimation by a panoply of models

Mathematical goal

Let $\mathbb{Q}$ be the unknown probability measure associated to the real random variable $\mathbf{Y}^*$ defined over $(\mathbb{R}^Q, \mathcal{B}(R^Q), \mathbb{Q})$. Our main goal is to predict one feature $\rho(\mathbb{Q})$ of the distribution $\mathbb{Q}$.

General description of the statistical problem

We want to develop robust estimation procedures of a feature ρ based upon the availability of a reference database $(\mathbf{Y}_1^*, \cdots, \mathbf{Y}_n^*)$, a panoply of numerical models $\mathcal{H} = \{h_1, \cdots, h_D\}$, with $h_i \in \mathcal{F}(\mathcal{X}_i \times \Theta_i, \mathcal{Y})$ and $\mathbf{X}_i$ following $\mathbb{P}_{\mathbf{X}_i}$ and a computational budget $\mathcal{B}$. $\mathcal{B}$ can be split into D computational budgets $\mathcal{B}_i$, each one corresponding to m_i simulations of the model h_i either all at once or in an adptative way.

Pb 3: Multi feature estimation by a single model

Mathematical goal

Let $\mathbb{Q}$ be the unknown probability measure associated to the real random variable $\mathbf{Y}^*$ defined over $(\mathbb{R}^Q, \mathcal{B}(R^Q), \mathbb{Q})$. Our main goal is to predict several feature $(\rho_j(\mathbb{Q}))_{j\in\mathcal{J}}$ of the distribution $\mathbb{Q}$.

General description of the statistical problem

We want to develop robust estimation procedures of several features $(\rho_j(\mathbb{Q}))_{j\in\mathcal{J}})$ based upon the availability of a reference database $(\mathbf{Y}_1^*, \cdots, \mathbf{Y}_n^*)$, a numerical model $h(\mathbf{x}, \theta)$, with $\mathbf{X}$ following $\mathbb{P}_{\mathbf{X}}$ and a computational budget $\mathcal{B}$ that can be spent either m times all at once or in an adptative way.

Pb 4: Multi feature estimation by a panoply of models

Mathematical goal

Let $\mathbb{Q}$ be the unknown probability measure associated to the real random variable $\mathbf{Y}^*$ defined over $(\mathbb{R}^Q, \mathcal{B}(R^Q), \mathbb{Q})$. Our main goal is to predict several feature $(\rho_j(\mathbb{Q}))_{j\in\mathcal{J}}$ of the distribution $\mathbb{Q}$.

General description of the statistical problem

We want to develop robust estimation procedures of several features $(\rho_j(\mathbb{Q}))_{j\in\mathcal{J}})$ based upon the availability of a reference database $(\mathbf{Y}_1^*, \cdots, \mathbf{Y}_n^*)$, a panoply of numerical models $\mathcal{H} = \{h_1, \cdots, h_D\}$, with $h_i \in \mathcal{F}(\mathcal{X}_i \times \Theta_i, \mathcal{Y})$ and $\mathbf{X}_i$ following $\mathbb{P}_{\mathbf{X}_i}$ and a computational budget $\mathcal{B}$. $\mathcal{B}$ can be split into D computational budgets $\mathcal{B}_i$, each one corresponding to m_i simulations of the model h_i either all at once or in an adptative way.

Outline

Conclusion

- **Ψ-estimators $\widehat{\theta}_{\Psi}$**
  - Constants improvement in risk bound inequalities
  - Central Limit Theorems
- **Duality estimation-prediction**
  - Rigorous analysis, functional study of contrast functions
  - More academic results
- **Extension to model selection**
  - For a given purpose (quantile study, threshold prob. etc...), what model to choose ?
  - Formalize the notion of "model granularity"
  - $\neq$ classical model selection $\rightarrow$ we know the "best" model... but too expensive

Outline

- A. Van der Vaart, J. Wellner, Weak convergence and empirical processes, 1996
- P. Massart, Concentration Inequalities and Model Selection, Ecole d'Ã©tÃ© de Saint-Flour, 2003
- N. Rachdi, PhD thesis, University of Toulouse, 2011
- F. Bach, formation EADS
- S. Arlot, PhD thesis, University of Paris-Sud, 2007