Modélisation d'un code numérique par un processus gaussien, application au calcul d'une courbe de probabilité de dépasser un seuil

Séverine Demeyer, Frédéric Jenson, Nicolas Dominguez

CEA, LIST, F-91191 Gif-sur-Yvette, France, frederic.jenson@cea.fr

Résumé.
La modélisation statistique d'un code numérique par processus gaussien permet de définir un cadre bayésien d'analyse d'un code numérique. Dans l'objectif de la propagation des incertitudes, le couplage du processus gaussien avec un plan d'expériences numérique permet de prendre en compte des relations complexes (corrélations linéaires, non linéarité,...) entre les variables, à partir d'un nombre d'appels au code limité, afin d'évaluer un indicateur en sortie du code. Cette démarche est ici adaptée au domaine du Contrôle Non Destructif (CND) pour lequel elle constitue une méthode efficace et une avancée conceptuelle de traitement des incertitudes. Dans un premier temps on présente les enjeux relatifs à une modélisation statistique en CND dans le but d'obtenir des courbes de probabilité de détection de défauts. Puis on présente une méthode d'estimation des processus gaussiens par échantillonnage de Gibbs permettant une construction originale de ces courbes *a posteriori*. Enfin la démarche complète est illustrée sur le cas d'une inspection d'une plaque de titane par une méthode d'inspection par courants de Foucault.

Mots-clés. Code numérique, propagation des incertitudes, plan d'expériences numériques, processus gaussien, contrôle non destructif, probabilités de détection

Abstract.
Statistical processing of numerical experiments with Gaussian process yields a Bayesian framework for the analysis of numerical codes. Pertaining to uncertainty propagation, the combination of numerical designs of experiments with Gaussian process allows to take into account complex relationships to establish output indicators. This methodology is here adapted to the Non Destructive Testing (NDT) field where it proves to be both an effective method and a conceptual advance for the management of uncertainty. In a first time, issues pertaining to the statistical analysis of NDT data are reviewed and the output indicator (curve of probability of detection) is presented. In a second time a Gibbs sampling algorithm is derived to sample from the posterior distributions of parameters, which gives an innovative method to compute posterior probability curves. Finally the full methodology is applied to the inspection of a titanium flat area with Eddy currents.

Keywords. Computer experiments, uncertainty propagation, numerical design, Gaussian process, Gibbs algorithm, non destructive testing, probability of detection

1 Contexte: enjeux de la modélisation statistique en Contrôle Non Destructif (CND)

La qualification d'une méthode d'inspection en CND repose sur l'évaluation de sa capacité à détecter des défauts effectivement présents dans la pièce inspectée. L'un des indicateurs de sa capacité de détection est une courbe, dite courbe POD, représentant la probabilité de détection d'un défaut (en tant que probabilité qu'un signal issu d'une méthode d'inspection dépasse le seuil de détection) en fonction d'un paramètre critique caractérisant le défaut contrôlé, compte tenu des incertitudes sur les conditions de l'inspection. Traditionnellement les données permettant de construire les courbes POD empiriques sont obtenues à partir de campagnes expérimentales longues et coûteuses réalisées selon un référentiel normatif qui vise à garantir l'exploitation statistique des résultats. Pour ces raisons, on cherche désormais à obtenir des courbes POD numériques à partir de données d'inspection simulées en sortie d'un code numérique complexe modélisant l'inspection. Les courbes POD numériques sont actuellement obtenues sous une hypothèse de linéarité entre la sortie du code, obtenue par propagation des incertitudes, et le paramètre critique en entrée, par l'utilisation de méthodes par maximum de vraisemblance implémentées dans l'article référence de Berens [1]. Ces méthodes, initialement développées pour traiter les données issues d'inspections réelles ont pour limite de ne pas exploiter de modélisation statistique du code numérique notamment en terme de prédiction et d'optimisation du nombre d'appels au code.

La méthodologie développée et recommandée dans ce papier pour l'obtention d'une POD numérique repose sur les quatre étapes suivantes 1) simulation de points d'entrée du code selon un plan d'expérience numérique, 2) propagation des incertitudes dans le code numérique, 3) modélisation de la réponse du système par un processus gaussien (choix des fonctions d'espérance et de corrélation) (section 2), 4) construction point par point de la courbe POD (section 3).

