

REDEFINITION DE LA COURBE DE PROBABILITE DE DETECTION (POD) COMME FONCTION DE REPARTITION ALEATOIRE

Thomas BROWNE^{1,2,3} & Jean-Claude FORT^{2,3}

1-EDF R&D, Chatou, France ; 2-Université Paris Descartes, Paris, France ; 3-Institut de
Mathématiques de Toulouse, Toulouse, France.

thomas.browne@edf.fr

jean-claude.fort@parisdescartes.fr

Résumé. Les probabilités de détection de défauts (POD) sont des outils standards dans l'industrie pour évaluer la performance d'un procédé de contrôle donné. Elles se modélisent mathématiquement par une courbe déterministe qui est une fonction de la taille du défaut étudié. Les aléas pris en compte pour le calcul de la probabilité sont les propriétés géométriques de la structure étudiée ($X \in \mathbb{R}^d$) ainsi qu'un bruit d'observation δ . La courbe de POD possède les caractéristiques d'une fonction de répartition. Le but est ici de faire évoluer ce modèle en une fonction de répartition aléatoire π_X . Pour chaque réalisation x des paramètres X , π_X lui associe la POD correspondante, *i.e.* la courbe de probabilités calculées seulement sur δ sous la condition $X = x$. Pour quantifier une distribution aléatoire de f.d.r. il est nécessaire de définir un nouveau cadre de travail. On cherche à répondre aux questions suivantes : comment comparer les possibles réalisations ? Peut-on par exemple parler de courbe moyenne ? A l'aide de la distance de Wasserstein ainsi que la notion de moyenne de Fréchet on propose des définitions de POD-moyenne, de POD-quantiles par analogie avec les notions d'espérance et de quantiles de variables aléatoires réelles. On s'intéresse également à la définition d'estimateurs pour ces grandeurs ainsi qu'à leurs propriétés asymptotiques. Pour des besoins industriels il est important de distinguer les variables de X les plus influentes. Dans ce but on propose une extension des méthodes d'analyse de sensibilité de type Sobol.

Mots-clés. POD, distance de Wasserstein, moyenne de Fréchet, analyse de sensibilité

Abstract. Probabilities of detection are standard tools to evaluate the performance of a control process. They are modeled by a deterministic curve that is a function of the size of the flaw. The probability is computed for a size a by considering both the randomness of the structure's geometrical properties ($X \in \mathbb{R}^d$) and the observation noise δ . The POD curve is similar to a cumulative distribution function. Our goal is to change this model into a random cdf π_X . To each realization x of X , π_X maps the matching POD, *i.e.* the POD curve computed on δ under the condition $X = x$. Then we must set a new framework in order to quantify a cdf random distribution. We ought to answer the

following questions : how to compare the different realizations? Is it possible to define an average curve for instance? By using the Wasserstein distance and the Fréchet-mean we offer definitions for curves such as a POD-mean, POD-quantiles, which are based on extensions of the mean and quantiles for real random variables. We also focus on the definition of estimators for these curves and study their asymptotical properties.

Keywords. POD, Wasserstein distance, Fréchet mean, sensitivity analysis

1 Redéfinition de la POD (Probability of detection)

Il est courant dans l'industrie de réaliser des contrôles non destructifs par courants de Foucault ou ultrasons afin d'assurer l'intégrité d'une structure étudiée [1]. Les courbes de probabilité de détection sont un outil standard pour évaluer la performance des examens non destructifs (NDT : Non Destructive Tests) [2]. Le cadre de travail est présenté ci-dessous :

- $Y \in \mathbb{R}$: mesure du signal obtenu après un NDT sur une structure à étudier.
- $a > 0$: longueur du défaut (mm) - paramètre d'intérêt.
- $X \in \mathbb{R}^d$: tous les autres paramètres influents sur la sortie Y , qui dépendent soit de la structure, tels que sa perméabilité et sa conductivité, soit du contrôle NDT même. On le considère comme un vecteur aléatoire dont les lois marginales sont indépendantes et de densité de probabilité : $(X_1, \dots, X_d) \sim (f_1, \dots, f_d)$.
- $\delta \in \mathbb{R}$: le bruit d'observation sur Y dû à la nature de l'expérience et aux paramètres non maîtrisés de la structure (sa microstructure par exemple), dont la pdf (resp. cdf) est f_δ (resp. F_δ). Lors d'un modèle de régression, δ inclut également l'erreur du modèle.
- t_s : le seuil du signal à partir duquel on considère la présence d'un défaut, *i.e.* $Y(a, X, \delta) > t_s$ implique la détection de défaut.

