Une utilisation originale de l'analyse de sensibilité (à but orienté) pour aider à choisir son maillage en présence de sources d'incertitudes

Gaël Poëtte

Une utilisation originale de l'analyse de sensibilité (à but orienté) pour aider à choisir son maillage en présence de sources d'incertitudes

Gaël Poëtte

[a] *CEA DAM CESTA, F-33114 Le Barp, France,*
[b] *Institut de Mathématiques de Bordeaux, 351 Cours de la Libération 33405 Talence cedex*

Abstract

Ce manuscrit est un document de travail. Il est en cours de rédaction (un travail bibliographique plus poussé sur certains points reste encore nécessaire). Son objectif est de faciliter certains échanges. Il présente une application originale des outils récents d'analyse de sensibilité (à but orienté) de la litérature. Il met en particulier en commun les résultats de 5 ouvrages [4, 9, 11, 5, 7]. La raison de cette mise en commun est explicitée dans le document. Les indices ainsi construits permettent d'effectuer une sensibilité à but orienté (sur une probabilité de défaillance) relative aux paramètres incertains d'une équation aux dérivées partielles en prenant en compte l'erreur de discrétisation de celle-ci. Y est également abordée la question de pouvoir faire une analyse de sensibilité sur des groupes de paramètres (enjeu important dès lors que les paramètres incertains sont corrélés par exemple). La démarche est déroulée sur un exemple jouet simple, facilitant la compréhension de la démarche, l'interprétation des indices et la reproductibilité des résultats.

Keywords: Analyse de sensibilité à but orienté, HSIC, problème jouet, entrées corrélées, groupes de paramètres, document de travail

1. Introduction

Dans ce document, l'application d'outils d'analyse de sensibilité à but orienté est détaillée sur un problème jouet. L'idée est de dérouler une analyse de sensibilité à but orienté sur ce modèle et de constuire une méthodologie compatible avec des problèmes industriels réels mais dans des conditions plus faciles d'accès (code rapide, modèle connu, interprétations plus aisées etc.). L'objectif de ce document est

- de montrer sur un exemple simple tout le potentiel des outils mathématiques décrits dans [5, 9, 11, 7];

Email address: gael.poette@cea.fr, gael.poette@bordeaux-inp.fr (Gaël Poëtte)

- de démontrer, sur un exemple parfaitement connu, la vérification de l'implémentation mais également la fiabilité des outils pour la prise de décision;
- de documenter un exemple facilement reproductible (une sorte de format "travaux dirigés");
- de montrer que les outils présentés et utilisés[1] permettent de prendre en compte dans l'analyse de sensibilité le paramètre de discrétisation numérique (permettant de savoir si l'on est par exemple trop sensible au maillage pour interpréter les résultats ou non);
- de présenter une méthodologie pour effectuer une analyse de sensibilité sur des groupes de paramètres (la bibliographie reste notamment à enrichir à ce sujet).

Dans la section 2 qui suit, le modèle que nous étudierons tout au long de ce document est présenté.

2. Un problème jouet représentatif de quelques difficultés industrielles réelles

Nous présentons un modèle jouet dans les paragraphes qui suivent. Nous effectuons une propagation d'incertitudes et une analyse de sensibilité sur ce modèle dans la section 3. Nous nous intéressons à un modèle simple ayant la forme suivante:

$$Y(t,x) = Y(t,x_0,x_1,x_2,x_3) = \left(x_0 + \frac{S}{\lambda(x_1,x_2)}\right)\exp\left(\lambda(x_1,x_2)t\right) - \frac{S}{\lambda(x_1,x_2)} + tCx_3, \quad \text{avec } \lambda(x_1,x_2) = -10 - (x_1+x_2). \tag{1}$$

Ce modèle est choisi parce qu'il a une forme générique que l'on retrouve dans de nombreuses études d'intérêt

- il possède une dépendance temporelle (typique des problèmes instationnaires de dynamique rapide, cf. les vitesses de surface libre étudiées dans [6] par exemple),
- il dépend d'un vecteur de paramètres $x = (x_0, x_1, x_2, x_3)$ qui peuvent être physique ou numérique et dont le rôle sera détaillé dans les lignes qui suivent.
- Lorsque $C = 0$, le modèle (1) coïncide avec la solution de l'EDP:

$$\begin{cases} \partial_t Y(t,x) = \lambda(x_1,x_2)Y(t,x) + S, \\ Y(0,x) = Y_0(x) = x_0. \end{cases} \tag{2}$$

[1] Ce qui suit, en terme d'analyse de sensibilité, peut être considéré comme une mise en commun des contenus de [4, 9, 11, 5, 7].

Ainsi, x_0 paramètrise la condition initiale et $\lambda(x_1, x_2)$ ressemble alors à une caractéristique matériau dont les paramètres physiques sont x_1, x_2. À noter que $\lambda(x_1, x_2)$ a volontairement été choisi de manière à ce que x_1 et x_2 jouent des rôles symétriques: nous verrons que les indices utilisés permettent de détecter ce comportement.

- Lorsque $C \neq 0$, le modèle (1) coïncide avec la solution de l'EDP précédente (2) avec une erreur de discrétisation (spatiale) $\Delta x = x_3$ telle que $\mathcal{O}(\Delta x) = \mathcal{O}(x_3)$ dont on supposerait connue la constante C et dont l'erreur serait proportionnelle au temps considéré[2], i.e. $\mathcal{O}(\Delta x) = tC\Delta x = tCx_3$. Ainsi $x_3 = \Delta x$ joue le rôle d'un paramètre de discrétisation numérique dont l'effet sur la solution s'amoindrit losque $x_3 = \Delta x \to 0$.

Nous allons maintenant aborder un problème de propagation d'incertitudes dans le modèle précédent.

3. Propagation d'incertitudes dans le modèle (1) et analyse de sensibilité

Dans cette section, nous effectuons une étude complète sur le modèle de la section 2. Nous commençons par modéliser (section 3.1) les incertitudes sur le vecteur de paramètre x, puis nous effectuons une propagation d'incertitudes, la commentons (section 3.2) et étudions la probabilité de dépassement d'un seuil du modèle, i.e. nous nous plaçons dans un contexte de garantie de performances (section 3.3). Enfin, nous construisons une sorte de méthodologie (i.e. utilisation combinée de plusieurs outils mathématiques et statistiques de la littérature) et mettons en évidence sa force, sa fiabilité (section 4).

3.1. Modélisation des incertitudes sur le vecteur des paramètres x

Supposons maintenant que les paramètres physiques x_0, x_1, x_2 sont incertains et qu'il faille choisir $x_3 = \Delta x$ de manière à éviter un effet de maillage trop important relativement aux incertitudes sur la solution.

Afin de modéliser les incertitudes sur les paramètres physiques, nous avons recours au cadre théorique probabiliste: nous supposons connues les distributions de ces paramètres (même si souvent, leur obtention représente un travail conséquent). Nous avons $X_i \sim \mathcal{U}([-1, 1]), \forall i \in \{0, 1, 2\}$, i.e. les paramètres x_0, x_1, x_2 sont modélisés par des variables aléatoires indépendantes uniformes sur $[-1, 1]$.

Contrairement aux paramètres physiques, dont les distributions peuvent être obtenues par métrologie par exemple ou par des expériences annexes, le paramètre $x_3 = \Delta x$ de maillage n'est pas réellement incertain. Simplement,

[2] A noter que cette forme est tout à fait réaliste du comportement des schémas numériques: l'erreur de discrétisation est souvent d'autant plus grande que le nombre d'itérations en temps est grand (et donc le temps final).

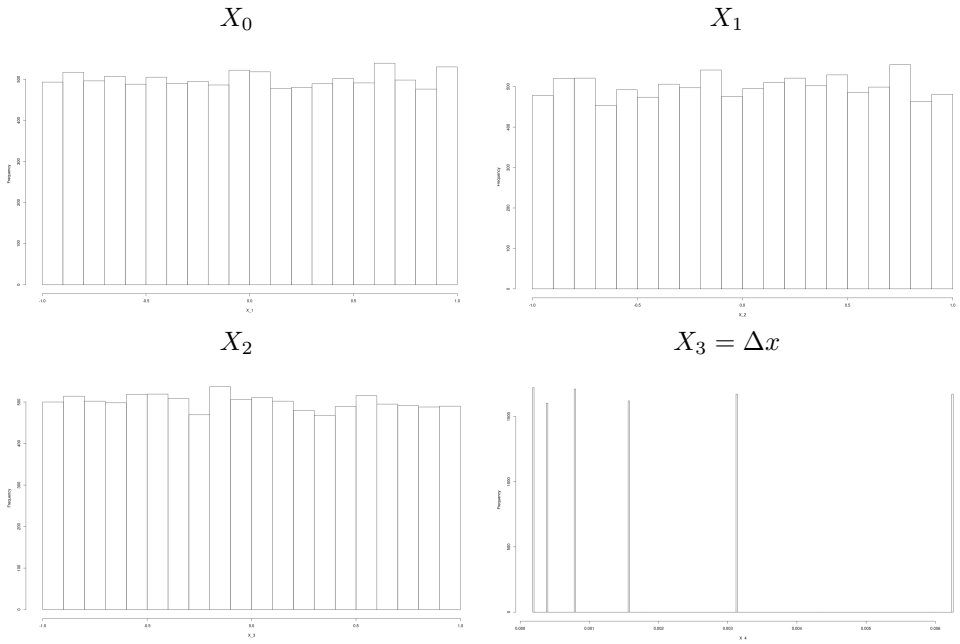

Figure 1: Histogrammes des variables aléatoires X_0, X_1, X_2 (uniformes dans $[-1, 1]$) et X_3 (discrète).

la valeur maximale de Δx permettant d'effectuer de manière fiable[3] une étude n'est pas connu *a priori*. Cette variable possède, à première vue, un statut particulier. Dans la suite, arbitrairement, nous allons supposer que $X_3 = \Delta x \sim \mathcal{M}(\{(p_0, \frac{1}{10\times 2^0}), ..., (p_N, \frac{1}{10\times 2^N})\})$, i.e. Δx suit une loi discrète à $N+1 = 10$ états dans $\{0.1, 0.05,\ 0.0125, 0.00625,\ 0.003125,\ 0.0015625, 0.00078125,\ 0.000390625\}$. Par soucis de concision, dans la suite, nous noterons $\{0.1, 0.05,\ 0.0125, 0.00625,\ 0.003125,\ 0.0015625, 0.00078125,\ 0.000390625\} = \{\Delta x_0, ..., \Delta x_9\}$. Concernant les probabilités $(p_i)_{i\in\{0,...,N\}}$, nous supposons que nous n'avons aucune raison de privilégier une discrétisation plutôt qu'une autre et prenons $p_i = \frac{1}{N+1}, \forall i \in \{0, ..., N\}$. L'intérêt de ce choix de distribution pour $X_3 = \Delta x$, arbitraire pour une variable qui n'a *a priori* rien d'aléatoire, sera mis en avant dans la suite.

