Leçon 1 - MAP 568

Josselin Garnier

2017-2018

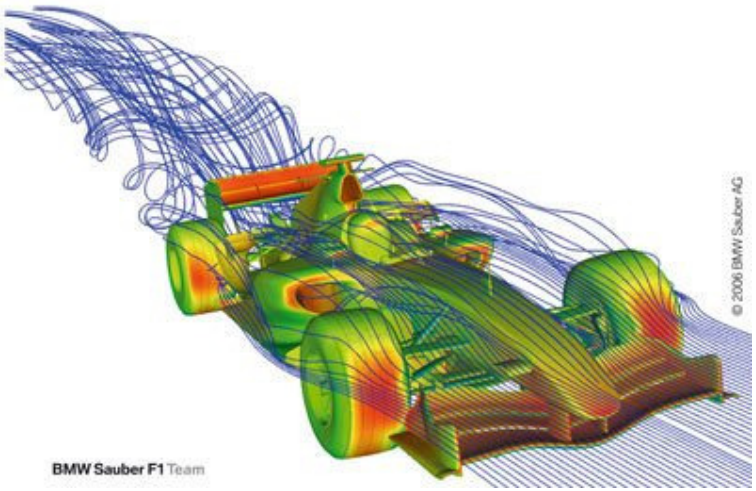
Préambule

- Cours 9 blocs (cours + PC).
- Contrôle : 22 mars.
- Projet de simulation
  - par binôme
  - distribué au cours 4
  - rapport+code à remettre le 22 mars.

As we know, there are known knowns. There are things we know we know. We also know there are known unknowns. That is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know.
Donald Rumsfeld (Secrétaire à la Défense des Etats-Unis, durant un briefing avec la presse, 12 février 2002)

Gestion des incertitudes et analyse de risque

Cadre : utilisation de simulations numériques pour la modélisation.



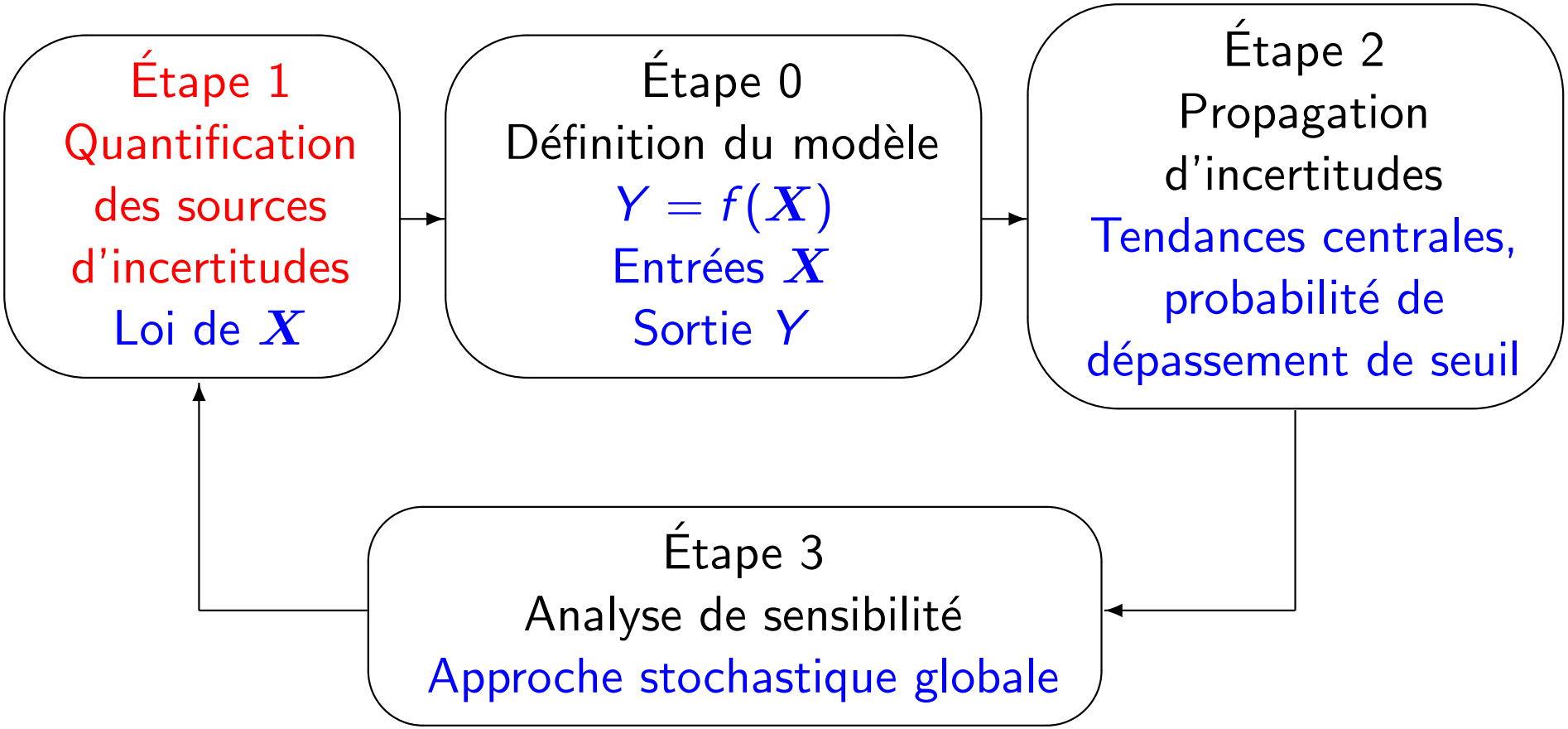
Elles permettent d'éviter le coût d'expériences réelles, pour :

- la conception d'un avant-projet,
- l'optimisation du projet final,
- la validation du projet abouti.

Gestion des incertitudes et analyse de risque

- Sources d'incertitudes : paramètres physiques, conditions environnementales, erreurs de fabrication, etc.
- Gestion des incertitudes : quantifier la confiance en les prédictions et les décisions issues de telles simulations.
- Objectif du cours : Présenter des méthodes mathématiques permettant de modéliser, de caractériser et d'analyser les incertitudes dans des simulations numériques.