Comme l'environnement est aléatoire, il est évident que pour deux défauts de même taille a , il est possible que seulement l'un des deux soit détecté : c'est la raison pour laquelle on parle de probabilité de détection. Dans ce contexte, la POD est une courbe dont l'expression est donnée par :

$$\forall a > 0 \quad \text{POD}(a) = \mathbb{P}(Y(a, X, \delta) > t_s)$$

Supposons que Y est une fonction croissante de δ [3]. Alors si on appelle $Y_{a,X}(\cdot)$ la fonction qui évalue le signal pour tout δ sachant a et X , on peut considérer sa fonction réciproque $Y_{a,X}^{-1}(\cdot)$:

$$\forall a > 0 \quad \text{POD}(a) = \mathbb{P}(\delta > Y_{a,X}^{-1}(t_s)),$$

ce qui signifie que la POD peut elle-même être vue comme une cdf. Selon ce résultat il est clair que la distribution de δ , *a priori* inconnue, est un point important du cadre de travail. Un enjeu majeur de cette thèse est d'élaborer un outil pour évaluer l'influence des

paramètres X sur les valeurs de la POD. Jusqu'ici les aléas sur X et δ sont pris au même niveau. Désormais on change légèrement le cadre de travail : considérons maintenant la fonction de répartition aléatoire π_X définie comme suit

$$\begin{aligned} \forall a > 0 \quad \pi_X(a) &= \mathbb{P}(Y(a, X, \delta) > t_s \mid X) \\ &= 1 - F_\delta(Y_{a,X}^{-1}(t_s)) \quad \text{sachant } X \end{aligned}$$

ce qui revient à dire que pour chaque réalisation $x \in \mathbb{R}^d$ de X on a une cdf correspondante qui se trouve être la POD sachant les paramètres x . Le but est maintenant de déterminer la "distribution" de cette cdf aléatoire à partir d'une collection de ses réalisations $(\pi_x)_x$. On en vient à définir un nouveau cadre de travail afin de comparer des cdf en utilisant la 2-distance de Wasserstein, qui s'exprime pour deux cdf F et G :

$$W_2^2(F, G) = \int_0^1 (F^{-1}(u) - G^{-1}(u))^2 du$$

où F^{-1} et G^{-1} sont des fonctions quantiles. Des outils peuvent être ainsi définis à partir de cette distance afin de quantifier π_X tels qu'une moyenne (moyenne de Fréchet), une variance ou des quantiles. Leur définition est en fait une extension du domaine des variables aléatoires réelles (cf figure 1).

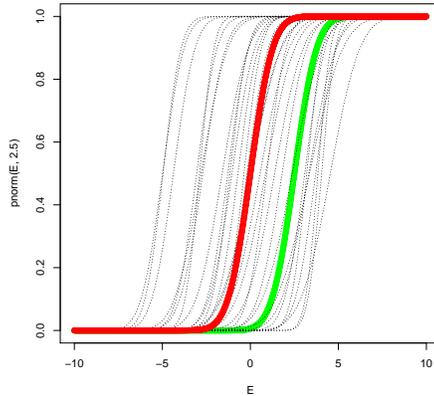


FIGURE 1 – En pointillés gris 30 réalisations de π_X , en lignes larges sa moyenne (en noir) et son 75%-quantile (en gris). On peut remarquer que la POD-moyenne et le POD-quantile sont aussi des cdf.

2 Moyenne de POD (POD-moyenne)

On définit la POD-moyenne comme la moyenne de Fréchet (par la 2-distance de Wasserstein) de la fonction de répartition aléatoire π_X . Dans le cadre des variables aléatoires

réelles, l'espérance d'une v.a.r. X ($\mathbb{E}[X]$) peut être vue comme le réel qui minimise le contraste :

$$\forall \theta \in \mathbb{R} \quad \varphi_X(\theta) = \mathbb{E}[|X - \theta|^2]. \quad (1)$$

Ce qui fournit immédiatement une définition de la variance dans ce cadre :

$$\mathbb{V}[X] = \varphi_X(\mathbb{E}[X]). \quad (2)$$

La distance considérée alors est l'écart quadratique classique. Dans le cas d'une cdf aléatoire, remplaçons la valeur absolue par la 2-distance de Wasserstein dans l'ensemble \mathcal{F}^2 des cdf des lois de carré intégrable. Un nouveau contraste peut alors être défini :

$$\begin{aligned} \forall F \in \mathcal{F} \quad \phi_{\pi_X}(F) = \mathbb{E}[W_2^2(F, \pi_X)] &= \int_0^1 \mathbb{E}[(\pi_X^{-1}(u) - F^{-1}(u))^2] du \\ &= \int_0^1 \varphi_{\pi_X^{-1}(u)}(F^{-1}(u)) du. \end{aligned} \quad (3)$$