2 Cadre bayésien de l'analyse d'un code numérique par processus gaussien

Soit d le nombre de variables en entrée d'un code numérique déterministe. On suppose que la sortie du code est observée aux points d'entrée $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ d'un sous-ensemble $\mathcal{X}$ de $\mathbf{R}^d$, obtenus par simulation à partir d'un plan d'expériences numériques de taille n, et est inconnue en dehors de ces points. On note $\boldsymbol{y}_n = (y_1, \ldots, y_n)$ les sorties (supposées unidimensionnelles) du code qui leur sont associées. On constitue ainsi la base de données $\mathcal{D} = \{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ servant à l'apprentissage des paramètres du modèle.

La modélisation de la sortie d'un code numérique, vue comme une fonction inconnue η des variables d'entrée, par un processus gaussien suppose que la sortie du code, $Y(.)$, est la réalisation d'un processus aléatoire gaussien η (voir Santner et al. [2]), c'est-à-dire

- $y_i = \eta(\boldsymbol{x}_i)$ pour $i = 1, \ldots, n$
- pour tout $L \geq 1$ et tous $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$ de $\mathcal{X}$ le vecteur $(\eta(\boldsymbol{x}_1), \ldots, \eta(\boldsymbol{x}_L))$ suit une distribution normale multivariée qui modélise l'incertitude sur les sorties $(\eta(\boldsymbol{x}_1), \ldots, \eta(\boldsymbol{x}_L))$.

Un processus gaussien est ainsi entièrement déterminé par sa fonction d'espérance $E(\eta(\boldsymbol{x}))$ et sa fonction de covariance $Cov\left(\eta(\boldsymbol{x}), \eta(\boldsymbol{x}')\right)$. Ces deux grandeurs, vues comme des fonctions de la fonction aléatoire η, sont également aléatoires.

Le modèle de processus gaussien s'écrit

$$\eta(\boldsymbol{x}_1), \ldots, \eta(\boldsymbol{x}_n) | \boldsymbol{\beta}, \sigma^2, \mathcal{D} \sim \mathrm{N}\left(\boldsymbol{F}_n \boldsymbol{\beta}, \sigma^2 \boldsymbol{R}_n\right) \tag{1}$$

où $\boldsymbol{F}_n = (f_j(\boldsymbol{x}_i))$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ la matrice $n \times p$ connue de fonctions de régression pour les données d'apprentissage, $\boldsymbol{\beta}$ est le vecteur $p \times 1$ des coefficients de régression, σ_Z^2 est la variance du processus gaussien, $\boldsymbol{R}_n = (\boldsymbol{R}_{i,j}) = (R(\boldsymbol{x}_i - \boldsymbol{x}_j))$ est la matrice $n \times n$ des corrélations entre les données d'apprentissage où R est une fonction de corrélation.

La matrice $\boldsymbol{F}_n$ permet de spécifier des relations non linéaires.

On observe que $E(\boldsymbol{y}_n | \boldsymbol{\beta}) = \boldsymbol{F}_n \boldsymbol{\beta}$ et $Cov\left(\boldsymbol{y}_i, \boldsymbol{y}_j\right) | \sigma^2 = \sigma^2 \boldsymbol{R}_{i,j}$.

La spécification de distributions *a priori* sur les paramètres de régression $\boldsymbol{\beta}$ et σ^2 permet de définir une distribution *a priori* pour la fonction $Y(.)$.

D'après (1), on choisit les distributions *a priori* conjuguées suivantes

$$\boldsymbol{\beta} | \sigma^2 \sim \mathrm{N}\left(\boldsymbol{\beta_0}, \sigma^2 \boldsymbol{V}_0\right) \tag{2}$$

$$\sigma^2 \sim \mathrm{Inv} - \mathrm{Chi}^2\left(\nu_0, s_0^2\right) \tag{3}$$

L'estimation des paramètres du processus gaussien (1) est réalisée sous les distributions *a priori* (2) et (3) par l'implémentation d'un algorithme de Gibbs alternant les tirages dans les distributions conditionnelles *a posteriori* suivantes