Gestion des incertitudes et analyse de risque

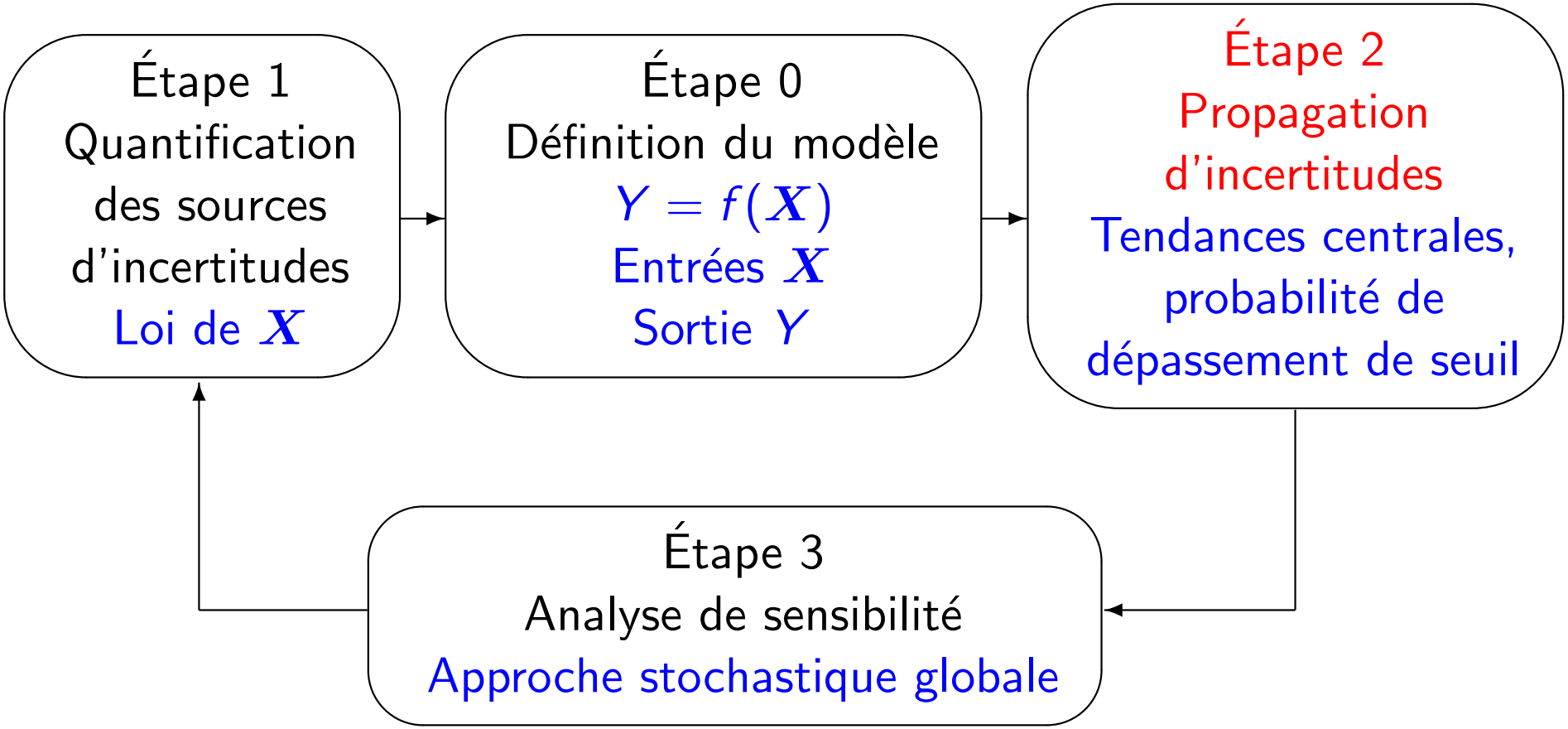


Partie I. Les sources d'incertitudes

• Deux types d'entrée "incertaines" :
- variables stochastiques (pour les codes et les expériences) : ces variables ont une variabilité naturelle résultant de phénomènes aléatoires (typiquement, une quantité associée à une pièce technologique, soumises aux aléas d'un procédé de fabrication).
- variables épistémiques (pour les codes surtout) : ces variables possèdent une valeur mais elle nous est inconnue, à cause d'un manque de connaissance (typiquement, une constante d'une loi physique).

• Les variables d'entrées sont traitées comme des variables aléatoires, de lois de probabilités données.

• Les variables d'entrées sont traitées comme des variables aléatoires, de lois de probabilités données.
Pour déterminer la (densité de) la loi d'une variable aléatoire, on prend en compte des jugements d'expert, des informations a priori, et/ou des données (un échantillon).
Méthodes (cours 1) :
- méthodes non-paramétriques à noyaux,
- méthodes paramétriques d'ajustement à une loi analytique,
- méthodes entropiques,
- méthodes bayésiennes.


Étape 1
Quantification
des sources
d'incertitudes
Loi de $\boldsymbol{X}$
Étape 0
Définition du modèle
$Y = f(\boldsymbol{X})$
Entrées $\boldsymbol{X}$
Sortie Y
Étape 2
Propagation
d'incertitudes
Tendances centrales,
probabilité de
dépassement de seuil
Étape 3
Analyse de sensibilité
Approche stochastique globale

Partie II. Propagation d'incertitudes

- Contexte : code de calcul ou expérience modélisé par

$$Y = f(\boldsymbol{X})$$

avec Y =variable de sortie
$\boldsymbol{X} = (X_i)_{i=1,\dots,d}$ variables d'entrée, de loi donnée
f = boîte noire

- But : estimation d'une quantité

$$\mathbb{E}[\psi(Y)]$$

avec une barre d'erreur et le minimum de simulations/expériences.

- Exemples (pour Y à valeurs réelles) :
- $\psi(y) = y \rightarrow$ moyenne de Y, i.e. $\mathbb{E}[Y]$
- $\psi(y) = y^2 \rightarrow$ variance de Y, i.e. $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$
- $\psi(y) = \mathbf{1}_{[a,\infty[}(y) \rightarrow$ probabilité de dépasser le seuil a, i.e. $\mathbb{P}(Y \geq a)$

Cumul quadratique

- But : Evaluer $\mathbb{E}[Y]$ et $\mathrm{Var}(Y)$ en connaissant le vecteur moyenne $\boldsymbol{\mu} = (\mathbb{E}[X_i])_{i=1}^d$ et la matrice de covariance $\mathbf{C} = (\mathrm{Cov}(X_i, X_j))_{i,j=1}^d$ des variables d'entrée.
- En supposant les C_{ii} petits et f régulier :

$$\mathbb{E}[Y] \simeq f(\boldsymbol{\mu}), \qquad \mathrm{Var}(Y) \simeq \nabla f(\boldsymbol{\mu})^T \mathbf{C} \nabla f(\boldsymbol{\mu}).$$

Preuve : Si $f \in \mathcal{C}^2$, alors $f(\boldsymbol{x}) = f(\boldsymbol{\mu}) + \nabla f(\boldsymbol{\mu}) \cdot (\boldsymbol{x} - \boldsymbol{\mu}) + O(|\boldsymbol{x} - \boldsymbol{\mu}|^2)$.

- On juste besoin de connaître le gradient de f en $\boldsymbol{\mu}$ (par différences finies ou par calcul adjoint).
- Rapide, analytique, permet de calculer approximativement des tendances centrales de la sortie (moyenne, variance).
- Convenable pour des petites variations des paramètres d'entrée et un modèle régulier (qu'on peut linéariser).
- Approche "locale". En général, pas de contrôle de l'erreur.

Méthodes de quadrature

- La quantité à estimer est une intégrale d-dimensionnelle :

$$I = \mathbb{E}[\psi(Y)] = \mathbb{E}[\psi(f(\boldsymbol{X}))] = \int_{\mathbb{R}^d} \psi(f(\boldsymbol{x}))p(\boldsymbol{x})d\boldsymbol{x}$$

où $p(\boldsymbol{x})$ est la densité de probabilité de $\boldsymbol{X}$.

- Les méthodes de quadrature (cours 2) réclament :

- des conditions de régularité sur $\boldsymbol{x} \to \psi(f(\boldsymbol{x}))$,
- une dimension d petite (même avec des grilles creuses de type Smolyak),
- beaucoup d'appels au code.

Méthode de Monte Carlo

- Estimateur de $I = \mathbb{E}[\psi(Y)]$, $Y = f(\boldsymbol{X})$:

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} \psi(f(\boldsymbol{X}^{(k)}))$$

où les $(\boldsymbol{X}^{(k)})_{k=1,\dots,n}$ sont n réalisations indépendantes de même loi que $\boldsymbol{X}$.

- Estimation non-biaisée :

$$\mathbb{E}[\hat{I}_n] = I$$

avec une erreur :

$$\mathbb{E}\left[(\hat{I}_n - I)^2\right]^{1/2} = \frac{1}{\sqrt{n}} \mathrm{Var}\left(\psi(f(\boldsymbol{X}))\right)^{1/2}$$

Possibilité d'obtenir des intervalles de confiance.