Les histogrammes des réalisations de X_0, X_1, X_2, X_3 sont présentés figure 1. À noter que la prise en compte de variables discrètes (cf. histogramme en bas à droite de la figure 1) dans certaines analyses de sensibilité n'est pas un problème simple, cf. exemple 1 dans [5].

[3] Maillage le plus fin possible mais permettant de lancer suffisamment de calculs pour que les statistiques soient elles aussi convergées.

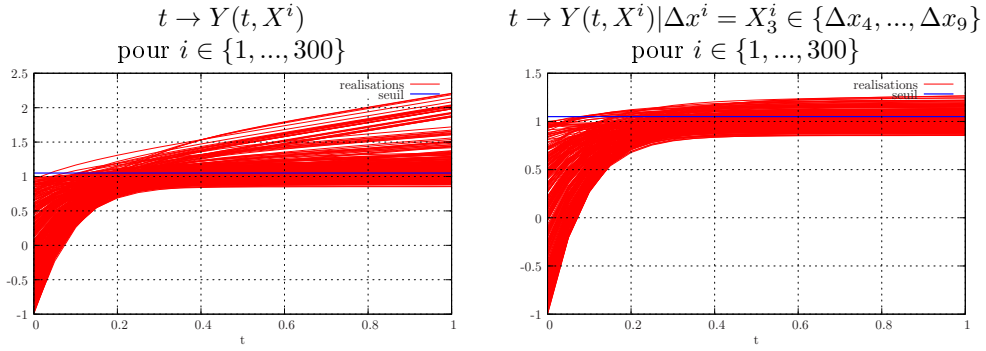

Figure 2: Évolutions temporelles $t \to Y(t, X^i)$ pour $N_{MC} = 300$ réalisations de $(X^i)_{i \in \{1,..N_{MC}\}}$. Gauche: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_0, ..., \Delta x_9\}$. Droite: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_4, ..., \Delta x_9\}$ (i.e. maillages plus fins).

3.2. Propagations d'incertitudes

Nous allons maintenant étudier le processus stochatique $Y(t, X)$ dont les composantes aléatoires de X suivent les distributions précitées (et illustrées figure 1). Pour cela, nous effectuons une propagation d'incertitudes par méthode Monte-Carlo (MC) à N_{MC} points. L'idée, bien connue, est

- d'échantillonner X, i.e. de générer N_{MC} réalisations indépendantes, identiquement distribuées $(w^i, X^i)_{i \in \{1,...,N_{MC}\}}$ de X (par exemple, les réalisations utilisées pour tracer les histogrammes de la figure 1), où les poids $w^i = \frac{1}{N_{MC}}, \forall i \in \{1, ..., N_{MC}\}$ pour un plan d'expériences MC;
- de lancer N_{MC} appels au modèle (1) afin de récupérer $(w^i, Y(t, X^i))_{i \in \{1,...,N_{MC}\}}$;
- de post-traiter les points $(w^i, Y(t, X^i))_{i \in \{1,...,N_{MC}\}}$ en fonction des besoins (cf. les discussions qui suivront sur le choix de l'observable *statistique* d'intérêt sur Y).

La figure 2 présente les résultats de deux propagations d'incertitudes, i.e. les évolutions temporelles $t \to Y(t, X^i), i \in \{1, ..., 300\}$, pour 300 réalisations de X:

- La première propagation d'incertitudes, figure 2 de gauche, a été effectuée dans les conditions précitées.
- La seconde propagation d'incertitudes, figure 2 de droite, a été effectuée en post-traitant les données et en conditionnant $X_3 = \Delta x$ aux états discrets $\{\Delta x_4, ..., \Delta x_9\}$ au lieu de $\{\Delta x_0, ..., \Delta x_9\}$. En d'autres termes, la seconde étude est une restriction de la première à des maillages fins.

En comparant les deux propagations d'incertitudes précédentes, il est possible de constater que le maillage, ici, a un effet important. Notamment, par exemple, sur la probabilité de dépasser le seuil $Y_0 = 1.05$. Ce seuil est représenté

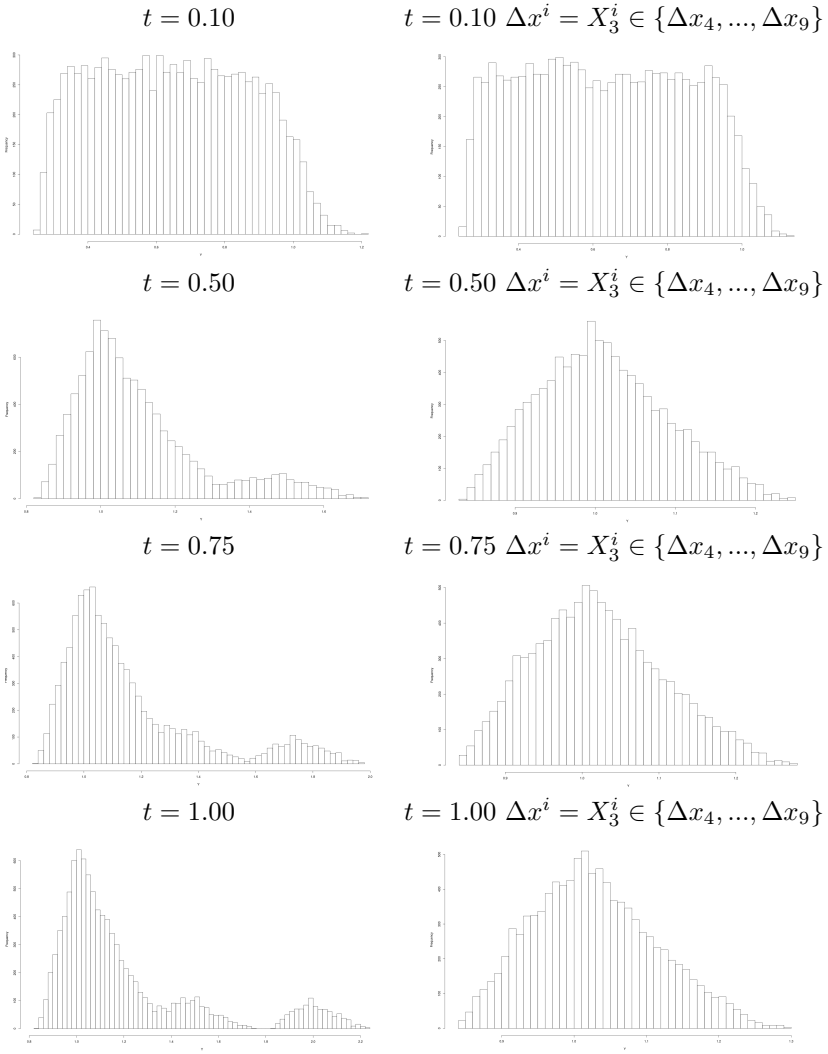

Figure 3: Histogrammes de $Y(t, X)$ pour $t \in \{0.10, 0.50, 0.75, 1.00\}$. Gauche: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_0, ..., \Delta x_9\}$. Droite: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_4, ..., \Delta x_9\}$ (i.e. maillages plus fins).

par la courbe bleue sur la figure 2: en effet, sur l'étude de droite, les évolutions temporelles ont tendance à beaucoup dépasser la valeur 1.05 alors que si l'on ne garde que les réalisations n'utilisant que des maillages fins (figure 2 droite), la proportion de courbes dépassant le seuil est beaucoup plus faible. Ceci est d'autant plus visible en observant les histogrammes de la variable aléatoire $Y(t, X)$ à plusieurs temps $t \in \{0.10, 0.50, 0.75, 1.00\}$ sur la figure 3. D'une part, les histogrammes obtenus sur tous les maillages tendent à devenir multi-