Cela nous ramène à la définition de la moyenne de Fréchet de la POD pour la distance de Wasserstein, que l'on appelle la POD-moyenne [3] C'est la cdf qui minimise le contraste précédent ϕ_{π_X} . Il est évident que selon (3), la POD-moyenne, notée $\mathcal{M}(\pi_X)$, est une cdf dont la fonction réciproque vérifie en tout point :

$$\forall u \in [0, 1] \quad \mathcal{M}(\pi_X)^{-1}(u) = \arg \min_{\theta \in \mathbb{R}} \varphi_{\pi_X^{-1}(u)}(\theta). \quad (4)$$

On peut alors conclure que $\mathcal{M}(\pi_X)$ est définie comme suit :

$$\forall u \in [0, 1] \quad \mathcal{M}(\pi_X)^{-1}(u) = \mathbb{E}[(\pi_X)^{-1}(u)], \quad (5)$$

et d'après (2) on définit une variance pour la POD

$$\begin{aligned} \mathcal{V}(\pi_X) &= \phi_{\pi_X}(\mathcal{M}(\pi_X)) \\ &= \int_0^1 \mathbb{V}[\pi_X^{-1}(u)] du. \end{aligned} \quad (6)$$

Exemple : supposons que la mesure du signal Y ait l'expression suivante

$$Y(a, X, \delta) = a + X + \delta,$$

avec $a \in [0, 10]$, $\delta \sim \mathcal{N}(0, 1)$ iid, X le paramètre unidimensionnel tel que $X \sim \mathcal{U}(-6, -2)$. Nous considérons un seuil $t_s = 1mV$. Ainsi l'expression de la POD est donnée par :

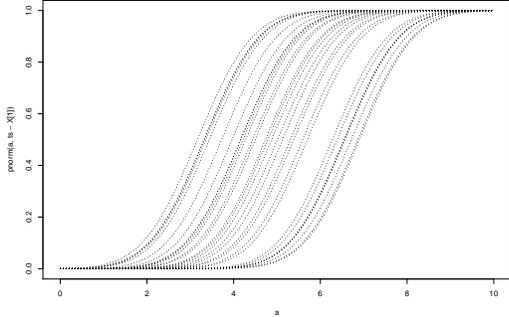
$$\begin{aligned} \forall a \in [0, 10] \quad \pi_X(a) &= \mathbb{P}(a + X + \delta > t_s \mid X) \\ &= \mathbb{P}(\delta < a + X - t_s \mid X) \quad \text{car } \delta \text{ est symétrique} \\ &= F_{\mathcal{N}(t_s - X, 1)}(a). \end{aligned} \quad (7)$$

où $F_{\mathcal{N}(0,1)}$ est la cdf de $\mathcal{N}(0,1)$. On peut conclure que : $\forall u \in [0, 1] \quad (\pi_X)^{-1}(u) = F_{\mathcal{N}(0,1)}^{-1}(u) + t_s - X$. En considérant (5) :

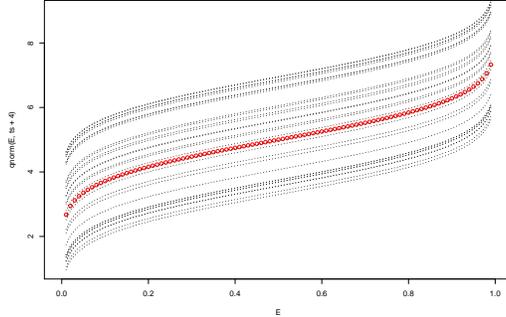
$$\begin{aligned} \forall u \in [0, 1] \quad \mathcal{M}(\pi_X)^{-1}(u) &= \mathbb{E}[(\pi_X)^{-1}(u)] \\ &= F_{\mathcal{N}(0,1)}^{-1}(u) + t_s - \mathbb{E}[X] \end{aligned} \quad (8)$$

La construction de la POD-mean est illustrée en 2.

N.B. : On peut aboutir à la POD-mean par la fonction réciproque de $\mathcal{M}(\pi_X)^{-1}$, qui est naturellement $F_{\mathcal{N}(5,1)}$. On peut remarquer que la moyenne de la POD, $F_{\mathcal{N}(t_s-X,1)}$, peut aussi être écrite $F_{\mathcal{N}(t_s-\mathbb{E}[X],1)}$. C'est une conséquence directe de la 2-distance de Wasserstein.