- Avantages :

- pas de régularité requise sur f, ψ.
- vitesse de convergence indépendante de la dimension.

Méthodes de Monte Carlo avancées

$$\mathbb{E}\left[(\hat{I}_n - I)^2\right]^{1/2} = \frac{1}{\sqrt{n}} \mathrm{Var}\left(\psi(f(\boldsymbol{X}))\right)^{1/2}$$

- **Techniques de réduction de variance**

On cherche à réduire la "constante" tout en gardant le $1/\sqrt{n}$.
Différentes méthodes possibles.
Particulièrement utile pour l'estimation de probabilités d'événements rares.

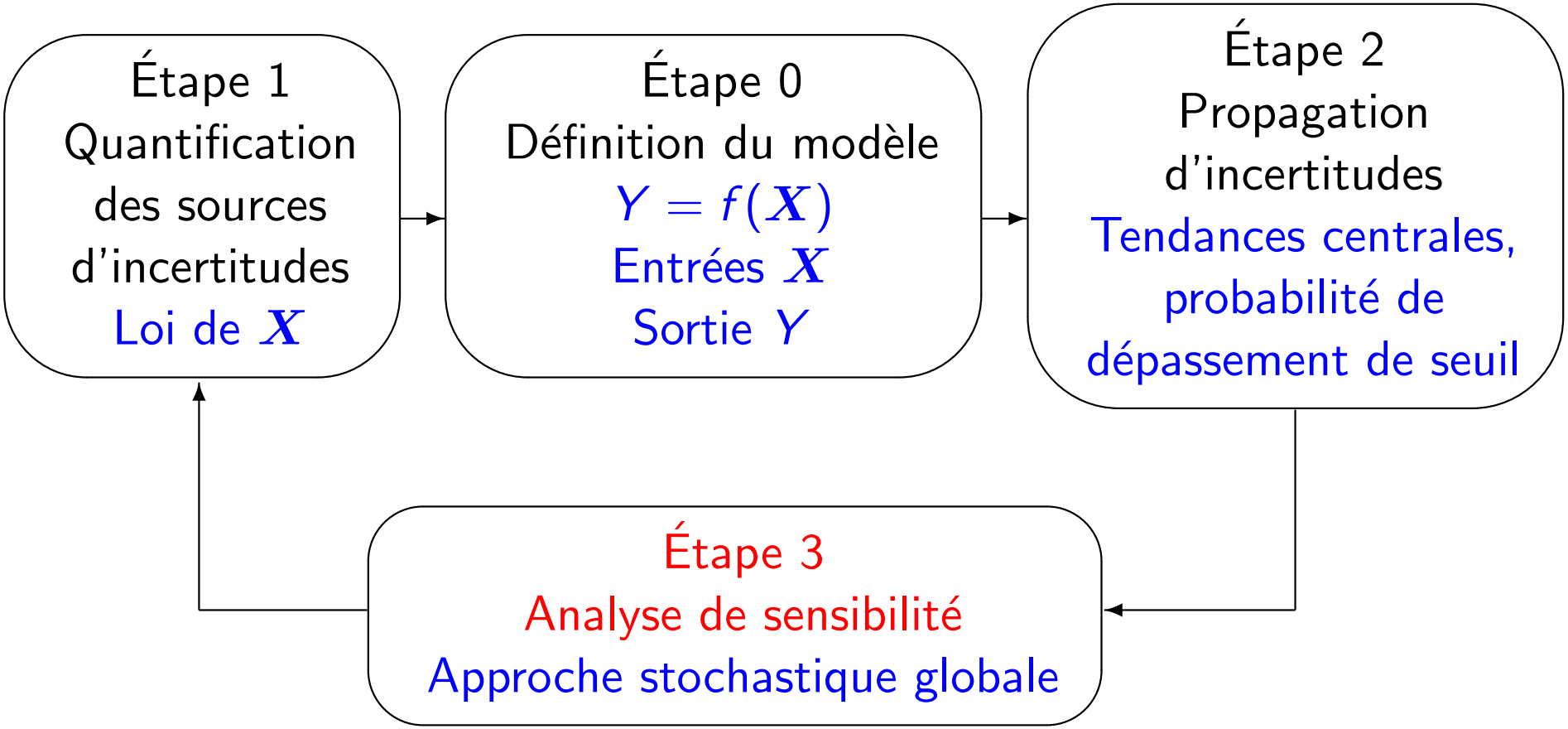
- **Quasi Monte Carlo**

On tire l'échantillon de manière moins aléatoire que Monte Carlo (pour combler les trous essentiellement).
Cette technique :
- réduit la variance si $\boldsymbol{x} \to \psi(f(\boldsymbol{x}))$ a un peu de régularité et/ou de monotonie; on peut aller jusqu'à une erreur en $C_d(\log n)^d/n$,
- marche en dimension pas trop grande,
- représente un intermédiaire entre Monte Carlo et quadrature usuelle,
- ne donne pas d'intervalle de confiance.

Propagation d'incertitudes par métamodèles

On remplace f par un métamodèle (modèle réduit, surface de réponse) f_r et on applique une des techniques précédentes (quadrature, Monte Carlo).
$\rightarrow$ On peut faire beaucoup d'appels au métamodèle.
$\rightarrow$ Le choix du métamodèle est critique.
$\rightarrow$ La méthode par cumul quadratique est un cas particulier (métamodèle affine).

- Le métamodèle doit être construit en appelant le code f en un minimum de points.
- Le contrôle de l'erreur n'est pas simple. On verra des utilisations du métamodèle qui ne nécessitent pas un contrôle fin.


Étape 1
Quantification
des sources
d'incertitudes
Loi de $\boldsymbol{X}$
Étape 0
Définition du modèle
$Y = f(\boldsymbol{X})$
Entrées $\boldsymbol{X}$
Sortie Y
Étape 2
Propagation
d'incertitudes
Tendances centrales,
probabilité de
dépassement de seuil
Étape 3
Analyse de sensibilité
Approche stochastique globale

Partie III. Analyse de sensibilité

- Contexte : code de calcul ou expérience modélisé par

$$Y = f(\boldsymbol{X})$$

avec Y =variable de sortie réelle
$\boldsymbol{X} = (X_1, ..., X_d)$ variables d'entrée, de loi donnée
f = boîte noire

- But : expliquer la variabilité de la réponse Y en fonction des X_i.
- Objectifs principaux :

- améliorer la compréhension du phénomène.
- réduire l'incertitude d'un modèle, en identifiant les variables les plus influentes ($\rightarrow$ il devient prioritaire de réduire la variabilité de ces entrées).
- simplifier ou alléger le modèle, en fixant les variables les moins influentes.

• But de l'analyse de sensibilité :
- qualitative : identifier les paramètres importants parmi les nombreux paramètres (screening ou criblage).
- quantitative : déterminer la part de la variance de la sortie Y due à une variable d'entrée ou un sous-ensemble de variables d'entrée X_i.

Analyse qualitative : Criblage

• Un modèle comportant beaucoup de variables d'entrée est difficile à explorer.
Souvent, seulement quelques entrées sont influentes.
Objectif : identifier rapidement les quelques h entrées influentes parmi les d entrées, avec n calculs.

• Méthodes possibles :

- méthode OAT (One factor At a Time) : calcul de gradient avec $n = d + 1$ (ou $n = 1$ avec méthode adjointe).
- méthode de Morris (généralise OAT) avec $n = R(d + 1)$ (ou $n = R$ avec méthode adjointe).
- matrices super-saturées avec $n \leq d$ et $h \ll d$.
- criblage par groupes, bifurcation séquentielle avec $h \ll n \ll d$.