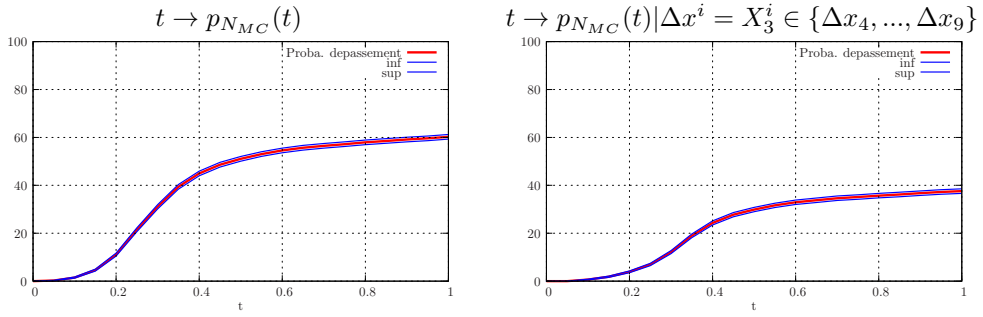

Figure 4: Évolutions temporelles $t \to p_{N_{MC}}(t)$ et intervalles de confiance associés. Gauche: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_0, ..., \Delta x_9\}$. Droite: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_4, ..., \Delta x_9\}$ (i.e. maillages plus fins).

modaux ou à supports non-continus (colonne de gauche de la figure 3), ce qui complique l'utilisation de métamodèles précis par exemple, alors que le comportement de la variable aléatoire est "plus" continu si le maillage est contraint aux paramètres de discrétisation fins (colonne de droite de la figure 3). De plus, la "masse" des réalisations allant au delà du seuil 1.05 est d'autant plus importante sur la colonne de gauche que sur la colonne de droite. En d'autres termes, l'observable statistique "probabilité de dépasser le seuil $Y_0 = 1.05$", notée $\mathbb{P}(Y(t,X) > Y_0) = 1.05)$ dans la suite, est très sensible au pas de discrétisation Δx et les maillages les plus fins ont tendance à diminuer son influence. Par la suite, *nous montrerons que les indicateurs statistiques considérés permettent de détecter et de quantifier ce genre d'effets.*

3.3. Probabilités de dépassement de seuil Y_0

Dans la section précédente, nous avons effectué une propagation d'incertitudes. Dans cette section, nous nous intéressons à l'observable statistique "probabilité de dépasser le seuil $Y_0 = 1.05$", notée $\mathbb{P}(Y(t,X) > Y_0) = 1.05)$. À partir des échantillons $(Y(t,X^i))_{i\in\{1,...,N_{MC}\}}$, il est possible d'estimer $\mathbb{P}(Y(t,X) > Y_0) = 1.05)$ ainsi que son intervalle de confiance et ce, pour plusieurs temps.

Notons $p_{N_{MC}}(t)$ l'estimation de $\mathbb{P}(Y(t,X) > Y_0) = 1.05)$ effectuée avec N_{MC} points MC au temps t. On sait alors qu'asymptotiquement [12], l'intervalle de confiance à 95% est donné par

$$\mathbb{P}(Y(t,X) > Y_0) = 1.05) \in \left[p_{N_{MC}}(t) - 1.96\frac{\sigma_{N_{MC}}(t)}{\sqrt{N_{MC}}}, p_{N_{MC}}(t) + 1.96\frac{\sigma_{N_{MC}}(t)}{\sqrt{N_{MC}}} \right],$$

avec

$$p_{N_{MC}}(t) = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} \mathbf{1}_{[Y_0,\infty[}(Y(t,X^i)) \text{ et } \sigma_{N_{MC}}(t) = \sqrt{p_{N_{MC}}(t)\left(1 - p_{N_{MC}}(t)\right)}.$$

La figure 4 présente les probabilités de dépassement de seuil dans les mêmes conditions que pour la figure 2 (avec tous les maillages à gauche et en conditionnant sur les maillages fins à droite) ainsi que l'intervalle de confiance à 95%. Comme intuité en observant les figures 2–3, la probabilité de dépassement de seuil est fortement impactée par le choix de la discrétisation: celle-ci est divisée par 1.5 en n'utilisant uniquement les maillages fins. En d'autres termes, le maillage risque certainement plus d'expliquer le fait de dépasser le seuil que les incertitudes sur les paramètres physiques. Ceci est problématique et dans la suite, nous allons travailler à pouvoir quantifier l'influence du maillage sur cette observable statistique tout en évitant un coût de calcul supplémentaire prohibitif.

4. Analyse de sensibilité sur la probabilité de dépassement de seuil

Dans la section qui précède, nous avons mis en évidence un comportement problématique que l'on souhaiterait pouvoir détecter *de manière fiable* et *à moindre coût* lors d'une étude de propagation d'incertitudes ou de garantie (probabilité de dépassement de seuil).

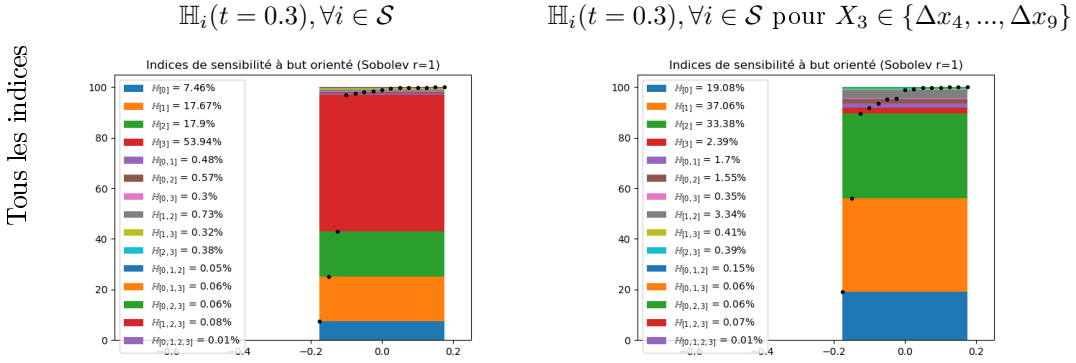

Figure 5: Indices HSIC appliqué au problème jouet (1) à $t = 0.3$. La colonne de gauche présente les indices calculés dans les conditions décrites dans la section 3.1. La colonne de droite présente les indices obtenus lorsque la variable de pas de discrétisation $X_3 = \Delta x$ est conditionnée au maillage plus fin $\{\Delta x_4, .., \Delta x_9\}$.

L'inclusion de la variable $X_3 = \Delta x$ parmi les paramètres incertains permet, dans un sens, d'adresser la question du "*moindre coût*" précitée: partant du principe qu'il nous faille propager beaucoup de paramètres incertains et que seul un plan d'expériences MC le permet[4], ce plan d'expériences étant peu sensible à la dimension, autant y ajouter la dimension supplémentaire $X_3 = \Delta x$ (d'où la nécessité de la modéliser probabilistiquement, cf. section 3.1).

[4]La méthode MC est la seule dont la vitesse de convergence est indépendante de la dimension et de la régularité et, même si cette vitesse est relativement lente, elle est compétitive

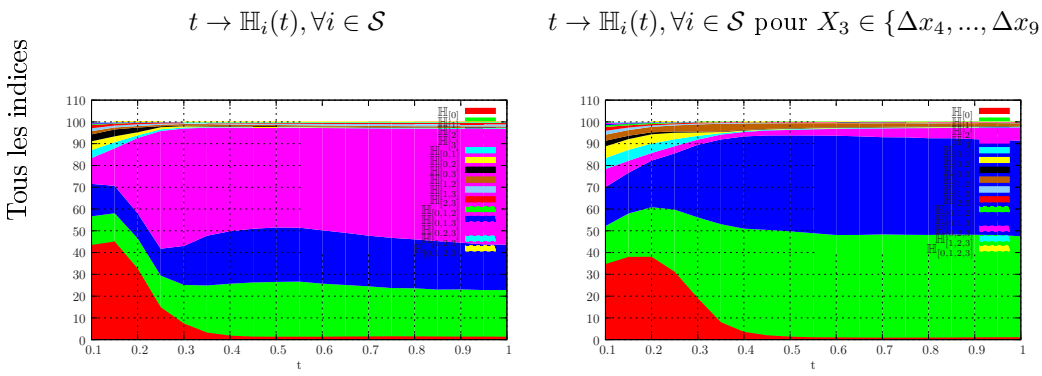

Figure 6: Évolution temporelle des indices HSIC appliqué au problème jouet (1). La colonne de gauche présente les indices calculés dans les conditions décrites dans la section 3.1. La colonne de droite présente les indices obtenus lorsque la variable de pas de discrétisation $X_3 = \Delta x$ est conditionnée au maillage plus fin $\{\Delta x_4, .., \Delta x_9\}$.

Reste la question de la *fiabilité*: effectuer une analyse de sensibilité à but orienté sur un mélange de variables continues et discrètes peut être complexe ou peu fiable selon les outils utilisés, cf. exemple 1 de [5]. Pour cette raison, nous combinons dans la suite plusieurs ingrédients *très récents* de la littérature, cf. [5, 9, 11, 7]. Sans trop rentrer dans les détails, nous avons recours à une combinaison de ces 4 papiers/travaux pour les raisons suivantes:

- pour être insensible au caractère discret ou non de la distribution, nous comptons sur les travaux décrits dans [5] consistant à effectuer une transformation iso-probabiliste et de calculer les indices à partir des variables aléatoires uniformes dans $[0, 1]$ ainsi obtenues $\mathcal{U}_0, \mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3$ pour X_0, X_1, X_2, X_3. Un exemple simple permettant d'insister sur l'importance de cette étape est donné dans la section Appendix B;

- pour avoir accès à un indice de sensibilité qui prend en compte tous les moments[5] de la distribution (indices dits 'moment-independent'), nous avons recours aux indices HSIC [9] qui peuvent, de manière fiable, être appliqués pour quantifier la sensibilité à une probabilité de dépassement de seuil et dont la complexité d'application *en terme de nombre d'appels au modèle à effectuer* est indépendante de la dimension[6].