(a) 30 réalisations de π_X en pointillés gris



(b) construction de la POD-moyenne (en points noirs) par la POD réciproque parmi 30 réalisations de π_X^{-1} (en pointillés gris)

FIGURE 2 – A gauche sont illustrées 30 réalisations de la POD comme cdf aléatoire. A droite, la réciproque de la POD-mean (en rouge) est obtenue comme la moyenne en tout point $[0, 1]$ de la POD réciproque, dont les réalisations sont en pointillés gris)

3 Analyse de sensibilité sur la POD

L'analyse de sensibilité a pour but de quantifier l'influence d'une v.a.r. X_i , pour $i \in \{1, \dots, d\}$, sur une v.a.r. Y où $Y = f(X_1, \dots, X_d)$, avec f une fonction mesurable. Par exemple l'indice de Sobol non normalisé [4] défini par

$$S_{X_i}^Y = \mathbb{V}_{X_i}(\mathbb{E}[Y | X_i]) \quad (9)$$

évalue l'influence de X_i sur l'espérance de Y . Les indices définis à partir de contrastes [5] généralisent les indices de type Sobol pour toute grandeur de Y (quantiles, probabilité de dépassement de seuil *etc.*) tant que la grandeur d'intérêt minimise le contraste choisi.

Par exemple si l'on considère à nouveau le contraste φ_Y défini en (1) on peut réécrire $S_{X_i}^Y$ de la manière suivante :

$$\begin{aligned} S_{X_i}^Y &= \mathbb{V}(Y) - \mathbb{E}_{X_i}[\mathbb{V}(Y | X_i)] \\ &= \min_{\theta \in \mathbb{R}} \varphi_Y(\theta) - \mathbb{E}_{X_i}[\min_{\theta \in \mathbb{R}} \varphi_Y(\theta) | X_i], \end{aligned} \quad (10)$$

avec φ_Y qui est minimisé par $\mathbb{E}[Y]$. On met ainsi en évidence le fait que les indices de Sobol sont pertinents pour étudier l'influence des variables d'entrées sur la moyenne de la sortie Y . Notre but est d'étendre les indices définis par contrastes à la POD. On montre par la suite la construction d'un tel indice pour évaluer l'influence des variables d'entrées X sur la POD-moyenne $\mathcal{M}(\pi_X)$.

De manière analogue à (10), on définit l'indice $\mathcal{S}_{X_i}^{\pi_X}$ par :

$$\mathcal{S}_{X_i}^{\pi_X} = \min_{F \in \mathcal{F}^2} \phi_{\pi_X}(F) - \mathbb{E}_{X_i} \left[\left(\min_{F \in \mathcal{F}^2} \phi_{\pi_X}(F) \right) \mid X_i \right] \quad (11)$$

où ϕ_{π_X} est le contraste défini en (3) sur \mathcal{F}^2 minimisé par $\mathcal{M}(\pi_X)$. Le premier terme correspond à la variance de POD $\mathcal{V}(\pi_X)$. Il apparait rapidement pour le deuxième terme :

$$\mathbb{E}_{X_i} \left[\left(\min_{F \in \mathcal{F}^2} \phi_{\pi_X}(F) \right) \mid X_i \right] = \int_0^1 \mathbb{E}_{X_i} [\mathbb{V}(\pi_X^{-1}(u) \mid X_i)] du. \quad (12)$$

Ainsi par (6) et (12), l'influence de X_i sur $\mathcal{M}(\pi_X)$ a pour expression :

$$\mathcal{S}_{X_i}^{\pi_X} = \int_0^1 S_{X_i}^{\pi_X^{-1}(u)} du, \quad (13)$$

où $S_{X_i}^{\pi_X^{-1}(u)}$ est l'indice de Sobol qui quantifie l'influence de X_i sur l'espérance de la v.a.r. $\pi_X^{-1}(u)$ pour $u \in]0, 1[$.

Bibliographie

- [1] L. Maurice and V. Costan and E. Guillot and P. Thomas (2012), Eddy current NDE performance demonstrations using simulation tools, Review of Progress in Quantitative Non Destructive Evaluation, Denver, Colorado, USA, V.32 : 464-471.
- [2] L. Gandosi and C. Annis (2010), Probability of Detection Curves : Statistical Best-Practice, ENIQ TGR Technical Document, V.41.
- [3] K. Kanishcheva (2014), Statistical estimation of the POD in a numerical experiments context, Master's degree thesis, Airbus Group, Ecole Polytechnique, Paris.
- [4] I. M. Sobol (1993), Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Experiment, V.1 : 407-414.
- [5] J. C. Fort, T. Klein, N. Rachdi (2013), New sensitivity analysis subordinated to a contrast, Communication in Statistics : Theory and Methods, *In press*.