Liens entre analyse de sensibilité et problèmes d'inversion stochastique bien posés

Mélanie Blazère[1] & Nicolas Bousquet[1,2*]

[1] *Institut de Mathématique de Toulouse, Université Paul Sabatier*
[2] *Département Management des Risques Industriels*
EDF R&D, Chatou, France / nicolas.bousquet@edf.fr

[*] *Orateur.*

Résumé. Les problèmes d'inversion stochastique correspondent à l'estimation d'une distribution de probabilité caractérisant un paramètre de nature aléatoire, décrit comme une entrée d'un opérateur (typiquement un modèle numérique), à partir de la connaissance de données observées vues comme des sorties bruitées. Si de tels problèmes sont caractérisés par des conditions d'identifiabilité fortes, des conditions dites de "problème bien posé" du type "signal sur bruit" doivent être intégrées de façon préliminaire à la définition même du problème, avant de considérer la collecte de données suffisamment informatives. En supplément des conditions classiques de Hadamard, une nouvelle condition de "problème bien posé" est établie, fondée sur la transmission de l'incertitude des entrées aux sorties de l'opérateur, qui peut être perçue comme le résultat prédictif d'une analyse de sensibilité si le problème d'inversion était résolu. Un lien peut alors être fait entre condition de Hadamard et les indices de sensibilité classiques. Cette nouvelle condition s'exprime par l'ajout d'une contrainte *a priori* dans le problème numérique d'inversion, qui peut être traité de façon fréquentiste ou bayésienne. Celle-ci s'exprime relativement simplement dans le cas où l'opérateur est linéaire ou linéarisable. Comme ce type de situation engendre souvent des manques de contraste dans les données, on peut percevoir le cas linéaire ou linéarisable comme un cas "dur". On suggère donc que la contrainte exprimée simplement dans le cas linéaire ou linéarisable soit une contrainte utilisée de façon générique, sous réserve que l'opérateur possède de bonnes propriétés de différentiabilité.

Mots-clés. Problèmes bien posés ; Hadamard ; information de Fisher ; indices de Sobol.

Abstract. Stochastic inversion problems arise when it is wanted to estimate the probability distribution of a stochastic input from indirect observable and noisy information and the limited knowledge of an operator that connects the inputs to the observable output. While such problems are characterized by strong identifiability conditions, well-posedness conditions of "signal over noise" nature are prior that should be respected to collect observations. In addition to well-known Hadamard' well-posedness condition, a new one is established based on the predictive transmission of input uncertainty to output, which can be interpreted as the result provided by a sensitivity analysis if the problem were solved. This new condition should take part within the input model itself, which adds

a constraint in established frequentist or Bayesian methodologies of stochastic inversion. While this article mainly deals with linearizable operators, the lack of constrast typical of linear problems suggest that the proposed condition should be used in more general settings, provided the operator owns diffferentiability properties.

Keywords. Well-posed problems ; Hadamard ; Fisher information ; Sobol indices.

1 Introduction

On considère la situation où des observations $\mathbf{y_n^*} = (y_i^*)_{i\in\{1,\dots,n\}}$ vivant dans un espace de dimension q sont supposées être des réalisations d'une variable aléatoire Y^* telle que

$$Y^* = Y + \varepsilon, \tag{1}$$
$$Y = g(X) \tag{2}$$

où X est un vecteur aléatoire de dimension p d'une distribution inconnue $\mathcal{F}$, ε est un bruit de mesure et/ou de modèle de distribution connue f_ϵ et g est une fonction déterministe de $\mathbb{R}^p$ vers $\mathbb{R}^q$. Dans de nombreux cas d'étude industriels, g est décrite comme une fonction "boîte noire" ou un code numérique, qui ne peut être explorée qu'au moyen de la simulation. Plusieurs cadres statistiques permettent d'inférer sur $\mathcal{F}$ à partir de la connaissance de $\mathbf{y_n^*}$ et f_ϵ. La *calibration bayésienne* (Gamblin *et al.*, 2015; Kennedy et O'Hagan, 2001) suppose que X est un paramètre rendu aléatoire dans un cadre bayésien, pour lequel une mesure *a priori* $\pi(X)$ est disponible, et que $\mathcal{F}$ peut être estimée par la loi *a posteriori* $\pi(X|\mathbf{y_n})$. Dans de tels cas, X et Y sont considérées comme des variables épistémiques. L'*inversion stochastique* (Celeux *et al.*, 2010) suppose que (Y, X) sont intrinsèquement aléatoires et que $\mathcal{F}$ est une distribution qui ne dégénère pas lorsque $n \to \infty$. Ce problème d'inversion est alors noté