Analyse quantitative : Indices de sensibilité

- Calcul d'indices représentant la contribution relative due à une variable ou à un groupe de variables d'entrée sur la variance de la réponse Y.
- Exemple :

$$S_i = \frac{\mathrm{Var}(\mathbb{E}[Y|X_i])}{\mathrm{Var}(Y)}.$$

Par le théorème de la variance totale :

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|X_i]) + \mathbb{E}[\mathrm{Var}(Y|X_i)],$$

on voit que $S_i \in [0,1]$.
Intuitivement, S_i est la part de la variance expliquée par X_i :
- Si S_i est proche de 0, alors $\mathrm{Var}(Y|X_i)$ est (en moyenne) proche de $\mathrm{Var}(Y)$: connaître X_i ne change pas l'amplitude des fluctuations de Y, donc Y dépend peu de X_i.
- Si S_i est proche de 1, alors $\mathrm{Var}(Y|X_i)$ est beaucoup plus petit que $\mathrm{Var}(Y)$: connaître X_i diminue fortement l'amplitude des fluctuations de Y, donc Y dépend fortement de X_i.

Quelques outils fondamentaux

Outils fondamentaux : théorème de Bayes

• Formule de Bayes.

Pour tous événements A et B (avec $\mathbb{P}(B) > 0$ et $\mathbb{P}(A) > 0$) :

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B)\mathbb{P}(A|B)}{\mathbb{P}(A)}$$

• Théorème de Bayes.

Si la loi de $\boldsymbol{X}$ conditionnellement aux paramètres $\boldsymbol{\beta}$ a pour densité $p(\boldsymbol{x}|\boldsymbol{\beta})$,
si la loi a priori sur les paramètres $\boldsymbol{\beta}$ a pour densité $p_{\text{prior}}(\boldsymbol{\beta})$,
alors la loi a posteriori des paramètres $\boldsymbol{\beta}$ étant données les observations $\boldsymbol{X} = \boldsymbol{x}$ a pour densité

$$p_{\text{post}}(\boldsymbol{\beta}|\boldsymbol{x}) \approx p_{\text{prior}}(\boldsymbol{\beta})p(\boldsymbol{x}|\boldsymbol{\beta})$$

où l'égalité $\approx$ est à une constante multiplicative près (dépendant de $\boldsymbol{x}$).
L'inverse de cette constante vaut : $\int p_{\text{prior}}(\boldsymbol{\beta}')p(\boldsymbol{x}|\boldsymbol{\beta}')d\boldsymbol{\beta}'$.

Espérance conditionnelle

Soit $(\boldsymbol{X}, Y)$ un vecteur aléatoire, $\boldsymbol{X}$ à valeurs dans $\mathbb{R}^n$ et Y à valeurs dans $\mathbb{R}$.

• L'espérance de Y est la constante qui approche au mieux la variable aléatoire Y :

$$\mathbb{E}[Y] = \operatorname*{argmin}_{a \in \mathbb{R}} \mathbb{E}\left[(Y-a)^2\right]$$

(au sens des moindres carrés).

- Espérance conditionnelle.

On observe $\boldsymbol{X}$. Quelle est la meilleure estimation de Y sachant $\boldsymbol{X}$?
On cherche donc :

$$\psi_0 = \underset{\psi:\mathbb{R}^n \to \mathbb{R}}{\operatorname{argmin}} \ \mathbb{E}\big[(Y - \psi(\boldsymbol{X}))^2\big]$$

Problème de minimisation complexe.

- On a :

$$\psi_0(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}].$$

Remarque 1 : si $\boldsymbol{X}$ et Y sont indépendants, alors $\mathbb{E}[Y|\boldsymbol{X}] = \mathbb{E}[Y]$.
Remarque 2 : si $(\boldsymbol{X}, Y)$ est un vecteur gaussien alors $\mathbb{E}[Y|\boldsymbol{X}]$ est une fonction affine en $\boldsymbol{X}$.
Remarque 3 : $\mathbb{E}[Y|\boldsymbol{X}]$ est la projection orthogonale (dans $L^2(\Omega)$) de Y sur le sous-espace infini-dimensionnel des variables aléatoires de L^2 de la forme $\psi(\boldsymbol{X})$.

Preuve de : le ψ optimal est $\psi_0(\boldsymbol{x}) = \mathbb{E}\big(Y|\boldsymbol{X}=\boldsymbol{x}\big) = \int_{\mathbb{R}} y p_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy$.
En notant $\tilde{\psi}(\boldsymbol{x}) = \psi(\boldsymbol{x}) - \psi_0(\boldsymbol{x})$, on a

$$\begin{aligned}\mathbb{E}\big((Y-\psi(\boldsymbol{X}))^2\big) &= \mathbb{E}\big((Y-\psi_0(\boldsymbol{X}))^2\big) - 2\mathbb{E}\big(\tilde{\psi}(\boldsymbol{X})(Y-\psi_0(\boldsymbol{X}))\big) \\ &\quad + \mathbb{E}\big(\tilde{\psi}(\boldsymbol{X})^2\big)\end{aligned}$$

Or

$$\begin{aligned}\mathbb{E}\big(\tilde{\psi}(\boldsymbol{X})\psi_0(\boldsymbol{X})\big) &= \int_{\mathbb{R}^n} \tilde{\psi}(\boldsymbol{x})\psi_0(\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \tilde{\psi}(\boldsymbol{x})\Big(\int_{\mathbb{R}} y p_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy\Big)p_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x} \\ &= \iint_{\mathbb{R}^n\times\mathbb{R}} \tilde{\psi}(\boldsymbol{x}) y p_{\boldsymbol{X},Y}(\boldsymbol{x},y)d\boldsymbol{x}dy = \mathbb{E}\big(\tilde{\psi}(\boldsymbol{X})Y\big)\end{aligned}$$

Donc

$$\mathbb{E}\big((Y-\psi(\boldsymbol{X}))^2\big) = \mathbb{E}\big((Y-\psi_0(\boldsymbol{X}))^2\big) + \mathbb{E}\big(\tilde{\psi}(\boldsymbol{X})^2\big)$$

dont le minimum (en $\tilde{\psi}$) est atteint pour $\tilde{\psi} = 0$.

• Théorème de conditionnement gaussien. Soit $\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}$ un vecteur aléatoire gaussien (avec $\boldsymbol{Y}_1$ de taille n et $\boldsymbol{Y}_2$ de taille p) :

$$\mathcal{L}\left(\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \right) \quad = \quad \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \right)$$

avec les vecteurs moyennes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ de tailles n_1 et n_2, les matrices de covariance $\mathbf{R}_{11}$ de taille $n_1 \times n_1$, $\mathbf{R}_{12}$ de taille $n_1 \times n_2$, $\mathbf{R}_{21} = \mathbf{R}_{12}^T$ de taille $n_2 \times n_1$, et $\mathbf{R}_{22}$ de taille $n_2 \times n_2$.
Alors la loi de $\boldsymbol{Y}_1$ conditionnellement à $\boldsymbol{Y}_2$ est gaussienne :

$$\mathcal{L}(\boldsymbol{Y}_1 | \boldsymbol{Y}_2 = \boldsymbol{y}_2) \quad = \quad \mathcal{N}(\boldsymbol{\mu}_1 + \mathbf{R}_{12}\mathbf{R}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21})$$

Preuve : La densité de la loi conditionnelle de $\boldsymbol{Y}_1$ sachant $\boldsymbol{Y}_2$ est :

$$p_{\boldsymbol{Y}_1|\boldsymbol{Y}_2}(\boldsymbol{y}_1|\boldsymbol{y}_2) = \frac{p_{\boldsymbol{Y}_1,\boldsymbol{Y}_2}(\boldsymbol{y}_1,\boldsymbol{y}_2)}{p_{\boldsymbol{Y}_2}(\boldsymbol{y}_2)} \propto p_{\boldsymbol{Y}_1,\boldsymbol{Y}_2}(\boldsymbol{y}_1,\boldsymbol{y}_2)$$

C'est (à constante multiplicative près) l'exponentielle d'une forme quadratique en $\boldsymbol{y}_1$: c'est donc la densité d'une loi gaussienne.
Reste à identifier la moyenne et la covariance.
Cas $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}^T \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = (\boldsymbol{y}_1 - \boldsymbol{\mu}_{\text{post}})^T \mathbf{R}_{\text{post}}^{-1} (\boldsymbol{y}_1 - \boldsymbol{\mu}_{\text{post}}) + \textit{cst}$$

$\rightarrow$ donne le résultat avec un peu d'algèbre linéaire.