$$\boldsymbol{\beta} | \mathcal{D}, \sigma^2 \sim \mathrm{N}\left(\left(\boldsymbol{F_n}^t \boldsymbol{R_n}^{-1} \boldsymbol{F_n} + \boldsymbol{V_0}^{-1}\right)^{-1} \left(\boldsymbol{F_n}^t \boldsymbol{R_n}^{-1} \boldsymbol{y_n} + \boldsymbol{V_0}^{-1} \boldsymbol{\beta}_0\right), \sigma^2 \left(\boldsymbol{F_n}^t \boldsymbol{R_n}^{-1} \boldsymbol{F_n} + \boldsymbol{V_0}^{-1}\right)^{-1}\right) \tag{4}$$

$$\sigma^2 | \mathcal{D}, \boldsymbol{\beta} \sim \mathrm{Inv} - \mathrm{Gamma}\left(\frac{\nu_0}{2} + \frac{n}{2}, \frac{\nu_0 s_0^2}{2} + \frac{1}{2}\left[\left(\boldsymbol{y_n} - \boldsymbol{F_n}\boldsymbol{\beta}\right)^t R_n^{-1} \left(\boldsymbol{y_n} - \boldsymbol{F_n}\boldsymbol{\beta}\right)\right]\right) \tag{5}$$

3 Calcul des courbes de probabilité de détection

En notant X_1 le paramètre critique d'un défaut vu comme la première composante du vecteur $\boldsymbol{x}$ en entrée du code numérique, la courbe POD moyenne est définie par

$$\begin{aligned} POD\left(x_1\right) &= P\left(\eta(\boldsymbol{x}) > s | X_1 = x_1, \mathcal{D}\right), \boldsymbol{x} \sim G\left(\boldsymbol{x}\right) \\ &= \int P\left(\eta(x_1, \boldsymbol{x_j^{'}}) > s | \mathcal{D}\right) f_{X_2,\ldots,X_d|X_1=x_1}\left(\boldsymbol{x_j^{'}} | x_1\right) \mathrm{d}\boldsymbol{x_j^{'}} \end{aligned} \tag{6}$$

où $G\left(x_1, \boldsymbol{x_j^{'}}\right) = f_{X_2,\ldots,X_d|X_1=x_1}\left(\boldsymbol{x_j^{'}} | x_1\right) \times f_{X_1}\left(x_1\right)$ est la distribution jointe des variables d'entrée $X_1, \ldots, X_d$ et la notation $\eta\left(.\right)|\mathcal{D}$ renvoie à la distribution prédictive *a posteriori* de η. Cette intégrale peut être approchée par

$$POD\left(x_1\right) \approx \frac{1}{n^{'}} \sum_{j=1}^{n^{'}} P\left(\eta(x_1, \boldsymbol{x}_j^{'}) > s | \mathcal{D}\right), \boldsymbol{x}_j^{'} \in \mathcal{D}_{\boldsymbol{x}_1}^{'} \tag{7}$$

où $\mathcal{D}_{\boldsymbol{x}_1}^{'}$ est un plan d'expériences LHS vu comme un pseudo-échantillon de la loi conditionnelle $X_2, \ldots, X_d | X_1 = x_1$ permettant l'intégration Monte Carlo accélérée de (6).

Puisque η est une fonction aléatoire, $POD\left(x_1\right)$ est une variable aléatoire et on évalue en fait la distribution de la probabilité (6) à partir de réalisations $\eta_{(i)}\left(.\right)|\mathcal{D}$ de $\eta\left(.\right)|\mathcal{D}$ obtenues à partir de réalisations $\left(\boldsymbol{\beta}_{(i)}, \sigma_{(i)}^2\right)$ de la loi jointe *a posteriori* de $(\boldsymbol{\beta}, \sigma^2)$, à savoir

$$\eta_{(i)}\left(\boldsymbol{x}\right)|\mathcal{D}, \boldsymbol{\beta}_{(i)}, \sigma_{(i)}^2 \sim \mathrm{N}_1\left(f_{\boldsymbol{x}}^t \boldsymbol{\beta}_{(i)} + r_{\boldsymbol{x}}^t \boldsymbol{R}_n^{-1}\left(\boldsymbol{y_n} - \boldsymbol{F}_n \boldsymbol{\beta}_{(i)}\right), \sigma_{(i)}^2\left(1 - r_{\boldsymbol{x}}^t \boldsymbol{R}_n^{-1} r_{\boldsymbol{x}}\right)\right) \tag{8}$$

où $f_{\boldsymbol{x}}^t = \left(f_j(\boldsymbol{x})\right), j = 1, \ldots, p$ et $r_{\boldsymbol{x}} = \left(R(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, R(\boldsymbol{x}, \boldsymbol{x}_n)\right)$.