- Nous utilisons les indices HSIC dans les conditions présentés dans [11]. Une décomposition ANOVA est alors possible: pour résumer l'intérêt de

dès lors que la dimension est supérieure à 10-20 par exemple.

[5] Les indices de Sobol ne quantifient l'impact des entrées uniquement sur la variance, i.e. le moment d'ordre 2, cf. [9] ou [4] pour un exemple pédagogique de risques encourus lorsque l'on utilise les indices de Sobol sur une probabilité de dépassement de seuil.

[6] Par exemple, pour les indices de Sobol ou les indices MMD [11], la complexité croit linéairement avec la dimension.

l'existence d'une telle décomposition, les indices peuvent être interprétés comme des pourcentages et peuvent rigoureusement être hiérarchisés. Les interactions d'ordre élevées peuvent même être estimées.

- Finalement, les indices HSIC nécessitent le choix d'un noyau (cf. [9, 11]). Pour ce choix, nous nous reposons sur le matériel décrit dans [7]: compte-tenu de l'analyse effectuée, le choix du noyau de Sobolev d'ordre $r = 1$ s'impose.

À noter que nous donnons plus de détails sur les indices utilisés dans la section Appendix A.

Dans la suite, nous appliquons les précédents indices sur le problème jouet des sections 1–3.2 et montrons, par l'exemple, qu'ils permettent d'effectuer une étude fiable (i.e. statistiquement convergée).

La figure 5 présente les indices de sensibilité, notés

$$\mathbb{H}_i(t), \forall i \in \mathcal{S},$$

avec $\mathcal{S}$ l'ensemble de toutes les combinaisons possibles d'indices

$$\mathcal{S} = \{\{0\}, \{1\}, .., \{0,1\}, ...\{0,1,2,3\}\},$$

sur la probabilité de dépassement de seuil $\mathbb{P}(Y(t,X) > Y_0) = 1.05)$ au temps $t = 0.3$. A gauche, les indices sont présentés sans contraintes sur $X_3 = \Delta x$ alors qu'à droite, la même variable est contrainte aux maillages les plus fins, i.e. $X_3 = \Delta x \in \{\Delta x_4, ..., \Delta x_9\}$. La figure 5 (gauche) peut être interprétée de la sorte: presque 54% de la probabilité de dépassement de seuil est expliquée par la variable de discrétisation $X_3 = \Delta x$. En d'autres termes, *pour les maillages choisis, la quantité d'intérêt est trop sensible au maillage pour être interprétable.* Compte-tenu de ces résultats, il est possible de resteindre *a posteriori* la variable $X_3 = \Delta x$ a des maillages plus fins. Aucun calcul supplémentaire n'est nécessaire, il suffit juste de post-traiter les points/résultats en conditionnant $X_3 = \Delta x$ à l'espace d'états $\{\Delta x_4, ..., \Delta x_9\}$ plutôt que $\{\Delta x_0, ..., \Delta x_9\}$. Le seul risque est d'avoir une statistique moins convergée (parce que moins de tirages sont disponibles). Les indices de sensibilité contraints aux maillages fins sont présentés sur la figure 5 (droite): l'effet de la variable de discrétisation $X_3 = \Delta x$ est négligeable sur la probabilité de dépassement de seuil. En effet, tous les indices $\mathbb{H}_u(t = 0.3), u \in \mathcal{S}$ tels que $3 \in u \in \mathcal{S}$ sont très faibles ($< 3.0\%$) traduisant un effet total très faible. Ainsi, il est possible d'interpréter physiquement les résultats de la figure 5: au temps $t = 0.3$, les variables X_0, X_1 et X_2 expliquent respectivement 19%, 37% et 34% de la probabilité de dépassement de seuil, pour un total de 90%. Les intéractions sont sans aucun doute très secondaires, en tout cas à ce temps.

Il est alors possible de tracer l'évolution de ces indices au cours du temps $t \to \mathbb{H}_i(t), i \in \mathcal{S}$, cf. figure 6, dans les mêmes conditions que précédemment. Sur

la figure 6 de gauche, il est possible de constater que l'effet du maillage est trop important, y compris pour des temps courts, avec environ 10% d'effet dès $t = 0.1$ et une tendance à l'augmentation. Pour cette raison, nous ne nous autoriserons pas à interpréter les résultats aux temps $t > 0.3$ et focalisons sur l'étude de la figure 6 de droite: l'effet du maillage (couleur magenta) est négligeable, tend à augmenter (ce qui est tout à fait en accord avec la tendance linéaire en temps du modèle (1), même si en pratique, nous ne sommes pas sensés avoir accès au modèle), sans jamais dépasser environ 8% tout en restant toujours faible devant les effets principaux. Ainsi, sur la figure 6 (droite), il est possible de retrouver le fait qu'au temps court, la variable X_0 paramétrant la condition initiale est prépondérante (en temps courts, se sont les valeurs les plus grandes de X_0 qui permettent à Y de se rapprocher Y_0). En temps plus long, l'effet de X_0 s'estompe (nous retrouvons le fait que l'état d'équilibre est peu dépendant de la condition initiale). Les effets des variables X_1, X_2 sont de plus en plus importants au cours du temps. Leurs effets sont quasi identiques, comme en témoigne leur rôle symétrique dans (1). À noter qu'en temps court, $t \in [0.1, 0.4]$, les intéractions entre X_0 et X_1 (cyan) et entre X_0 et X_2 (jaune) ne sont pas négligeables (avec environ 12% d'intéractions à $t = 0.1$): en effet, dans cet intervalle de temps, il faut à la fois que X_0 soit grand et que X_1, X_2 rendent λ raide pour que Y dépasse Y_0. Au temps final, 8% de la sensibilité de la probabilité de dépassement de seuil est expliquée par la variable de discrétisation $X_3 = \Delta x$ mais ces 8% restent négligeables devant les 85% des effets des variables X_1, X_2: en d'autres termes, dans ce cas, même si la variable de discrétisation $X_3 = \Delta x$ joue à 8%, les résultats restent interprétables physiquement.

Dans la section précédente, nous avons mis en avant le fait que les indices considérés permettent de répondre à nos besoins en terme d'aide à la décision: les outils sont fiables, permettent de détecter l'utilisation d'un maillage trop grossier, permettent de hiérarchiser de manière quantifiée (interprétation en terme de pourcentages) les contributeurs, de détecter des intéractions.

Toutefois, en dimension incertaine plus grande, une difficulté supplémentaire intervient: par exemple, si le nombre de paramètres incertains est égal à 110, le cardinal de $\mathcal{S}$ est $\#\mathcal{S} = 2^{110} \sim 10^{36}$. Les indices ne sont plus calculables en pratique dans ces conditions. Toutefois, souvent, nous n'avons pas besoin d'avoir les informations par paramètres, il nous suffit souvent de les avoir par groupes. La section qui suit présente comment il est possible d'éviter la difficulté précitée et de construire des indices de sensibilité par groupes de paramètres.

4.1. Analyse de sensibilité sur la probabilité de dépassement de seuil avec réduction de dimension

Dans cette section qui suit, nous reprenons la même étude mais en utilisant une astuce supplémentaire pour réduire la dimension du problème (à noter que l'astuce peut également être utilisée pour prendre en compte des corrélations dans les variables d'entrée). Supposons que, compte-tenu des besoins de l'étude, connaitre la part d'influence de X_1 et de X_2 ne soit pas nécessaire: il nous suffit de connaitre la part de $\lambda(X_1, X_2)$. Dans ce cas, s'intéresser à la variable aléatoire Λ permet de réduire la dimension à 3 plutôt que 4. Pour cela, il nous faut

déterminer la distribution de $\Lambda = \lambda(X_1, X_2)$. Ceci peut être effectué *a priori* et est même possible lorsque X_1 et X_2 ne sont pas indépendants, en pré-traitant les données et en préparant N_{MC} réalisations $(\Lambda^i = \lambda(X_1^i, X_2^i))_{i\in\{1,...,N_{MC}\}}$.

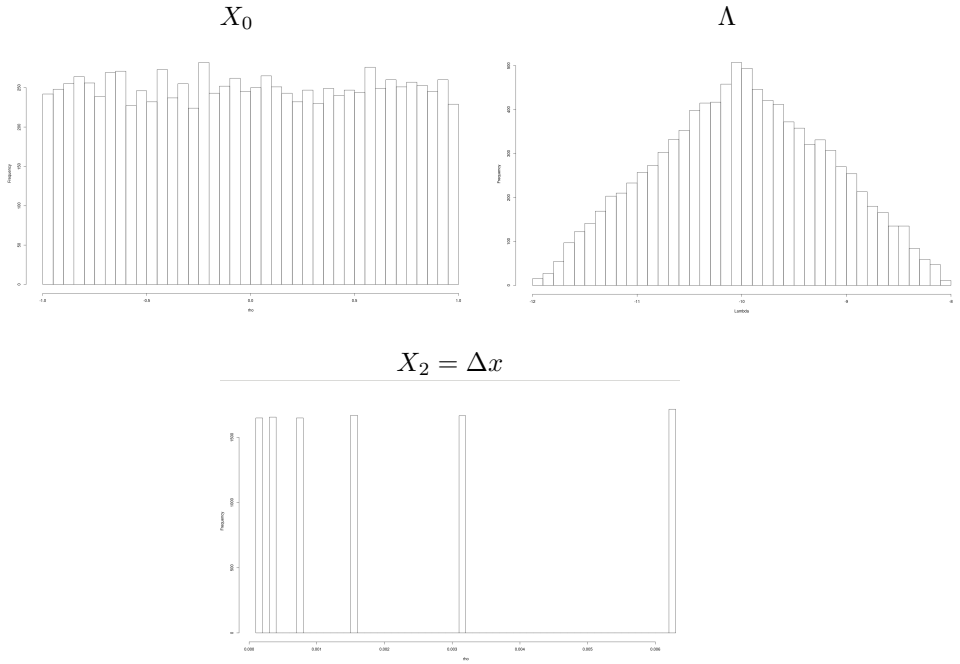

Figure 7: Histogrammes des variables aléatoires $X_0, \Lambda, X_2 = \Delta x$.