Exemple. On considère le vecteur gaussien ($n = p = 1$)

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

avec $\rho \in [-1, 1]$ coefficient de corrélation de Y_1 et Y_2.
La loi de Y_1 est

$$\mathcal{L}(Y_1) \quad = \quad \mathcal{N}(0, 1)$$

Si on observe $Y_2 = y_2$, alors :

$$\mathcal{L}(Y_1 | Y_2 = y_2) \quad = \quad \mathcal{N}(\rho y_2, 1 - \rho^2)$$

- la moyenne de Y_1 est attirée (si $\rho > 0$) ou rejetée (si $\rho < 0$) par l'observation.
- la variance de Y_1 est réduite (si $\rho \neq 0$).

Cas extrêmes : $\rho = 0$ (on n'apprend rien de l'observation de Y_2) et $\rho = 1$ (on sait tout une fois qu'on a observé Y_2).

Soit $(\boldsymbol{X}, Y)$ un vecteur aléatoire, $\boldsymbol{X}$ à valeurs dans $\mathbb{R}^n$ et Y à valeurs dans $\mathbb{R}$.

• Régression linéaire.

On observe $\boldsymbol{X}$. Quelle est la meilleure combinaison affine de $\boldsymbol{X}$ qui approche au mieux Y ?

On cherche donc à résoudre :

$$(\alpha_j^{reg})_{j=0}^n = \underset{(\alpha_j)_{j=0}^n \in \mathbb{R}^{n+1}}{\operatorname{argmin}} \mathbb{E}\Big[\big(Y - \alpha_0 - \sum_{j=1}^n \alpha_j X_j\big)^2\Big]$$

et on obtient $Y^{reg} = \alpha_0^{reg} + \sum_{j=1}^n \alpha_j^{reg} X_j$.

Problème de minimisation (quadratique) beaucoup plus simple que celui qui correspond à l'espérance conditionnelle.

Mais la régression est sous-optimale du point de vue de l'approximation : la meilleure combinaison affine de $\boldsymbol{X}$ n'est pas forcément la meilleure approximation de Y sachant $\boldsymbol{X}$.

• Résultat : Si $(\boldsymbol{X}, Y)$ est gaussien, alors l'espérance conditionnelle $\mathbb{E}[Y|\boldsymbol{X}]$ est la régression linéaire de Y sur $\boldsymbol{X}$.

Méthodes d'estimation classique en statistique

• Objectif : estimer un certain nombre de paramètres inconnus, comme par exemple les paramètres β de la densité $p_\beta(x)$ d'une variable aléatoire X, en fonction d'un échantillon de cette loi.

• n-échantillon : une suite de n réalisations $(X_k)_{k=1,\dots,n}$ tirées de manière indépendante selon la loi de densité p_{β^*}, avec β^* inconnu.

• On passe en revue trois méthodes classiques :

- estimation empirique
- estimation par maximum de vraisemblance
- estimation par maximum a posteriori (bayésienne)

- Estimateur empirique.

Dans le cas où les paramètres $\boldsymbol{\beta} = (\beta_q)_{q=1,\dots,Q}$ peuvent s'exprimer comme des moments :

$$\beta_q^* = \mathbb{E}[f_q(X)]$$

alors on peut proposer comme estimateur

$$\widehat{\beta}_q = \frac{1}{n}\sum_{k=1}^{n} f_q(X_k)$$

- L'estimateur est sans biais ($\mathbb{E}[\widehat{\beta}_q] = \beta_q^*$), convergent ($\widehat{\beta}_q \overset{n\to\infty}{\longrightarrow} \beta_q^*$ p.s), asymptotiquement normal :

$$\sqrt{n}(\widehat{\beta}_q - \beta_q^*) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0, \mathrm{Var}(f_q(X))\big)$$

L'erreur quadratique moyenne est en $1/n$:

$$\mathbb{E}\big[(\widehat{\beta}_q - \beta_q^*)^2\big] = \frac{1}{n}\mathrm{Var}(f_q(X))$$

(hypothèse : $\mathbb{E}[f_q^2(X)] < \infty$).

- Exemple : $\beta = (\mu, \sigma^2)$, $p_\beta(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \dfrac{(x-\mu)^2}{2\sigma^2} \Big)$.

On a :

$$\mu = \mathbb{E}[X], \qquad \sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Estimateurs empiriques :

$$\widehat{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k, \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \widehat{\mu}^2 = \frac{1}{n}\sum_{k=1}^{n} (X_k - \widehat{\mu})^2$$

Remarque : ici, $\widehat{\sigma}^2$ est biaisé, $\mathbb{E}[\widehat{\sigma}^2] = \frac{n-1}{n}\sigma^2$.
Estimateur non-biaisé : $\widetilde{\sigma}^2 = \frac{n}{n-1}\widehat{\sigma}^2$.

- Estimateur du maximum de vraisemblance (MV).

Vraisemblance des données $(X_k)_{k=1,\dots,n}$ sachant les paramètres $\boldsymbol{\beta}$:

$$\prod_{k=1}^{n} p_{\boldsymbol{\beta}}(X_k)$$

Formule de Bayes : vraisemblance de $\boldsymbol{\beta}$ sachant les données $\boldsymbol{X} = (X_k)_{k=1,\dots,n}$:

$$L_{\boldsymbol{X}}(\boldsymbol{\beta}) := \prod_{k=1}^{n} p_{\boldsymbol{\beta}}(X_k)$$

à une constante multiplicative près (qui ne dépend que de $\boldsymbol{X}$).

Estimateur MV : Consiste à trouver la valeur $\widehat{\boldsymbol{\beta}}$ qui maximise $L_{\boldsymbol{X}}(\boldsymbol{\beta})$.

• Sous de bonnes hypothèses (modèle régulier) :
L'estimateur MV est asymptotiquement sans biais, convergent, asymptotiquement normal, asymptotiquement efficace :

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \overset{n\to\infty}{\longrightarrow} \mathcal{N}\big(0, \mathbf{I}_{\boldsymbol{\beta}^*}^{-1}\big)$$

où $\mathbf{I}_{\boldsymbol{\beta}}$ est la matrice d'information de Fisher :

$$\mathbf{I}_{\boldsymbol{\beta}} = -\mathbb{E}_{\beta}\big[\nabla_{\boldsymbol{\beta}} \otimes \nabla_{\boldsymbol{\beta}} \ln p_{\boldsymbol{\beta}}(X)\big] = -\int \big[\nabla_{\boldsymbol{\beta}} \otimes \nabla_{\boldsymbol{\beta}} \ln p_{\boldsymbol{\beta}}(x)\big] p_{\boldsymbol{\beta}}(x) dx$$

Résultat (borne de Cramer-Rao) : quel que soit l'estimateur sans biais $\tilde{\boldsymbol{\beta}} = \boldsymbol{\psi}_n(X_1, \ldots, X_n)$, on a $\mathbf{Cov}(\tilde{\boldsymbol{\beta}}) \geq [n\mathbf{I}_{\boldsymbol{\beta}^*}]^{-1}$.
Un estimateur est dit efficace s'il atteint la borne de Cramer-Rao.