$$\hat{\mathcal{F}} = \mathcal{H}_g^{-1}(Y^*, Y, \varepsilon) \tag{3}$$

où $\mathcal{H}_g$ est un opérateur d'inversion induit par g. En général, on donne à $\mathcal{F}$ une forme paramétrique (souvent gaussienne, à une possible reparamétrisation près), et l'estimation de son vecteur de paramètre peut être faite dans un cadre classique (Barbillon *et al.*, 2011; Celeux *et al.*, 2010) ou bayésien Fu *et al.* (2015), possiblement au prix d'une linéarisation de g (Barbillon *et al.*, 2011), et par l'emploi d'algorithme à données manquantes. Dans les deux cadres, résoudre le problème d'inférence sur $\mathcal{F}$ nécessite de vérifier des conditions de problème bien posé et d'identifiabilité. La condition de problème bien posé au sens de Hadamard stipule que la solution $\hat{\mathcal{F}}$ doit exister, être unique et continûment dépendante des observations selon une topologie raisonnable. Elle se traduit par une faible valeur du nombre de condition (Belsley *et al.*, 1980) dans les problèmes linéarisés. Pour ce même type de problème, la condition d'identifiabilité est équivalente à l'injectivité de l'opérateur linéaire accompagnée de la contrainte $p \leq nq$ (Celeux *et al.*, 2010).

Une autre condition de problème bien posé, parlante pour les ingénieurs habitués au traitement des incertitudes, peut être introduite comme suit. Imaginons que le problème est résolu et $\mathcal{F}$ connue. Toute étude de sensibilité, par exemple conduite grâce aux indices de Sobol, devrait donc indiquer que la principale source d'incertitude expliquant les variations de Y^* est X et non ε. En pratique, ce type de diagnostic est établi *a posteriori*, comme vérification que la résolution numérique du problème d'inversion présente de bonnes propriétés. Cependant, cette condition doit agir comme une contrainte *a priori*, en particulier dans les cas où l'inversion stochastique est conduite dans un cadre bayésien, sur les paramètres de la solution $\hat{\mathcal{F}}$. Ceux-ci doivent être contraints par les caractéristiques de g et f_ε. Ce besoin a notamment été identifié dans (Fu *et al.*, 2015). Cela demande de pouvoir spécifier formellement ce que signifie "la principale source d'incertitude". L'information de Fisher sur Y a été proposée comme moyen d'action.

2 Une notion intuitive de "problème bien posé" au sens de Sobol

Par le prisme des outils classiques de l'analyse de sensibilité, le raisonnement exprimé dans l'introduction peut être formalisé comme suit :

Definition 1 *Soit (S_X, S_ε) les indices de Sobol du premier ordre, qui quantifient respectivement la majeure partie de l'incertitude sur Y^* expliquée par X et ε. Le problème (3) peut être dit bien posé au sens de Sobol si*

$$S_X \;>\; S_\varepsilon. \tag{4}$$

On en déduit le résultat suivant, qui fournit effectivement une contrainte sur les paramètres de $\mathcal{F}$.

Proposition 1 *Dans (2),supposons que g est différentiable dans le voisinage de $\mathbb{E}(X) := (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_p))$. Supposons que $X \sim \mathcal{N}(\mu, \Gamma)$, $\varepsilon \in \mathbb{R}^p \sim \mathcal{N}(0, \sigma^2 I_p)$ et notons $Dg_{\mathbb{E}(X)} := (\frac{\partial g}{\partial x_1}(\mathbb{E}(X_1)), \ldots \frac{\partial g}{\partial x_p}(\mathbb{E}(X_p)))$. Alors le problème d'inversion stochastique (2-3) est bien posé au sens de Sobol si et seulement si*

$$Dg^T_{\mathbb{E}(X)} \Gamma Dg_{\mathbb{E}(X)} \;>\; \sigma^2. \tag{5}$$

Très clairement, selon ce type raisonnement il existe alors autant de problèmes bien posés que d'indices de sensibilité, et de nombreux résultats similaires à (5) peuvent être produits. Une vision plus générique est exprimée dans la section suivante.

3 Une règle générique : problème bien posé au sens de Fisher

Décrire comment l'incertitude de X est transmise à Y peut être réalisé en considérant la dégradation de quantités d'information, comme l'information de Fisher (en ce sens qu'un gain d'information correspond à une incertitude éliminée du problème). Nous nous plaçons dans un cadre paramétrique où cette information est bien définie : $X \sim \mathcal{F} \equiv \mathcal{F}_\theta$ où θ vit dans un espace de dimension finie.

