

Analyse de sensibilité et réduction de dimension. Application à l'océanographie.

Alexandre Janon (Université Joseph Fourier, INRIA)

Soutenance de thèse

15 novembre 2012, Maison Jean Kuntzmann, Grenoble

Plan

- ▶ Partie 1: Contexte et problématique de la thèse
- ▶ Partie 2: Généralités sur l'analyse de sensibilité et la construction de métamodèles
- ▶ Partie 3: Résumé des contributions
- ▶ Conclusion et perspectives

Partie 1: Contexte et Problématique

Contexte de la thèse

- ▶ **Équipe MOISE:** développement de **systèmes de prévision** en **sciences de l'environnement** (météorologie, océanographie, étude des fleuves et des crues, climatologie).
- ▶ Ces systèmes de prévision utilisent des **modèles**, ayant pour objectif de décrire mathématiquement le comportement de fluides géophysiques (océan, atmosphère).

Problématique générale – Erreur de modèle

- ▶ La réalité du système physique étudié n'est pas toujours décrite exactement par le modèle mathématique, et **la prévision est entâchée d'une incertitude.**
- ▶ En effet, le modèle est le fruit des approximations suivantes:
 - ▶ simplifications physiques: négligence de l'effet de certaines forces;
 - ▶ approximations numériques: erreur de discrétisation;
 - ▶ estimation des paramètres du modèle, qui ne sont en général pas connus exactement.
- ▶ Dans cette thèse, nous nous centrons sur l'**incertitude** liée au **troisième type d'approximation.**

Problématique – Notion de sensibilité

- ▶ Nous souhaitons quantifier l'impact de l'incertitude attachée aux paramètres du modèle sur l'état du système prédit par le modèle (ou sur la ou les quantités d'intérêt calculées en sortie).
- ▶ Plus particulièrement, **nous souhaitons identifier les paramètres « sensibles »**.
- ▶ L'**indice de sensibilité** d'un paramètre quantifie l'importance de l'influence de son incertitude sur la sortie. Il s'agit de la part de la variabilité de la sortie expliquée par la variabilité du paramètre.

Problématique – Intérêt de l'analyse de sensibilité

- ▶ L'**analyse de sensibilité** consiste à calculer les indices de sensibilité pour chacun des paramètres d'entrée, ce qui permet de **classer** ces derniers en fonction de leur influence sur la sortie.
- ▶ Les paramètres les plus influents sont ceux sur lesquels l'incertitude doit être réduite – dans la mesure du possible – en priorité afin d'apporter une réduction de l'incertitude sur la sortie la plus importante.
- ▶ À l'inverse, les paramètres les moins influents peuvent être fixés à une valeur nominale, ce qui permet de simplifier le modèle.

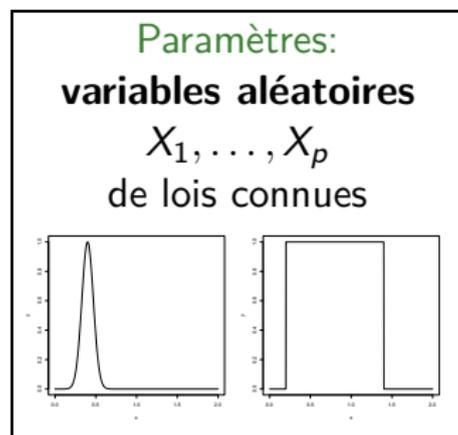
Problématique – Quantification de l'erreur d'estimation

- ▶ **On s'intéresse au calcul effectif des indices de sensibilité.**
- ▶ Il existe plusieurs méthodes permettant de calculer numériquement des valeurs approchées de ces indices, à partir d'un grand échantillon de la quantité d'intérêt, pour différentes valeurs des paramètres d'entrée.
- ▶ Lorsque ces calculs prennent un temps prohibitif, il devient nécessaire de remplacer le modèle original par une approximation, appelée ***métamodèle***, nécessitant moins de temps de calcul pour être résolu.
- ▶ **Comment quantifier l'impact du remplacement du modèle par le métamodèle sur l'estimation des indices de sensibilité ?**
Cette quantification peut faire appel à une *borne d'erreur*, à spécifier, du métamodèle.