- Exemple : $\boldsymbol{\beta} = (\mu, \sigma^2)$, $p_{\boldsymbol{\beta}}(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \dfrac{(x-\mu)^2}{2\sigma^2}\Big)$.

On maximise $L_{\boldsymbol{X}}(\boldsymbol{\beta})$:

$$L_{\boldsymbol{X}}(\boldsymbol{\beta}) \approx \frac{1}{\sigma^n} \exp\Big(- \sum_{k=1}^{n} \frac{(X_k - \mu)^2}{2\sigma^2}\Big)$$

en trouvant le point $\widehat{\boldsymbol{\beta}} = (\widehat{\mu}, \widehat{\sigma}^2)$ tel que $\nabla_{\boldsymbol{\beta}} \ln L_{\boldsymbol{X}}(\boldsymbol{\beta}) = \mathbf{0}$:

$$\frac{\partial}{\partial \mu}\Big(- \sum_{k=1}^{n} \frac{(X_k - \mu)^2}{2\sigma^2}\Big) = 0, \qquad \frac{\partial}{\partial \sigma^2}\Big(- \sum_{k=1}^{n} \frac{(X_k - \mu)^2}{2\sigma^2} - \frac{n}{2} \ln \sigma^2\Big) = 0$$

$$\hookrightarrow \qquad \widehat{\mu} = \frac{1}{n} \sum_{k=1}^{n} X_k, \qquad \widehat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \widehat{\mu})^2$$

- On peut parfois battre la borne de Cramer-Rao en $1/n$.

Exemple : $(X_k)_{k=1,\dots,n}$ i.i.d. de loi uniforme sur $[0, \theta^*]$, θ^* inconnu.
En maximisant

$$L_{\boldsymbol{x}}(\theta) = \prod_i p_\theta(x_i) = \theta^{-n} \mathbf{1}_{[0,+\infty[}(\min_i x_i) \mathbf{1}_{]-\infty,\theta]}(\max_i x_i)$$

on trouve l'estimateur MV :

$$\hat{\theta} = \max_{k=1,\dots,n} (X_k)$$

On a

$$\mathbb{P}(\hat{\theta} \leq t) = \mathbb{P}(X_1 \leq t)^n = \begin{cases} 1 \text{ si } t > \theta^* \\ \left(\frac{t}{\theta^*}\right)^n \text{ si } t \leq \theta^* \end{cases}$$

Donc $\hat{\theta}$ est à densité $p(t) = n t^{n-1} \theta^{*-n} \mathbf{1}_{[0,\theta^*]}(t)$. On trouve

$$\mathbb{E}[\hat{\theta}] = \frac{n}{n+1}\theta^*, \qquad \mathbb{E}\left[(\hat{\theta} - \theta^*)^2\right] = \frac{2}{(n+1)(n+2)}\theta^{*2}$$

En corrigeant le biais: $\check{\theta}_n = \frac{n+1}{n}\hat{\theta}$ on obtient un estimateur de θ non-biaisé tel que $\mathrm{Var}(\check{\theta}_n) = \frac{1}{n(n+2)}\theta^{*2}$.

- Estimateur du maximum a posteriori (MAP).
- Très proche du maximum de vraisemblance.

Différence : prend en compte un a priori sur les paramètres à estimer.

- Cadre bayésien.

Estimateur MAP = mode de la loi a posteriori

$$p_{\text{post}}(\boldsymbol{\beta}) \approx L_{\boldsymbol{X}}(\boldsymbol{\beta}) p_{\text{prior}}(\boldsymbol{\beta})$$

où $L_{\boldsymbol{X}}(\boldsymbol{\beta})$ est la vraisemblance et $p_{\text{prior}}(\boldsymbol{\beta})$ la distribution a priori des paramètres $\boldsymbol{\beta}$.

L'estimateur du maximum de vraisemblance est l'estimateur MAP pour une distribution a priori uniforme.

- Exemple : $\beta = (\mu, \sigma^2)$, $p_\beta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

Prior sur μ : $p_{\text{prior}}(\mu) = \frac{1}{\sqrt{2\pi\mu_0^2}} \exp(-\frac{\mu^2}{2\mu_0^2})$.

On maximise $L_{\boldsymbol{X}}(\beta) p_{\text{prior}}(\beta)$ en trouvant le point $\widehat{\beta}$ tel que $\nabla_\beta \ln L_{\boldsymbol{X}}(\beta) p_{\text{prior}}(\beta) = \mathbf{0}$:

$$\frac{\partial}{\partial \mu}\Big(-\sum_{k=1}^{n} \frac{(X_k - \mu)^2}{2\sigma^2} - \frac{\mu^2}{2\mu_0^2}\Big) = 0, \qquad \frac{\partial}{\partial \sigma^2}\Big(-\sum_{k=1}^{n} \frac{(X_k - \mu)^2}{2\sigma^2} - \frac{n}{2} \ln \sigma^2\Big) = 0$$

$$\hookrightarrow \qquad \widehat{\mu} = \frac{1}{1 + \frac{\widehat{\sigma}^2}{\mu_0^2}} \frac{1}{n} \sum_{k=1}^{n} X_k, \qquad \widehat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \widehat{\mu})^2$$

$\hookrightarrow$ Equation cubique à résoudre.

Sources d'incertitudes

Sources d'incertitudes

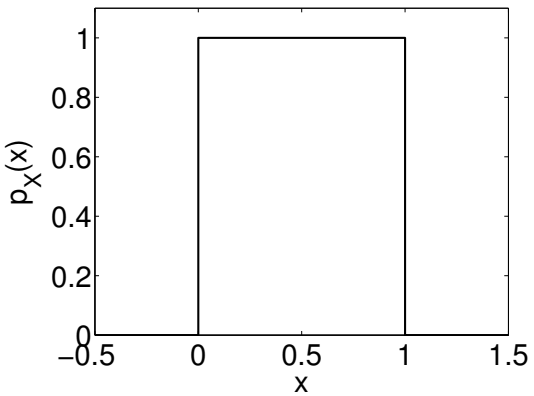
On peut distinguer deux types d'entrée "incertaines" :

- variables stochastiques : ces variables ont une variabilité naturelle résultant de phénomènes aléatoires (typiquement, une quantité soumise à des fluctuations dans un procédé de fabrication).
- variables épistémiques : ces variables possèdent une valeur mais elle nous est inconnue, à cause d'un manque de connaissance (typiquement, une constante d'une loi physique).