La figure 7 présente les distributions de X_0 (idem figure 1), de $\Lambda = \lambda(X_1, X_2)$ et de $X_3 = \Delta x$ (idem figure 1). Pour l'étude considérée, connaitre la distribution de Λ n'est pas difficile: il s'agit d'une somme de deux variables aléatoires uniformes de support tangent, la distribution est triangulaire. Il est possible de l'échantilloner aisément à partir d'un tirage uniforme en inversant sa *cumulative density function* (CDF):

$$\mathcal{T} = F_\Lambda^{-1}(\mathcal{U}) = \mathbf{1}_{]-\infty,0]}(2\mathcal{U}-1)\left(-2+2\sqrt{2}\sqrt{\mathcal{U}}\right)+\mathbf{1}_{[0,\infty}(2\mathcal{U}-1)\left(2-2\sqrt{2-2\mathcal{U}}\right),$$

où $\mathcal{U} \sim \mathcal{U}([0,1])$.

La propagation d'incertitudes donne, à l'erreur statistique près, les mêmes résultats et les indices de sensibilité à but orienté au cours du temps sont présentés figure 8. À noter que pour cette étude, étant donné qu'il n'y a plus que 3 variables X_0, Λ et $X_2 = \Delta x$, l'ensemble sur lequel les indices sont calculés est $\mathcal{S} = \{\{0\}, \{1\}, \{2\}, \{0,1\}, ..., \{0,1,2\}\}$. Il y a donc moins d'indices à évaluer mais par contre, les indices avec la numérotation 1 (i.e. incluant les effets de Λ) incluent les effets, indiscernables, de X_1 et X_2. Ainsi, la figure 8, présentant l'évolution temporelle des indices de sensibilité, peut être directement comparée

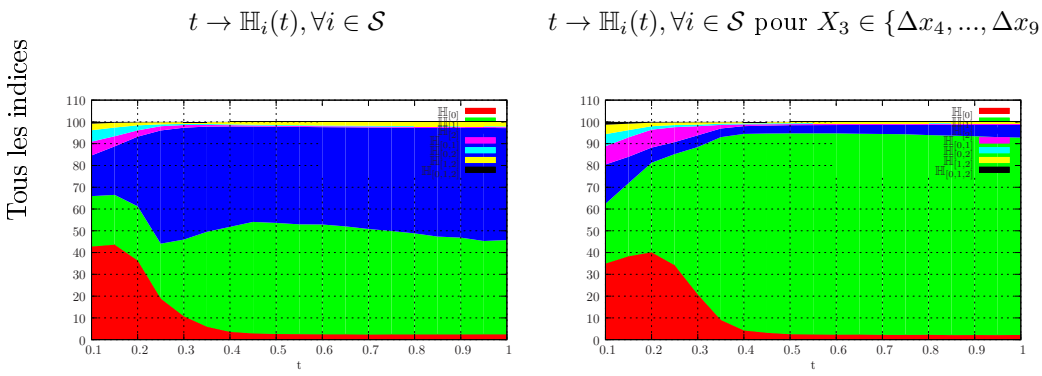

Figure 8: Évolution temporelle des indices HSIC appliqué au problème jouet (1) en ne considérant que la variable Λ au lieu de X_1, X_2. Avec la renumérotation, $X_2 = \Delta x$ ici. La colonne de gauche présente les indices calculés dans les conditions décrites dans la section 3.1. La colonne de droite présente les indices obtenus lorsque la variable de pas de discrétisation $X_3 = \Delta x$ est conditionnée au maillage plus fin $\{\Delta x_4, .., \Delta x_9\}$.

à la figure 6: dans les deux études, le rôle de X_0 est le même, de même que le rôle de Δx. Les rôles de X_1, X_2 sur la figure 6 sont "répartis" au sein du paramètre Λ, numéroté 1, sur la figure 8. En d'autres termes, les deux études sont équivalentes, au bruit statistique près. Il est important de noter que le fait de retrouver les effets de X_1 et X_2 dans Λ est également étroitement liée au fait de pouvoir interpréter les indices comme des pourcentages: sans cela, il n'aurait, en théorie, pas été possible de retrouver des résultats équivalents.

Pour cette seconde étude, avec les indices utilisés ici, la dimension a été réduite (passage de la dimension 4 à la dimension 3) au seul prix de ne plus pouvoir discerner les rôles respectifs de X_1 et X_2.

4.2. Analyse de sensibilité sur la probabilité de dépassement de seuil avec réduction de dimensions mais en plus général encore

Dans la section précédente, la dimension a pu être réduite en considérant la variable $\Lambda = \lambda(X_1, X_2)$ et en effectuant l'étude en dimension 3 plutôt qu'en dimension 2. La seule contrainte pour se faire a été de pouvoir echantilloner Λ simplement à partir d'une variable uniforme. *Pour d'autres lois, ce travail peut être beaucoup plus compliqué.* Nous présentons ici une méthode relativement simple (mais calculatoire) permettant d'effectuer l'étape précédente. Cette nouvelle méthode est basée sur le fait de focaliser sur la variable aléatoire "modèle de λ" plutôt que Λ. Nous la détaillons dans ce qui suit.

Supposons que nous ayons facilement accès à un ensemble de tirages MC de Λ, i.e. nous avons $(\Lambda^i = \lambda(X_1^i, X_2^i))_{i \in \{1,...,N_{MC}\}}$. Implicitement, ces échantillons sont pondérés de masses $w_i^{MC} = \frac{1}{N_{MC}}, \forall i \in \{1, ..., N_{MC}\}$, i.e. il serait plus rigoureux d'écrire l'échantillon $(w_i^{MC}, \Lambda^i)_{i \in \{1,...,N_{MC}\}}$. Nous proposons de construire un nouvel échantillon pondéré $(w_i, \Lambda^i)_{i \in \{1,...,N\}}$ à N points avec $N \ll N_{MC}$. Supposons que nous ayons accès à ce nouvel échantillon et qu'il

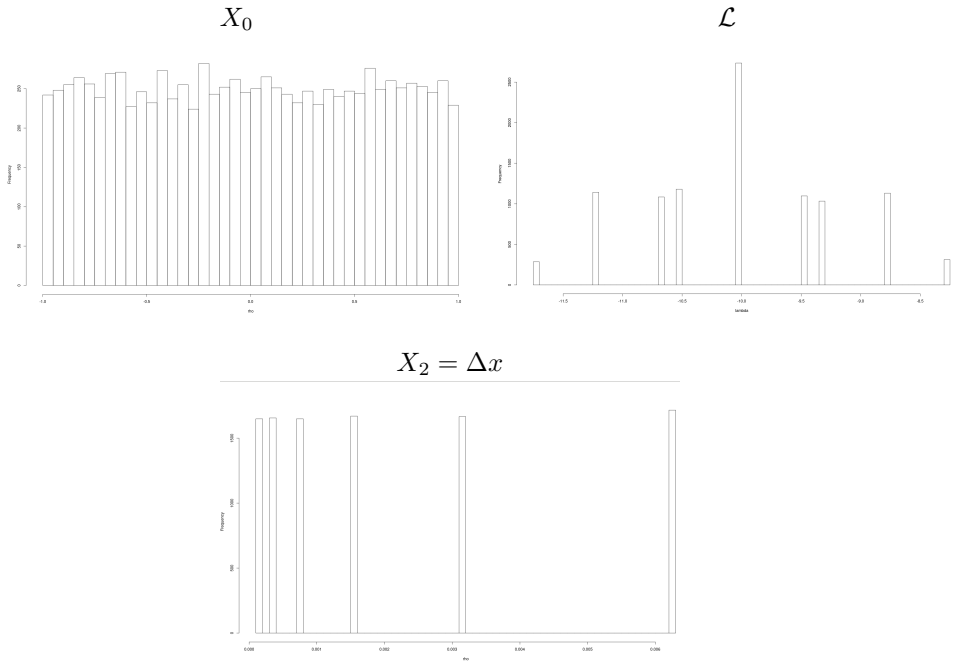

Figure 9: Histogrammes des variables aléatoires $X_0, \mathcal{L}, X_2 = \Delta x$.

respecte certaines contraintes, notamment que pour un certain nombre $N_m + 1$ de moment $(s_k = \mathbb{E}[\Lambda^k])_{k\in\{0,...,N_m\}}$, on ait

$$s^k \approx s_k^N = \sum_{i=1}^{N} w_i (\Lambda^i)^k \approx s_k^{N_{MC}} = \sum_{i=1}^{N_{MC}} \frac{1}{N_{MC}} (\Lambda^i)^k.$$

Nous proposons alors de considérer la variable aléatoire discrète "modèle de λ" à N états $(\Lambda^i)_{i\in\{1,...,N\}}$ chacun de probabilité $(w_i)_{i\in\{1,...,N\}}$. Cette variable aléatoire discrète, notée $\mathcal{L}$, est telle que

$$\mathbb{E}[\mathcal{L}^k] = \sum_{i=1}^{N} w_i (\Lambda^i)^k = s_k^N.$$

Ainsi, si N est suffisamment grand pour que $s_k \approx s_k^N$ pour un certain nombre de moments k, on aura $\mathbb{E}[\mathcal{L}^k] \approx s_k$. Une façon de construire $\mathcal{L}$, par exemple, est d'utiliser des points de Gauss-Legendre. Pour $\Lambda = \lambda(X_1, X_2)$, cela revient à prendre les points de Gauss-Legendre $(w_1^i, X_1^i)_{i\in\{1,...,N_1\}}$ pour X_1, les points de Gauss-Legendre $(w_2^i, X_2^i)_{i\in\{1,...,N_2\}}$ pour X_2, de les tensoriser pour obtenir $(w_k^\Lambda, \Lambda^k)_{k\in\{1,...,N=N_1\times N_2\}}$ avec $w_k^\Lambda = w_1^i \times w_j^1, \Lambda^k = \lambda(X_1^i, X_2^j), \forall (i,j) \in \{1, ..., N_1\} \times \{1, ..., N_2\}$.