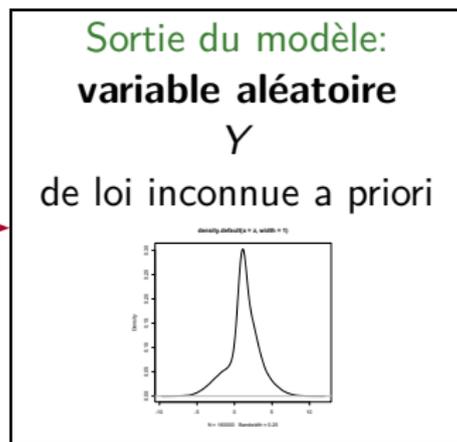
Partie 2: Analyse de sensibilité et construction de métamodèles

- ▶ 2.a. Indices de sensibilité: définition, estimation, estimation de l'erreur
- ▶ 2.b. Métamodèles: généralités, méthode base réduite, borne d'erreur

Quantification d'incertitude [Santner, Cacuci, Saltelli...]



code de calcul



- ▶ $E(Y)$: sortie moyennée sur les valeurs possibles des paramètres, pondérées par leur proba.
- ▶ $\text{Var}(Y)$: dispersion de la sortie autour de cette moyenne lorsque X_1, \dots, X_p varient

Analyse de sensibilité globale [op.cit.]

- ▶ **Hypothèses:** Y scalaire; X_1, \dots, X_p indépendants.
- ▶ Quelle est la contribution de *chaque* paramètre à l'incertitude sur la sortie ?
- ▶ Espérance conditionnelle $\mathbf{E}(Y|X_i)$: variable aléatoire:
 - ▶ ne dépendant que de X_i
 - ▶ donnant, à X_i fixé, la moyenne de Y sur les X_j ($j \neq i$)
 - ▶ meilleure approximation de Y ne dépendant que de X_i .
- ▶ Sa variance $\mathbf{Var}(\mathbf{E}(Y|X_i))$ est un réel quantifiant la dispersion de (la meilleure approx. de) Y lorsque seul X_i varie.
- ▶ En renormalisant, on obtient **l'indice de sensibilité (de Sobol)** de X_i :

$$0 \leq S_i = \frac{\mathbf{Var}(\mathbf{E}(Y|X_i))}{\mathbf{Var} Y} \leq 1$$

c'est la part de variance de Y «expliquée» par l'incertitude sur X_i .

→ Calcul numérique de S_i ?

Estimation de S_i [Sobol, Saltelli]

- ▶ On note $Y = f(\mathbf{X})$ où $\mathbf{X} = (X_1, \dots, X_p)$.
- ▶ On peut écrire:

$$S_i = \frac{\mathbf{Cov}(Y, Y')}{\mathbf{Var}Y},$$

pour $Y' = f(X'_1, \dots, X'_{i-1}, X_i, X'_{i+1}, \dots, X'_p)$, où $\mathbf{X}' = (X'_1, \dots, X'_p)$ est une copie indépendante de \mathbf{X} .

- ▶ Soit $\{\mathbf{X}^k\}_{k=1, \dots, N}$ et $\{\mathbf{X}'^k\}_{k=1, \dots, N}$: deux échantillons iid de la loi de \mathbf{X} .
- ▶ On estime S_i par Monte-Carlo:

$$\hat{S}_i = \frac{\frac{1}{N} \sum_{k=1}^N y_k y'_k - \left(\frac{1}{N} \sum_{k=1}^N y_k \right) \left(\frac{1}{N} \sum_{k=1}^N y'_k \right)}{\frac{1}{N} \sum_{k=1}^N (y_k)^2 - \left(\frac{1}{N} \sum_{k=1}^N y_k \right)^2}$$

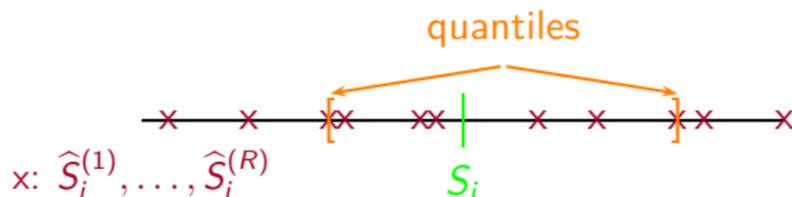
avec $y_k = f(\mathbf{X}^k)$, $y'_k = f(X_1'^k, X_2'^k, \dots, X_{i-1}'^k, X_i^k, X_{i+1}'^k, \dots, X_p'^k)$

- ▶ \hat{S}_i converge p.s. vers S_i quand $N \rightarrow +\infty$.
→ Quantification de l'erreur $|S_i - \hat{S}_i|$?
- ▶ Le calcul de \hat{S}_i nécessite $2N$ évaluations de f .
→ Comment réduire le temps nécessaire à chaque évaluation ?