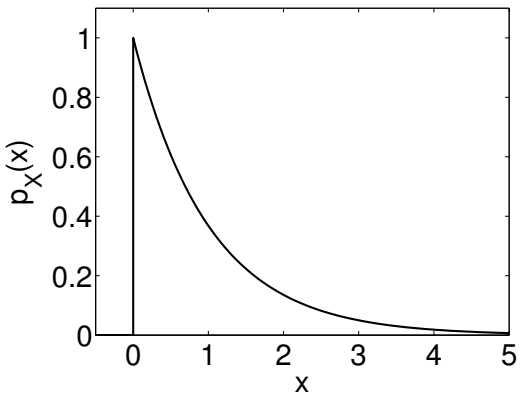
Sources d'incertitudes

- Dans la suite on modélisera les variables d'entrée comme des variables aléatoires de lois de probabilités données. La loi de probabilité d'une variable aléatoire réelle X caractérise la probabilité $\mathbb{P}(X \in [a, b])$ pour tout $a < b$.
- Dans le cadre de ce cours, on considérera (pour simplifier) uniquement des variables aléatoires d'entrée réelles continues. Dans ce cas la loi est donnée par la densité $(p(x))_{x \in \mathbb{R}}$ telle que

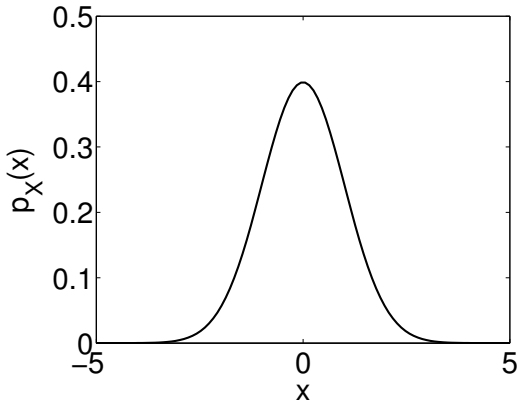
$$\mathbb{P}(X \in [a, b]) = \int_a^b p(x)dx.$$



uniforme



exponentielle



gaussienne

Estimateur non-paramétrique à noyau

- Soit X une variable réelle de densité $(p(x))_{x\in\mathbb{R}}$ inconnue.

On dispose d'un échantillon $(x_i)_{i=1,\dots,n}$.

- Estimateur à noyau de la densité :

$$\hat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big),$$

où K est le noyau et h la taille du noyau.

- Le noyau doit être une fonction réelle positive paire normalisée de telle sorte que $\int_{-\infty}^{\infty} x^2 K(x)dx = 1$ et $\int_{-\infty}^{\infty} K(x)dx = 1$; on prend souvent un noyau gaussien $K(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})$.
- La taille du noyau h est le paramètre critique de la méthode, elle régit le niveau de lissage.

• Règle de Silverman : Le h optimal minimisant l'erreur quadratique moyenne :

$$\mathcal{E} = \mathbb{E}\Big[\int_{\mathbb{R}} (\hat{p}(x) - p(x))^2 dx\Big],$$

est

$$h_{opt} = \frac{\|K\|_{L^2}^{2/5}}{\|p''\|_{L^2}^{2/5}} n^{-1/5} + o(n^{-1/5}),$$

quand p est régulière.

• En pratique : $h = \hat{\sigma} n^{-1/5}$ avec $\hat{\sigma}$ l'écart-type empirique de l'échantillon.

Preuve : L'erreur quadratique locale est la somme de deux termes (variance et biais) :

$$\mathbb{E}\big[(\hat{p}(x)-p(x))^2\big] = \mathrm{Var}\big(\hat{p}(x)\big) + \big(\mathbb{E}[\hat{p}(x)]-p(x)\big)^2.$$

Comme les x_i sont i.i.d. de loi de densité p, la variance est égale à :

$$\begin{aligned}
\mathrm{Var}\big(\hat{p}(x)\big) &= \frac{1}{nh^2}\mathrm{Var}\big(K(\frac{x_1-x}{h})\big) \\
&= \frac{1}{nh^2}\Big[\int K(\frac{y-x}{h})^2 p(y)dy - \Big(\int K(\frac{y-x}{h})p(y)dy\Big)^2\Big] \\
&= \frac{1}{nh}\Big[p(x)\int K(z)^2dz + o(h)\Big],
\end{aligned}$$

et le terme de biais est :

$$\begin{aligned}
\mathbb{E}[\hat{p}(x)]-p(x) &= \frac{1}{h}\int K(\frac{y-x}{h})\big(p(y)-p(x)\big)dy \\
&= \frac{h^2}{2}p''(x)\int K(z)z^2dz + o(h^2).
\end{aligned}$$

On trouve donc que l'erreur quadratique moyenne est asymptotiquement la somme de deux termes (pour h petit) :

$$\mathcal{E} = \frac{\|K\|_{L^2}^2}{nh}\big(1+o(1)\big) + \frac{\|p''\|_{L^2}^2 h^4}{4}\big(1+o(1)\big).$$

- Le premier terme est le terme de variance : quand on prend h trop petit, la densité estimée est trop bruitée.
- Le second terme est le terme de biais : quand on prend h trop grand, on lisse trop la densité estimée.

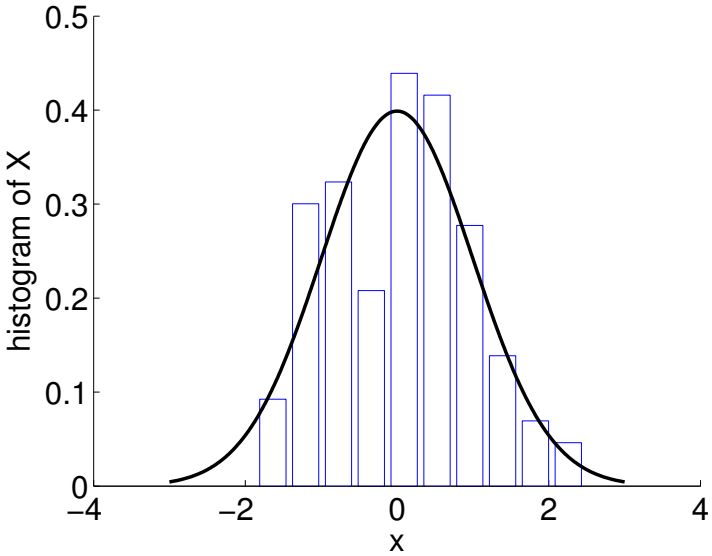
La minimisation de cette somme donne le résultat.

Résultats supplémentaires :

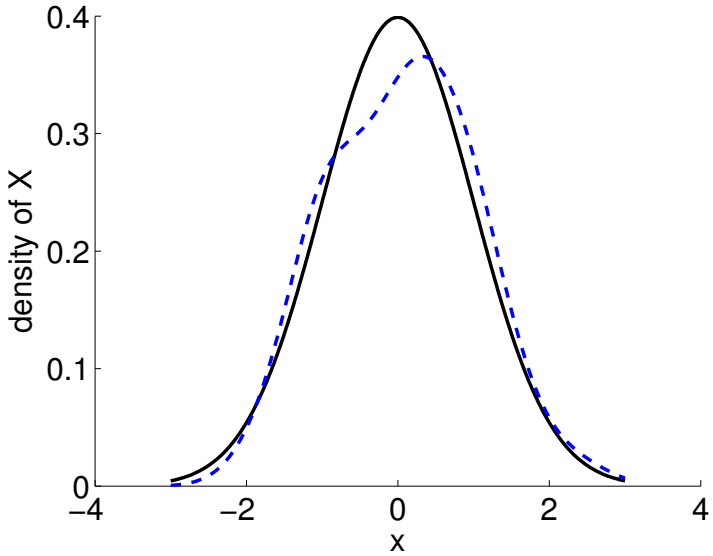
- Il n'existe pas d'estimateur non-paramétrique qui converge plus vite que l'estimateur à noyau avec le choix de la fenêtre $h \simeq n^{-1/5}$.
- La vitesse de convergence ($n^{-4/5}$ pour l'erreur quadratique moyenne) est plus faible que la vitesse typique des méthodes paramétriques (typiquement n^{-1}).
- La méthode à noyau supporte assez mal de monter en dimension : si $\boldsymbol{X}$ est à valeurs dans $\mathbb{R}^d$, alors on peut proposer un estimateur à noyau de sa densité de la forme

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h})$$

En reprenant les calculs précédents on trouve que le h optimal est de la forme $h \simeq n^{-1/(4+d)}$ et l'erreur quadratique moyenne est en $n^{-4/(4+d)}$.



histogramme



méthode à noyau

Comparaison entre estimation d'une densité par un histogramme et une méthode à noyau (200 données tirées selon une loi gaussienne $\mathcal{N}(0,1)$).