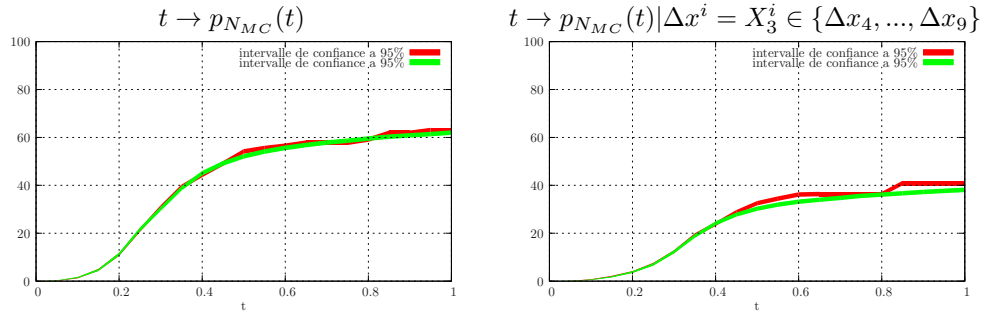

Figure 10: Évolutions temporelles $t \to p_{N_{MC}}(t)$ et intervalles de confiance associés. Gauche: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_0, ..., \Delta x_9\}$. Droite: la variable de discrétisation spatiale $X_3 = \Delta x$ est échantillonnée dans $\{\Delta x_4, ..., \Delta x_9\}$ (i.e. maillages plus fins).

Tous les indices

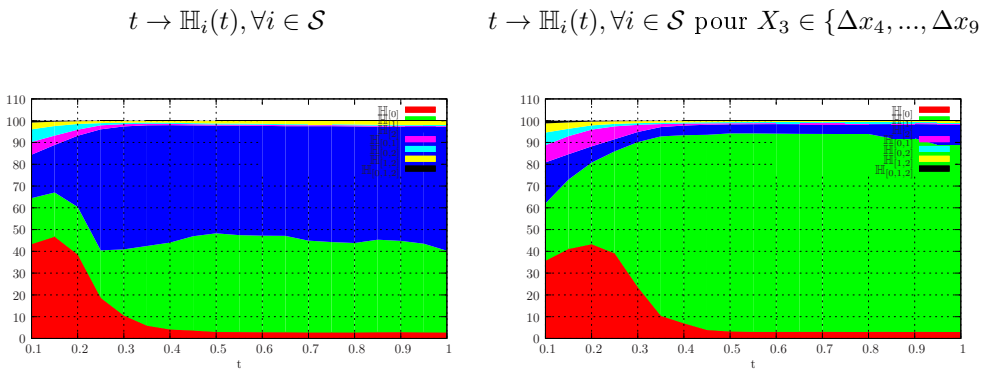

Figure 11: Évolution temporelle des indices HSIC appliqué au problème jouet (1). La colonne de gauche présente les indices calculés dans les conditions décrites dans la section 3.1. La colonne de droite présente les indices obtenus lorsque la variable de pas de discrétisation $X_3 = \Delta x$ est conditionnée au maillage plus fin $\{\Delta x_4, .., \Delta x_9\}$.

La figure 9 présente les histogrammes de X_0, $\mathcal{L}$ et $X_2 = \Delta x$. Les variables aléatoires X_0 et $X_2 = \Delta x$ sont traitées comme dans les sections précédentes. Par contre, $\mathcal{L}$ est discrète sur $N = N_1 \times N_2 = 6 \times 6 = 36$ états pondérés: moins d'états apparaissent du fait de la symétrie entre X_1 et X_2 dans l'expression de λ.

La figure 10 présente les résultats de l'évaluation de la probabilité de dépassement de seuil au cours de temps avec une propagation d'incertitudes obtenues en échantillonnant les variables $X_0, \mathcal{L}, \Delta x$ (courbe verte) au lieu de $X_0, \Lambda, \Delta x$ (référence, courbe rouge). Les résultats sont sensiblement les mêmes, même si plus d'états auraient pu être considérés pour une meilleure statistique.

Reprenons alors la même étude d'analyse de sensibilité que précédemment mais sur $X_0, \mathcal{L}, \Delta x$: celle-ci est présentée figure 11. Cette figure 11 peut être directement comparée à la figure 8: les résultats sont identiques. Il a donc été

possible de réduire la dimension en n'ayant recours qu'à un échantillon pondéré relativement facile à construire plutôt que de trouver la loi de Λ ainsi que d'inverser sa fonction de répartition. Cette méthodologie de réduction de dimension exploite intensivement le fait de pouvoir, de manière fiable cf. travaux de [5] et section Appendix B, prendre en compte des variables discrètes avec les indices HSIC.

Remarque Dans les paragraphes précédents, nous avons réduit la dimension de manière calculatoire mais en ayant recours à des points de Gauss. Ces points ne sont utilisables qu'en petite dimension. Mais la méthodologie reste indépendante du type de points utilisés: il est aisé de remplacer les points de Gauss par un LHS (Latin Hypercube Sampling) par exemple. Les résultats sont très proches des précédents et ne sont pas documentés dans ce rapport pour éviter les redondances. •

5. Conclusion

Dans ce document, nous avons présenté des outils récents (leur combinaison n'a, à notre connaissance, jamais été publié dans litérature) adaptés à une application à une analyse de sensibilité pour la garantie. A noter qu'il faudrait envisager de passer par une phase de parallélisation du posttraitement calculant les indices (la complexité est en N^2 où N est le nombre d'échantillons et pour N grand et selon les cas, le posttraitement peut être coûteux).

6. Remerciements

Merci à Marc Sancandi pour les discussions fructueuses sur la réduction de dimension. Merci à Pierre Minvielle pour ses conseils la notion probabiliste de garantie.

References

[1] K. Fukumizu *et al.* Kernel measures and conditional dependences. *Advances in neural information processing systems*, pages 489–4996, 2008.

[2] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[3] David Hébert. Étuude de convergence avec le code Hochimin: prise en coompte d'un dépôt d'énergie et d'un empilement de deux matériaux. Technical report, CEA, 2021.

[4] Collectif LSDR. Analyse de sensibilité à but orienté & Calibration sous incertitudes: exemples, enjeux & difficultés. Technical Report DR 41, CEA DAM CESTA, 2021.

[5] Paul Novello, Gaël Poëtte, David Lugato, and Pietro M Congedo. Goal-Oriented Sensitivity Analysis of Hyperparameters in Deep Learning. *Journal of Scientific Computing*, page 94:45, 2023.

[6] Mattheo Saldanha. Évaluation des méthodes de calibration bayesienne pour le durcissement et les expériences de sollicitations mécaniques: applications aux tirs DEMETER. Technical report, CEA, sous la supervision de F. Malaise & G. Poëtte, 2021.

[7] G. Sarrazin, A. Marrel, S. Da Veiga, and V. Chabridon. Towards more interpretable kernel-based global sensitivity analysis: new insights into Sobolev kernels and their feature maps. Technical Report DO 08, CEA/DES/IRESNE/DER/SESI/LEMS/NT, 2022.

[8] Adrien Spagnol, Rodolphe Le Riche, Sébastien Da Veiga, and Olivier Roustant. Global sensitivity analysis for optimization with variable selection. In *PGMO Days 2017*, Saclay, France, November 2017.

[9] S. Da Veiga, F. Gamboa, and B. Iooss annd C. Prieur. *Basics and Trends in Sensitivity analysis, theory and practice in R.* Computational Science and Engineering. SIAM, 2020.

[10] Sébastien Da Veiga. Global sensitivity analysis with dependence measures. 2013.

[11] Sébastien Da Veiga. Kernel-based anova decomposition and shapley effects – application to global sensitivity analysis. 2021. arXiv:2101.05487v1.

[12] Clément Walter. *Using Poisson Processes for rare event estimations.* PhD thesis, Université de Paris VII, November 2016.

Appendix A. Les indices de sensibilité à but orienté HSICs et leur décomposition ANOVA

Dans cette section, nous rappelons brièvement comment sont construits les indices utilisés dans ce document, leur principe et les choix effectués (notamment au niveau du noyau). La seule originalité de cette section consiste en le fait de mettre en commun ce qui est présent dans plusieurs publications [10, 8, 9, 11, 5, 7]. Cette section ne remplace en aucun cas la lecture et la compréhension des documents [10, 8, 9, 11, 5, 7].