Ainsi, à $\boldsymbol{x}_j^{'}$ et $\eta_{(i)}$ donnés, on obtient la réalisation de la variable aléatoire $POD\left(x_1\right)$

$$POD_{(i,j)}\left(x_1\right) = 1 - \Phi\left(\frac{s - f_{\left(x_1, \boldsymbol{x_j^{'}}\right)}^t \boldsymbol{\beta}^{(i)} - r_{\left(x_1, \boldsymbol{x_j^{'}}\right)}^t \boldsymbol{R}_n^{-1}\left(\boldsymbol{y}_n - \boldsymbol{F}_n \boldsymbol{\beta}^{(i)}\right)}{\sigma_{(i)} \sqrt{\left(1 - r_{\left(x_1, \boldsymbol{x_j^{'}}\right)}^t \boldsymbol{R}_n^{-1} r_{\left(x_1, \boldsymbol{x_j^{'}}\right)}\right)}}\right) \tag{9}$$

La courbe POD moyenne est calculée point par point comme la courbe moyenne de l'ensemble des points $POD_{(i,j)}\left(x_1\right), i = 1 \ldots N, j = 1, \ldots, n^{'}, x_1 \sim f_{X_1}$. On définit la courbe de confiance inférieure à 5% par la courbe quantile à 5%.

4 Cas d'application en CND

On cherche à établir la courbe de probabilité de détection d'une fissure dans une plaque de titane inspectée par une méthode par courants de Foucault [3], dont le paramètre critique est sa longueur (X_1) compte-tenu de trois sources d'incertitude: sa hauteur (X_2), la position initiale de la sonde (X_3) et la hauteur d'entrefer (distance entre la sonde et la plaque à inspecter) (X_4) sous les hypothèses suivantes reposant sur des avis d'experts :

- $X_1 \sim \text{Unif}\,(0.05, 5)$ telle que l'étendue de la distribution comprenne la gamme critique de longueurs d'entailles (en mm),
- $X_2 = 0.5X_1 + 0.1X_1\text{N}\,(0,1)$ qui traduit l'expertise que la hauteur moyenne de l'entaille (en mm) est égale à la moitié de sa longueur avec une incertitude égale à 20% de sa hauteur moyenne,
- $X_3 \sim \text{Unif}\,(12.5 - 0.05, 12.5 + 0.05)$ où 12.5 mm est la valeur théorique de la position initiale de la sonde se déplaçant avec un pas de 0.1 mm générant une incertitude de 0.05 mm sur la position de départ,
- $X_4 \sim 0.08 + 0.8\sin(\alpha)\,,\ \alpha \sim \text{N}\left(0, \frac{\pi}{180}\right)$ où 0.08 mm est la valeur théorique de l'entrefer.

Un plan de type LHS (hypercube latin) de $n = 50$ expériences vérifiant les hypothèses ci-avant a été obtenu en appliquant les fonctions quantiles adéquates aux points du plan généré par la fonction optimumLHS du logiciel statistique R [4].

La propagation des incertitudes a été réalisée par appel au code numérique CIVA développé au sein des équipes du CEA/LIST (http://www-civa.cea.fr). La sortie relevée du code est le maximum d'amplitude (en mV) de la partie imaginaire du signal. Le seuil de détection a été fixé à 44 mV.

La représentation graphique des sorties en fonction de X_1 montre une log-linéarité. On modélise alors $(\log(y_1), \ldots, \log(y_n))$ par rapport à $F_n = \left(1\log(x_{1[1]}),\ \ldots,\ 1\log(x_{n[1]})\right)^t$ avec la fonction de corrélation $R\,(\boldsymbol{x}_i - \boldsymbol{x}_j) = \exp\,(-1.5\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2)$.

