Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2014/2015

Sujet 9: propagation des incertitudes et des distributions par la méthode de Monte-Carlo

---

On souhaite étudier comment l'incertitude portant sur les variables d'entrées d'un modèle se propage au travers d'un modèle donné sur la/les variable(s) en sortie. La démarche suit globalement la progression suivante[1]:

• identification des variables d'entrée incertaines $\mathbf{X} = (X^{(1)}, ..., X^{(K)})$. L'incertitude peut être de nature métrologique (qualité des outils employés, de l'expérience effectuée) ou épistémique (connaissance que l'on a sur la grandeur d'intérêt). Insistons sur le fait que *mesurande* et *grandeur mesurée* sont deux notions différentes. Le mesurande est ce que l'on veut mesurer tandis que la grandeur mesurée est ce que l'on peut mesurer. L'état des connaissances correspondant est décrit au mieux par une distribution sur l'ensemble des valeurs possibles du mesurande.

• définition du modèle (lois de la physique, code de calcul...) sous la forme:

$$y = \varphi(\mathbf{x})$$

en notant $\mathbf{x} = (x^{(1)}, ..., x^{(K)})$.

• définition de la loi de probabilité jointe des variables d'entrée $\mathbf{X} = (X^{(1)}, ..., X^{(K)})$, notée $P^{\mathbf{X}}$.

- type A: c'est les cas où des données sont disponibles sur la variable d'entrée considérée sous la forme d'un échantillon i.i.d. représentatif.
  On peut ajuster une loi paramétrique choisie dans une famille donnée ou bien on estime la distribution de façon non-paramétrique (histogramme, estimateur à noyau,... pourvu que l'on ait beaucoup de données disponibles!). Le GUM préconise d'estimer l'incertitude, par exemple $u(X^{(1)})$ portant sur la variable d'entrée $X^{(1)}$, à partir d'un échantillon i.i.d. de taille n noté $(X_1^{(1)}, ..., X_n^{(n)})$ par l'écart-type de la moyenne empirique, ce qui donne

  $$\widehat{u}_n(X^{(1)}) = \widehat{\sigma}_n(\overline{(X^{(1)})}_n) = \frac{\widehat{\sigma}_n(X^{(1)})}{\sqrt{n}}$$

  en notant $\overline{(X^{(1)})}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^{(1)}$ et $\widehat{\sigma}_n(X^{(1)}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(1)} - \overline{(X^{(1)})}_n)^2$.

- type B: c'est le cas de l'absence de données. On a recours à des méthodes d'élicitation. Cela consiste à traduire formellement l'avis d'experts "métier" en une loi de probabilité. Une manière de construire des lois de probabilité à partir d'informations "minimalistes" consiste à s'appuyer sur la méthode du maximum d'entropie (MaxEnt). L'entropie de

[1] voir aussi le GUM = *Guide to the expression of Uncertainty in Measurement*, International organization for standardization, 1995

Shannon (1948) et Jaynes (1957) [2] d'une distribution de densité f par rapport à une mesure μ positive et σ-finie est définie par:

$$H(f) = -\int f(x)\log f(x)d\mu(x)\,.$$

L'entropie de Shannon s'interprète comme une mesure de l' inverse de l'information apportée sur $\mathbf{X}$ par sa loi de probabilité au sens où l'entropie est minimale en cas d'information "parfaite" (ie lorsqu'il n'y a aucun doute sur la valeur prise par $\mathbf{X}$ dans l'espace des valeurs possibles) et où l'entropie est maximale dans le cas où l'information apportée par la loi de probabilité est la plus vague possible (ie les valeurs possibles pour $\mathbf{X}$ sont équiprobables). Le principe du maximum d'entropie consiste à choisir, parmi toutes les lois possibles, celle qui apporte le minimum d'information, c'est donc celle qui maximise l'entropie. Ce principe se justifie heuristiquement par la recherche d'objectivité qui consiste ici à ne rajouter aucune information à celles fournies par les experts. Ainsi, si un expert donne J informations sur la loi de $\mathbf{X}$ sous la forme:

$$\int \psi_j(\mathbf{x})dP^{\mathbf{X}}(\mathbf{x}) = c_j\,.$$

Le principe MaxEnt devient un problème d'optimisation sous contraintes puisque l'on recherche une fonction f qui maximise $H(.)$ et qui respecte les $(J+1)$ conditions suivantes:

$$\begin{cases} \int f(\mathbf{x})d\mu(\mathbf{x}) = 1, & \\ \int \psi_j(\mathbf{x})dP^{\mathbf{X}}(\mathbf{x}) = c_j, & \text{pour } j = 1, ..., J. \end{cases}$$

information fournie	distribution minimisant l'entropie
$X \in [a,b]$ p.s., pas de valeur préférable aux autres	$X \sim \mathcal{U}(a,b)$: loi uniforme = rectangulaire
$\mathbb{E}[X] = \mu$ et $X > 0$ p.s.	$X \sim \mathcal{E}1/\mu)$: loi exponentielle
$\mathbb{E}[X] = \mu$ et $\mathrm{Var}(X) = \sigma^2$	$X \sim \mathcal{N}(\mu, \sigma^2)$: loi gaussienne

D'autres choix sont possibles, par exemple:

- loi triangulaire quand l'expert fournit un intervalle $[a,b]$ et un mode c:

$$f_{a,b,c}(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}, & \text{si } x \in [a,c], \\ \dfrac{2(b-x)}{(b-a)(b-c)}, & \text{si } x \in [c,b], \\ 0, & \text{sinon.} \end{cases}$$

- loi beta pour une proportion dans le continuum $]0,1[$,
- loi dérivée d'arc-sinus: variation en forme de "U" entre deux limites a et b:

$$f_{a,b}(x) = \frac{1}{\pi\sqrt{(x-a)(b-x)}} I(x \in [a,b]),$$

$$F_{a,b}(x) = \frac{2}{\pi}\arcsin\left(\sqrt{\frac{x-a}{b-a}}\right),$$

[2] Jaynes E.T. (1957) Information theory and statistical mechanics, *Physical review*, 106:620-630.

- loi de Poisson pour des comptages dans un intervalle de temps(ou d'espace) donné (nombre de requêtes à un serveur en $1s$),
- loi log-normale, Gamma, chi-deux pour des variables positives à distribution asymétrique,
- loi exponentielle pour des temps d'attente sans mémoire (durée de vie de systèmes sans usure),
- loi de Weibull pour des durées de vie d'un matériel qui se dégrade ($\alpha > 1$, prothèse) ou qui se bonifie ($\alpha < 1$, résistance de béton en début de vie),
- loi de Gumbell pour modéliser des phénomènes climatiques potentiellement extrêmes (niveau d'une rivière), fortement asymétrique autour du mode, les fortes valeurs restant tout à fait probables,
- loi normale multivariée pour modéliser des entrées gaussiennes corrélées.
  NB: la modélisation de la dépendance autre que gaussienne fait généralement appel à la théorie des copules...

NB: Il existe d'autres méthodes basées sur la théorie bayésienne.

• Etude de la propagation des incertitudes et des distributions, ici, par méthode stochastique (Monte-Carlo) et présentation du tracé de la distribution de Y et/ou des résumés statistiques de la distribution de Y:

- choix de la taille de l'échantillon simulée, notée N.
- simulation de N vecteurs indépendants et distribués comme $\mathbf{X} = (X^{(1)i}, ..., X^{(K)})$ ie de loi $P^{\mathbf{X}}$ identifiée précédemment. On note $\mathbf{X}_i = (X_i^{(1)}, ..., X_i^{(K)})$ pour $i = 1, ..., N$ ces vecteurs.
- calculer $Y_i = \varphi(\mathbf{X}_i)$ pour $i = 1, ..., N$.
- estimer
  - l'espérance $\mathbb{E}[Y]$ par $\widehat{\mathbb{E}[Y]}_N = \overline{Y}_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \varphi(\mathbf{X}_i)$
  - la variance $\mathrm{Var}(Y)$ par $S_N'^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N} \left(Y_i - \overline{Y}_N\right)^2$ ou de manière équivalente l'écart-type $\sigma(Y)$ par $S_N' = \sqrt{S_N'^2}$
  - l'incertitude-type (ou standard) $u(Y)$ par $\widehat{u(Y)}_N = \dfrac{S_N'}{\sqrt{N}}$
  - l'incertitude relative $\dfrac{u(Y)}{\mathbb{E}[Y]}$ par:
  $$\frac{\widehat{u(Y)}_N}{\widehat{\mathbb{E}[Y]}_N}.$$
  - les moments d'ordre supérieur (notamment le coefficient d'asymétrie).
  - l'intervalle de fluctuation pour Y au niveau de confiance approché de $(1-\alpha)$ par les quantiles empiriques de $(Y_i)_{i=1,...,N}$ d'ordre respectif $\alpha/2$ et $1-\alpha/2$ notés $\widehat{q}_{\alpha/2}$ et $\widehat{q}_{1-\alpha/2}$

– étude de la propagation des distributions: Cela consiste en la détermination de la loi de probabilité de la sortie $Y = \varphi(\mathbf{X})$. On estime la distribution de Y par sa fonction de répartition empirique définie pour $t \in \mathbb{R}$ par

$$t \to \frac{1}{N} \sum_{i=1}^{N} I(Y_i \leq t)$$

ou par un estimateur de la densité (histogramme, estimateur à noyau,...).

Exercice 1.
Considérons les modèles suivants

1. $y = x_1 + x_2$
2. $y = x_1 x_2$
3. $y = \dfrac{x_1}{x_2}$
4. $y = x_1 + x_2 + x_3 + x_4$ avec x_4 beaucoup plus grand que les autres, puis avec l'incertitude sur x_4 beaucoup plus grande que les autres
5. $y = b_0 + b_1 \exp(b_2 x_1) + b_3 \sin(b_4 x_2)$

Etudier la propagation d'incertitude et de distribution

1. dans les modèles 1), 2) et 3), lorsqu'en entrée, on considère
   (a) deux variables X_1 et X_2 indépendantes de loi gaussienne,
   (b) deux variables X_1 et X_2 corrélées positivement de loi gaussienne,
   (c) deux variables X_1 et X_2 corrélées négativement de loi gaussienne,
   (d) deux variables X_1 et X_2 indépendantes de loi uniforme,
   (e) deux variables X_1 et X_2 indépendantes de loi Gamma (avec une asymétrie marquée),
   (f) deux variables X_1 et X_2 indépendantes de loi estimées à partir d'un échantillon de taille $n = 150$ dans les deux cas,
   (g) deux variables X_1 et X_2 indépendantes de loi estimées à partir d'un échantillon de taille respective $n_1 = 20$ et $n_2 = 200$,
2. dans le modèle 4), lorsqu'en entrée, on considère
   (a) des variables X_1, X_2, X_3 et X_4 indépendantes avec $\mathbb{E}[X_4]$ beaucoup plus grande que les espérances des trois autres variables,
   (b) des variables X_1, X_2, X_3 et X_4 indépendantes avec $u(X_4)$ beaucoup plus grande que les incertitudes des trois autres variables,
3. dans le modèle 5), lorsqu'en entrée, on considère
   (a) deux variables X_1 et X_2 indépendantes de loi gaussienne,

(b) deux variables X_1 et X_2 corrélées positivement de loi gaussienne,

(c) deux variables X_1 et X_2 corrélées négativement de loi gaussienne.

NB: on admettra que pour simuler des variables X_i selon une loi de fonction de répartition F_X inconnue mais estimée par la fonction de répartition empirique, il suffit de tirer des variables indépendamment et avec remise dans l'échantillon des observations $(X_1, ..., X_n)$, ce qui s'effectue au moyen de l'instruction `sample` du logiciel `R`.