Appendix A.1. Une métrique sur des distributions: pour un indice ne focalisant pas sur un moment en particulier

Soient X, Y deux variables aléatoires de mesures de probabilité respectives $\mathrm{d}\mathcal{P}_X, \mathrm{d}\mathcal{P}_Y \in \mathcal{X}$. Dans [2], les auteurs montrent que

$$\mathrm{d}\mathcal{P}_X = \mathrm{d}\mathcal{P}_Y \iff \mathbb{E}[f(X)] - \mathbb{E}[f(Y)] = 0, \forall f \in \mathcal{C}(\mathcal{X}),$$

où $\mathcal{C}(\mathcal{X})$ est l'espace des fonctions continues bornées par 1 sur $\mathcal{X}$. Construisons

$$\gamma(\mathcal{F}, \mathrm{d}\mathcal{P}_X, \mathrm{d}\mathcal{P}_Y) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}[f(X)] - \mathbb{E}[f(Y)] \right|,$$

et où $\mathcal{F}$ est une classe de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ bornées par 1. Alors γ est appelé un IPM (Integral Probability Metric). En prenant $\mathcal{F} = \mathcal{C}(\mathcal{X})$ et en exploitant l'équivalence précédente, il est possible de construire une métrique permettant de quantifier à quel point $\mathrm{d}\mathcal{P}_X$ et $\mathrm{d}\mathcal{P}_Y$ sont proches ou non.

Mais γ n'a que peu d'intérêt pratique lorsque $\mathcal{F} = \mathcal{C}(\mathcal{X})$: comment tester toutes les fonctions $f \in \mathcal{C}(\mathcal{X})$? Le MMD (Maximum Mean Discrepancy) est alors un IPM défini sur une classe de fonctions restreinte $\mathcal{F} = \mathcal{F}_{\mathcal{H}} \subset \mathcal{C}(\mathcal{X})$ où

$$f : \mathcal{X} \longrightarrow \mathcal{H},$$

où $\mathcal{H}$ est un RKHS (Reproducing Kernel Hilbert Space). Si $\mathcal{F}_{\mathcal{H}}$ est suffisamment dense dans $\mathcal{C}(\mathcal{X})$ alors le tour est joué... Mais quel est l'intérêt de passer par ce RKHS? Un RKHS est "livré" avec son noyau k. À noter que pour l'argument de densité précédent, le choix du noyau et donc du RKHS est important, cf. [10, 11]: le noyau doit être *caractéristique*. Ce noyau est une fonction

$$k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R},$$

définie par le fait que $\forall f \in \mathcal{F}_{\mathcal{H}}$

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}.$$

Qu'est ce que cela apporte? Effectuons quelques calculs:

$$\begin{aligned}
\gamma_k(X,Y) = \gamma(\mathcal{F}_{\mathcal{H}}, \mathrm{d}\mathcal{P}_X, \mathrm{d}\mathcal{P}_Y) &= \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \mathbb{E}[f(X)] - \mathbb{E}[f(Y)] \right|, \\
&\overset{\text{def.}}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \int f(X) \mathrm{d}\mathcal{P}_X - \int f(Y) \mathrm{d}\mathcal{P}_Y \right|, \\
&\overset{\text{def.}}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \int \langle f, k(X, \cdot) \rangle_{\mathcal{H}} \, \mathrm{d}\mathcal{P}_X - \int \langle f, k(Y, \cdot) \rangle_{\mathcal{H}} \, \mathrm{d}\mathcal{P}_Y \right|, \\
&\overset{\text{lin.}}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \left\langle f, \int k(X, \cdot) \mathrm{d}\mathcal{P}_X \right\rangle_{\mathcal{H}} - \left\langle f, \int k(Y, \cdot) \mathrm{d}\mathcal{P}_Y \right\rangle_{\mathcal{H}} \right|, \\
&\overset{\text{lin.}}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \left\langle f, \int k(X, \cdot) \mathrm{d}\mathcal{P}_X - \int k(Y, \cdot) \mathrm{d}\mathcal{P}_Y \right\rangle_{\mathcal{H}} \right|, \\
&\overset{\|f\| \leq 1}{=} \left\| \int k(X, \cdot) \mathrm{d}\mathcal{P}_X - \int k(Y, \cdot) \mathrm{d}\mathcal{P}_Y \right\|_{\mathcal{H}}, \\
\gamma_k^2(X,Y) &= \int \int k(x,y) (\mathrm{d}\mathcal{P}_X(x) - \mathrm{d}\mathcal{P}_Y(x)) (\mathrm{d}\mathcal{P}_X(y) - \mathrm{d}\mathcal{P}_Y(y)).
\end{aligned} \tag{A.1}$$

La dernière ligne de l'équation précédente montre que $\gamma_k(X,Y)$ peut être évaluée. Pour cela, il suffit d'effectuer une intégration numérique de k par rapport aux distributions $\mathrm{d}\mathcal{P}_X, \mathrm{d}\mathcal{P}_Y$. In particulier, supposons que l'on ait accès à deux

échantillons $\mathbb{X} = \{X^1, ..., X^m\} \sim \mathrm{d}\mathcal{P}_X$ et $\mathbb{Y} = \{Y^1, ..., Y^n\} \sim \mathrm{d}\mathcal{P}_Y$, alors un estimateur non biaisé de (A.1) est donné par

$$\gamma_k^2(\mathbb{X}, \mathbb{Y}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} k(X^i, X^j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} k(Y^i, Y^j) - \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} k(X^i, Y^j).$$

La complexité du calcul est alors en $\mathcal{O}((n+m)^2)$: le post-traitement peut être coûteux lorsque n, m sont grands.

La question maintenant est: quel noyau prendre pour k? Quel observable choisir? i.e. quoi choisir pour X et Y de manière à construire un indice pertinent pour la garantie? Nous présentons des éléments de réponse dans les sections qui suivent.

Appendix A.2. Les propriétés désirées du noyau k: le choix du noyau de Sobolev $r = 1$

Dans cette section, nous expliquons pourquoi nous avons choisi le noyau de Sobolev $r = 1$ pour nos études. La raison avec l'analyse poussée est détaillée dans [7]. Le noyau de Sobolev $r = 1$ est donné par

$$k(x, y) = 1 + B_1(x)B_1(y) + \frac{1}{2}B_2(|x - y|).$$

avec les B_i les polynomes de Bernouilli:

$$B_1(x) = x - \frac{1}{2}, B_2(x) = x^2 - x + \frac{1}{6}.$$

Parmi les propriétés du noyau k de Sobolev $r = 1$ utilisé, nous comptons:

- la symétrie;
- le fait qu'il soit défini positif;
- le noyau est *caractéristique* (pour tout r): ceci veut dire que la classe de fonction dans laquelle les distributions d'intérêt sont transformées est suffisamment riche pour distinguer tous les moments des distributions [1, 7]. Cette proprété est importante pour que les indices permettent de détecter l'indépendence de X et Y;
- Le noyau repecte les hypothèses 2 et 3 de [11] assurant l'existence d'une décomposition ANOVA (permettant une interprétation en terme de pourcentage):

$$\begin{aligned} \mathbb{H}(X, Y) &= \sum_{u \in \mathcal{S}} \mathbb{H}_u, \text{ sera le coefficient de renormalisation et} \\ \forall u \in \mathcal{S}, \mathbb{H}_u &= \sum_{i \in u} (-1)^{|i|-|u|} \mathbb{H}(X_i, Y) \text{ avec } \forall i \in \mathcal{S},\ \mathbb{H}(X_i, Y) = \gamma_k^2(X_i, Y), \end{aligned}$$

où $|i|$ dénote le cardinal de l'ensemble i.

- Le fait de choisir $r = 1$ permet de discerner le plus de moments possible pour un échantillon donné [7].

Maintenant que les indices, ainsi que quelques unes de leurs proprétés, ont été présentés, focalisons sur les observables d'intérêt (i.e. que choisir pour X et Y?).

Appendix A.3. Que choisir pour X et Y pour faire une analyse de sensibilité sur $\mathbb{P}(Y(t,X) > Y_0)$?

On souhaite pouvoir identifier si la probabilité de dépassement de seuil $\mathbb{P}(Y(t,X) > Y_0)$ dépend de X_0, X_1 etc. Supposons les mesures de probabilité de X, Y, X et Y continues, i.e nous avons $\mathrm{d}\mathcal{P}_X(x) = \mathcal{P}_X(x)\mathrm{d}x$, $\mathrm{d}\mathcal{P}_Y(y) = \mathcal{P}_Y(y)\mathrm{d}y$, $\mathrm{d}\mathcal{P}_{X,Y}(x,y) = \mathcal{P}_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$. Les indices HSIC veulent dire "Hilbert Schmidt Independence Criteron" et ont été historiquement appliqués pour tester la distance entre une distribution jointe $\mathcal{P}_{X,Y}$ et le produit des marginales $\mathcal{P}_X, \mathcal{P}_y$. En effet, $X \perp Y \iff \mathcal{P}_{X,Y} = \mathcal{P}_X\mathcal{P}_Y$ et donc $\gamma_k^2(\mathcal{P}_{X,Y}, \mathcal{P}_X\mathcal{P}_Y) = 0$. La puissance de ces indices se révéle donc pour tester statistiquement l'hypothèse d'indépendance entre deux distributions.

Dans le cadre de l'estimation d'une probabilité de dépassement de seuil, leur application [10, 8, 5] consiste à estimer la vraisemblance de l'hypothèse d'indépendance de la variable X_1 (et de tous les groupes de variables d'indices dans $\mathcal{S}$) et de la variable $Z(t,X) = \mathbf{1}_{[Y_0,\infty[}(Y(t,X))$: ainsi, cf. les calculs effectués dans [5, 8]:

$$\forall i \in \mathcal{S}, \mathbb{H}_u = \gamma_k^2(X_i, Z(t,X)) \propto \gamma_k^2(X_i, X|Z(t,X) = 1).$$

Il suffit donc de prendre des produits de noyau de Sobolev $r = 1$ pour chaque composante des n échantillons de X_i et des m échantillons issus de la distribution de $X|Z(t,X) = 1$.