Estimateur paramétrique

- Soit X une variable réelle de densité $(p(x))_{x\in\mathbb{R}}$ inconnue. On dispose d'un échantillon $(x_i)_{i=1,\dots,n}$.
- On admet que cette densité appartient à une famille paramétrique :

$$\mathcal{P} = \left\{(p_{\boldsymbol{\theta}}(x))_{x\in\mathbb{R}}, \boldsymbol{\theta} \in \Theta\right\},$$

où $\Theta \subset \mathbb{R}^q$ pour un certain $q \in \mathbb{N}^*$ et $p_{\boldsymbol{\theta}}(x)$ est une densité de probabilité pour tout $\boldsymbol{\theta} \in \Theta$.
I.e., on suppose qu'il existe $\boldsymbol{\theta}^* \in \Theta$ tel que la densité inconnue est égale à $p_{\boldsymbol{\theta}^*}(x)$

- On estime $\boldsymbol{\theta}^*$ par une méthode d'estimation usuelle, méthode des moments ou méthode du maximum de vraisemblance par exemple.

• Si la famille de densités est régulière (vis-à-vis des paramètres), alors l'erreur d'estimation est en $1/n$:

$$\big(\mathbb{E}\big[(\hat{\theta}_i - \theta_i^*)(\hat{\theta}_j - \theta_j^*)\big]\big)_{i,j=1}^{q} = \frac{1}{n}(\mathbf{I}_{\boldsymbol{\theta}^*})^{-1} + o(\frac{1}{n})$$

• L'erreur quadratique moyenne

$$\mathcal{E} = \mathbb{E}\Big[\int_{\mathbb{R}} (p_{\hat{\boldsymbol{\theta}}}(x) - p_{\boldsymbol{\theta}^*}(x))^2 dx\Big]$$

est elle aussi en $1/n$:

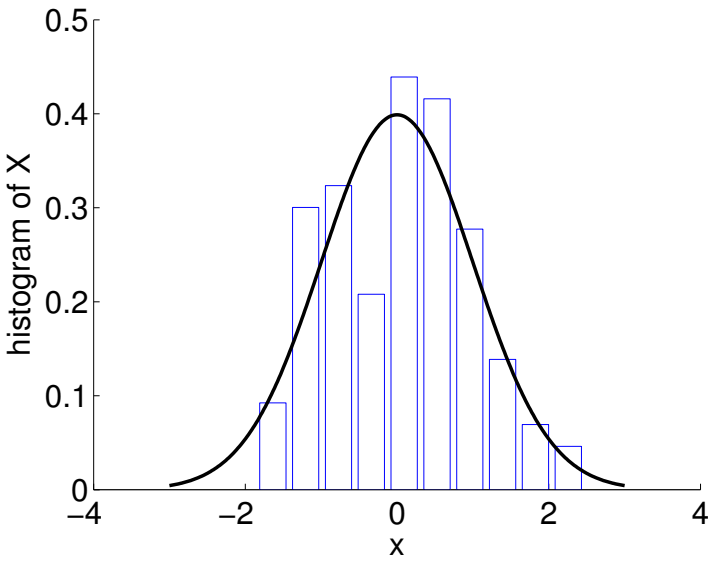
$$\mathcal{E} = \sum_{i,j=1}^{q} \int_{\mathbb{R}} \partial_{\theta_i} p_{\boldsymbol{\theta}^*}(x) \partial_{\theta_j} p_{\boldsymbol{\theta}^*}(x) dx \, \mathbb{E}\big[(\hat{\theta}_i - \theta_i^*)(\hat{\theta}_j - \theta_j^*)\big] + o(\frac{1}{n}).$$

Exemple : Famille de lois gaussiennes :

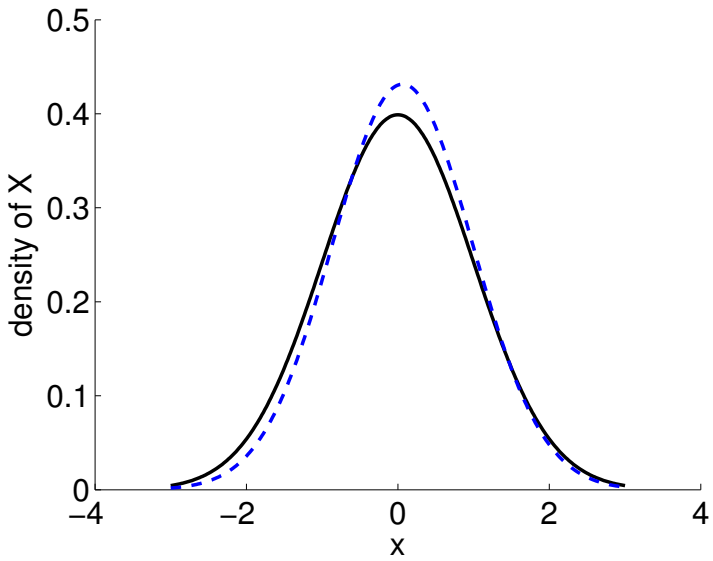
$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big)$$

Ici $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times]0, +\infty[$.
Estimateur des moments = estimateur MV =

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2.$$



histogramme



méthode des moments

Comparaison entre un histogramme (gauche) et la méthode des moments (droite) (200 données tirées selon une loi gaussienne $\mathcal{N}(0,1)$).

Exemple : $(X_k)_{k=1,\dots,n}$ i.i.d. de loi uniforme sur $[0,\theta^*]$, θ^* inconnu.
L'estimateur MV :

$$\hat{\theta} = \max_{k=1,\dots,n}(X_k)$$

satisfait :

$$\mathbb{E}[\hat{\theta}] = \frac{n}{n+1}\theta^*, \qquad \mathbb{E}\big[(\hat{\theta}-\theta^*)^2\big] = \frac{2}{(n+1)(n+2)}\theta^{*2}$$

C'est un exemple de familles non-régulières (la densité n'est pas dérivable en θ), pour lesquelles on peut arriver à une erreur quadratique pour l'estimation du paramètre plus petite que $1/n$.
Mais comme la densité n'est pas régulière, on perd cet avantage lorsqu'on considère l'erreur quadratique moyenne $\mathcal{E}$:

$$\mathbb{E}\big[(p_{\hat{\theta}}(x)-p_{\theta^*}(x))^2\big] = \mathbb{E}\Big[\big(\frac{1}{\theta^*}-\frac{1}{\hat{\theta}}\big)^2\mathbf{1}_{[0,\hat{\theta}[}(x)+\frac{1}{\theta^{*2}}\mathbf{1}_{[\hat{\theta},\theta^*]}(x)\Big],$$

puis (pourvu que $n \geq 2$) :

$$\mathcal{E} = \int \mathbb{E}\big[(p_{\hat{\theta}}(x)-p_{\theta^*}(x))^2\big]dx = \mathbb{E}\big[\big(\frac{1}{\theta^*}-\frac{1}{\hat{\theta}}\big)^2\hat{\theta}+\frac{1}{\theta^{*2}}(\theta^*-\hat{\theta})\big] = \mathbb{E}\big[\frac{1}{\hat{\theta}}\big]-\frac{1}{\theta^*},$$

qui est égal à $\mathcal{E} = \frac{1}{n-1}\frac{1}{\theta^*}$, en $1/n$.