L'analyse bayésienne est réalisée sous les distributions *a priori* conjuguées suivantes

$$\boldsymbol{\beta}|\sigma^2 \sim N_2\left((1,1)^t, \sigma^2\boldsymbol{Id}_2\right) \tag{10}$$

$$\sigma^2 \sim \text{Inv} - \text{Chi}^2\left(6, \frac{2}{3}\right) \tag{11}$$

où $\boldsymbol{Id}_2$ est la matrice identité 2×2 et les paramètres de la distribution Inverse-Chi2 modélisent la croyance *a priori* d'une valeur centrale de 1 avec un écart-type de 1.

Les courbes POD (moyenne et confiance à 5%) obtenues suivant la méthodologie des sections 2 et 3 sont illustrées à la figure 1.

A partir de la courbe POD moyenne, on déduit que la méthode d'inspection a 90% de chances de détecter un défaut de taille 1.4 mm, dite valeur a_{90}. La courbe quantile à 5% fournit la valeur plus conservative de 1.7 mm, dite valeur $a_{90/95}$, pour laquelle on a 95% de chances que la probabilité de détection soit d'au moins 90%.

5 Conclusion

Une méthodologie complète reposant sur la propagation des incertitudes a été proposée dans un cadre bayésien pour apporter une réponse innovante et flexible pour le traitement

de données de Contrôle Non Destructif dans le but d'évaluer des courbes de probabilité de dépassement d'un seuil. Cette approche repose sur l'utilisation directe des chaînes de Markov générées par l'algorithme de Gibbs, se généralise à l'évaluation de courbes plus complexes et permet des analyses plus complexe du code. Ainsi le cadre théorique défini a pour perspective d'être étendu à l'étude d'une problématique cruciale en CND qui est la calibration des distributions incertaines en entrée du code à partir de données CND réelles et simulées.

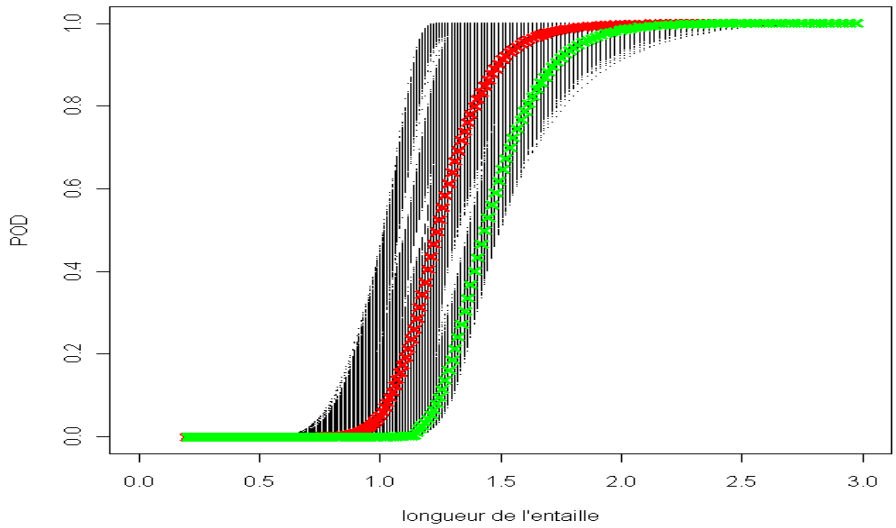

Figure 1: Représentation graphique de la variable aléatoire $POD_{(i,j)}(x_1)$ d'après l'équation (9) (points noirs) pour chaque valeur de X_1 du plan d'expérience $\mathcal{D}$ et construction point par point des courbes POD quantile à 50% (en rouge) et 5% (en vert).

Bibliographie

[1] Berens, A.P. (1988), NDE Reliability Data Analysis, ASM Metals Handbook, volume 17, 9th edition: Nondestructive Evaluation and Quality Control, ASM International, Materials Park, Ohio, 689–701.

[2] Santner, T. J. , Williams, B.J., Notz, W.I. (2003), *The Design and analysis of computer experiments*, Springer Verlag, Springer Series in Statistics.

[3] Jenson, F., Iakovleva, E., Dominguez, N. (2010), *Simulation supported POD : methodology and HFET validation case*, Review of Progress in QNDE, 30.

[4] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.