Finalement, dans cette section, tout a été présenté avec des mesures quelconques $\mathrm{d}\mathcal{P}_X$ pour les entrées. En pratique, dans [5], les auteurs mettent en évidence l'importance d'effectuer une transformation iso-probabiliste, i.e. de retransformer toutes les variables $F_X(X_i) = \mathcal{U}_i$ (où F_X est la CDF de X_i) en des uniformes avec $\mathcal{U} = \mathcal{U}(X) = (F_{X_1}(X_1), ..., F_{X_d}(X_d))$ et de définir les indices *via* celles-ci:

$$\forall i \in \mathcal{S}, \mathbb{H}_u = \gamma_k^2(\mathcal{U}_i, Z(t,\mathcal{U})) \propto \gamma_k^2(\mathcal{U}_i, \mathcal{U}|Z(t,\mathcal{U})) = 1).$$

Dans la section qui suit, nous présentons un exemple mettant en évidence l'importance d'effectuer cette transformation.

Appendix B. Importance d'utiliser la transformation iso-probabiliste de [5] pour l'analyse de sensibilité à but orienté

L'importance de la transformation est déjà illustrée dans [5] (exemple 1). L'exemple qui suit peut paraître redondant mais il est également moins abstrait.

Dans cette section, nous reprenons la toute première étude effectuée sur le modèle (1), i.e. en dimension 4 mais au lieu de revenir à une distribution uniforme pour la variable de maillage $X_3 = \Delta x$, nous utilisons directement ses valeurs discrètes.

La figure B.12 présente les résultats obtenus. Tout comme précédemment, la colonne de gauche présente les résultats en prenant en compte tous les maillages et la colonne de droite, les résultats en contraignant les maillages à des valeurs convergées pour l'étude considérée. Tout d'abord, la ligne du haut de la figure

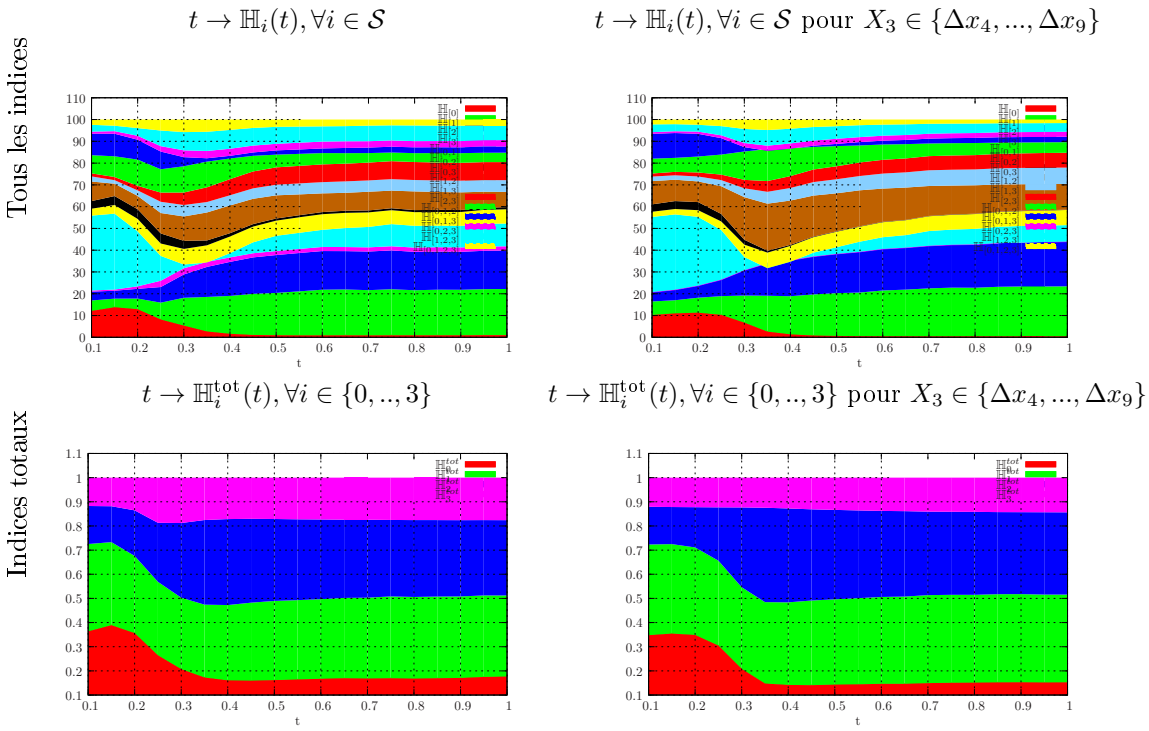

Figure B.12: Étude sans revenir à des distributions uniformes (i.e. sans faire comme dans [5]. Évolution temporelle des indices HSIC appliqué au problème jouet (1). La colonne de gauche présente les indices calculés dans les conditions décrites dans la section 3.1. La colonne de droite présente les indices obtenus lorsque la variable de pas de discrétisation $X_3 = \Delta x$ est conditionnée au maillage plus fin $\{\Delta x_4, .., \Delta x_9\}$.

B.12 présente les pourcentages des contributions de tous les indices de $\mathcal{S}$: dans les deux colonnes, les indices ont l'air de détecter beaucoup d'intéractions entre les variables. Dans ce genre de contexte, la lecture des indices est complexe et il est recommandé de focaliser sur les indices totaux, définis par

$$\forall i \in \{0, ..., d\}, \mathbb{H}_i^{\text{tot}}(t) = \sum_{u \in \mathcal{S}, i \in u} \mathbb{H}_u(t).$$

Ces indices sont présentés dans la seconde ligne de la figure B.12 (ils sont renor-

malisés par $\sum_{i=1}^{d=4} \mathbb{H}_i^{\text{tot}}(t)$ dans la figure, ce qui est peut classique mais pratique ici). Précisons ce que peut capturer l'indice total: pour la variable X_1 par exemple, l'indice total $\mathbb{H}_1^{\text{tot}}$ prend en compte son effet élémentaire $\mathbb{H}_1$ plus les effets de X_1 avec toutes les autres variables, i.e. plus les indices $\mathbb{H}_{1,2}$, $\mathbb{H}_{1,3}$, $\mathbb{H}_{1,4}$, $\mathbb{H}_{1,2,3}$, $\mathbb{H}_{1,3,4}$, $\mathbb{H}_{1,2,4}$, $\mathbb{H}_{1,2,3,4}$. Typiquement, si cet indice $\mathbb{H}_1^{\text{tot}}$ est faible, alors la variable X_1 est sûre d'être négligeable. Si cet indice est élevé, la variable est sûre d'être importante (même si elle l'est peut-être en intéraction, surtout si $\mathbb{H}_1^{\text{tot}} \gg \mathbb{H}_1$). La seconde ligne de la figure B.12 semble dire que dans les deux cas considérés, le maillage joue à peu près le même rôle, *relativement faible*, sur le probabilité de dépassement de seuil... Or le cas a précisément été construit pour que cette variable ait un effet, surtout dans la configuration de la colonne de gauche de la figure B.12! Cet exemple atteste de l'importance de passer par la transformation iso-probabiliste effectuée dans la section 4. Sans elle, les résultats ne sont pas fiables lorsque la variable est discrète notamment. Plus de détails permettant de comprendre ce qu'il se passe sont donnés dans l'exemple 1 de [5].

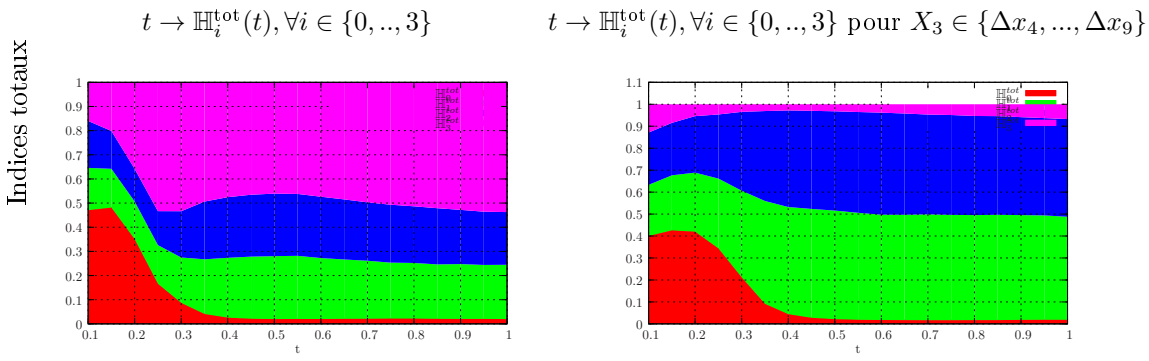

Figure B.13: Indices totaux avec transformation iso-probabiliste. Évolution temporelle des indices HSIC appliqué au problème jouet (1). La colonne de gauche présente les indices calculés dans les conditions décrites dans la section 3.1. La colonne de droite présente les indices obtenus lorsque la variable de pas de discrétisation $X_3 = \Delta x$ est conditionnée au maillage plus fin $\{\Delta x_4, .., \Delta x_9\}$.

Les indices totaux *avec* transformation iso-probabiliste sont présentés figure B.13: *dans la colonne de gauche, les indices totaux permettent de détecter une forte influence du maillage* tandis qu'à droite, l'influence est effectivement amoindrie, dès lors que les maillages sont contraints à des valeurs convergées. Avec la transformation iso-probabiliste, les indices sont fiables.