Notons alors $I_{g(X)}(\theta)$ et $I_{Y^*}(\theta)$ les informations de Fisher apportées respectivement par $g(X)$ and Y^* sur θ. Puisque l'impact de ε est de dégrader l'information, il est alors postulé que

$$I_{g(X)}(\theta) \quad > \quad I_{Y^*}(\theta) \tag{6}$$

où $A > B$, pour deux matrices carrées A et B, signifie que $A - B$ est une matrice définie positive. Supposer que la plupart de l'information sur θ dans Y^* est transmise par $g(X)$ implique que la différence entre $I_{g(X)}(\theta)$ et $I_{Y^*}(\theta)$, qui est une mesure de la perte d'information due au bruit ε, ne devrait pas être plus grande qu'une fraction $(1 - 1/c)I_{g(X)}(\theta)$ où $c > 1$. Il s'ensuit la règle

$$I_{g(X)}(\theta) > I_{Y^*}(\theta) > \frac{1}{c} I_{g(X)}(\theta). \tag{7}$$

Choisir $c = 2$ apparaît intuitif, mais des développements théoriques permettent de guider formellement le choix d'une valeur.

4 Application aux problèmes linéaires

L'obtention de conditions nécessaires et suffisantes pour s'assurer de (6) et (7) permet de dégager des contraintes utiles pour la modélisation $\mathcal{F}$. Considérons la situation classique où :

— g est réduite à un opérateur linéaire $H \in \mathbb{R}^{q \times p}$ de rang plein et $q \leq p$;
— $X \in \mathbb{R}^p \sim \mathcal{N}(\mu, \tau^2 I_p)$, où $\mu \in \mathbb{R}^p$;
— $\varepsilon \in \mathbb{R}^q \sim \mathcal{N}(0, \Sigma)$.

Proposition 2 *Une condition nécessaire pour (7) est*

$$(\sqrt{c} - 1)\tau^2 > \frac{1}{\underset{1 \leq i \leq q}{max} \{\lambda_i^\Psi\}}, \tag{8}$$

et une condition suffisante pour (7) est

$$(\sqrt{c}-1)\tau^2 > \frac{1}{\underset{1\leq i\leq q}{min}\{\lambda_i^\Psi\}}, \tag{9}$$

où $\left\{\lambda_i^\Psi\right\}_{1\leq i\leq q}$ *est le vecteur des valeurs propres de* $\Psi := \Sigma^{-1/2}HH^T\Sigma^{-1/2}$.

Des conditions suffisantes plus raffinées permettent d'établir un lien fort entre (7) et les contraintes qualitatives classiques sur le *nombre de condition*, qui permettent de postuler que l'interprétation donnée ici à la notion de problème bien posé est une réécriture partielle de celle de Hadamard, mais avec un sens plus immédiatement interprétable pour les spécialistes du traitement des incertitudes. Le résultat (8) permet en outre de déterminer une borne minimale de 4 pour c. Si le choix $c = 4$ est fait, alors les conditions de Fisher et de Sobol sont équivalentes.

Des résultats plus généraux peuvent être produits dans des situations où la covariance de X est anisotrope. Lorsque g est étendu aux modèles linéarisables, le point de linéarisation peut être défini comme la solution d'un problème de minimisation d'un écart de quantité d'information, sous une contrainte permettant de conserver le caractère bien posé du modèle linéaire approximant le vrai modèle. L'ensemble de ces résultats peut être perçu comme un premier pas sur l'explication des liens entre analyse de sensibilité et contraintes *a priori*, et leur utilisation pratique en modélisation.

Références

P. Barbillon, G. Celeux, A. Grimaud, Y. Lefebvre et E. Rocquigny (de) : Non linear methods for inverse statistical problems. 55:132–142, 2011.

D.A. Belsley, E. Kuh et R.E. Welsch : *"The Condition Number". In : Regression Diagnostics : Identifying Influential Data and Sources of Collinearity.* New York : John Wiley & Sons, 1980.

G. Celeux, A. Grimaud, Y. Lefebvre et E. Rocquigny (de) : Identifying intrinsic variability in multivariate systems through linearised inverse methods. 18:401–415, 2010.

S. Fu, G. Celeux, N. Bousquet et M. Couplet : Bayesian inference for inverse problems occurring in uncertainty analysis. 5:73–98, 2015.

G. Gamblin, M. Keller, P. Barbillon, A. Pasanisi et E. Parent : Adaptive numerical designs for the calibration of computer codes. 2015.

MC. Kennedy et A. O'Hagan : Bayesian calibration of computer models. 63:425–464, 2001.