Quantification de l'erreur Monte-Carlo

Approche "standard"

- ▶ $\hat{S}_i = \hat{S}_i(\mathcal{E})$ où $\mathcal{E} = (\{\mathbf{X}^k\}, \{\mathbf{X}'^k\})$ désigne le couple d'échantillons de la loi de \mathbf{X} choisis aléatoirement.
- ▶ Pour quantifier l'erreur entre \hat{S}_i et S_i , on évalue $\hat{S}_i(\mathcal{E})$ pour plusieurs échantillons indépendants $\mathcal{E}^{(1)}, \dots, \mathcal{E}^{(R)}$.
- ▶ On obtient ainsi un ensemble de répliques $\mathcal{R} = \{\hat{S}_i^{(1)}, \dots, \hat{S}_i^{(R)}\}$ de \hat{S}_i .



- ▶ On en déduit un **intervalle de confiance (approché)** de niveau désiré.

Quantification de l'erreur Monte-Carlo

Approche bootstrap

- ▶ **Problème:** le calcul des R réplifications de \widehat{S}_i nécessite $2N \times R$ évaluations de la fonction f .
- ▶ Dans l'approche bootstrap:

- ▶ on tire un couple d'échantillons:

$$\mathcal{E} = (\{\mathbf{X}^k\}_{k=1,\dots,N}, \{\mathbf{X}'^k\}_{k=1,\dots,N})$$

- ▶ pour $r = 1, \dots, R$, on calcule la $r^{\text{ème}}$ réplification $\widehat{S}_i^{(r)}$ sur le couple de *rééchantillons bootstrap*:

$$\mathcal{E}^{(r)} = (\{\mathbf{X}^k\}_{k \in L_r}, \{\mathbf{X}'^k\}_{k \in L_r})$$

où L_r est une *liste* tirée avec remise dans $\{1, \dots, N\}$;

- ▶ l'ensemble de réplifications est alors utilisé comme précédemment.
- ▶ Le calcul des R réplifications peut s'effectuer à partir des $2N$ évaluations de f sur les points contenus dans \mathcal{E} .
- ▶ Le bootstrap est une approximation de l'approche précédente.

Partie 2: Analyse de sensibilité et construction de métamodèles

- ▶ **2.a.** Indices de sensibilité: définition, estimation, estimation de l'erreur
- ▶ **2.b.** Métamodèles: généralités, méthode base réduite, borne d'erreur

Métamodèles

- ▶ **Accélèrent** le calcul de la variable d'état u pour $K \gg 1$ valeurs du paramètre.
- ▶ Les K appels au code classique sont remplacés par deux phases:

Phase offline: coûteuse, faite une seule fois.

Collecte d' « informations » sur l'ensemble $\{u(\mathbf{X}); \mathbf{X} \in \mathcal{X}\}$ dans lequel « vit » la variable d'état.

Phase online: peu coûteuse, faite K fois.

Pour chaque valeur du paramètre, on utilise les données de la phase offline pour accélérer la résolution approchée de l'EDP.

- ▶ Intéressant si:

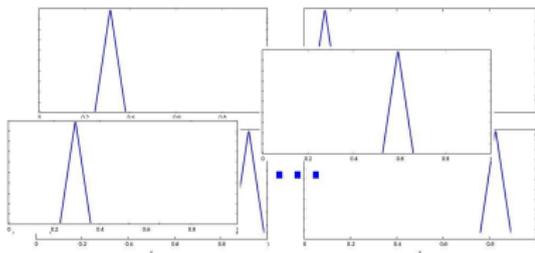
$$\textit{TempsOffline} + K \times \textit{TempsOnline} < K \times \textit{TempsClassique}$$

Métamodèles non intrusifs/intrusifs

- ▶ **Métamodèles non intrusifs:** on dispose seulement d'un échantillon $\{(\mathbf{X}_1, u(\mathbf{X}_1)), \dots, (\mathbf{X}_n, u(\mathbf{X}_n))\}$
 - ▶ \pm Interpolation ou régression:
 - ▶ Krigeage / interpolation par RKHS [Santner, Kleijnen, Schaback].
 - ▶ Décomposition en polynômes de chaos non intrusifs [Sudret-Blatman].
- ▶ **Métamodèles intrusifs:** on travaille sur l'équation satisfaite par la variable d'état.
 - ▶ Métamodèles par bases réduites [Maday-Patera], polynômes de chaos [Ghanem et al.].
 - ▶ *Inconvénient:* on doit connaître et pouvoir traiter cette équation.
 - ▶ *Avantage:* on peut espérer être plus performant.
- ▶ Dans la suite on considère la méthode base réduite.

Métamodèle base réduite: réduction de dimension

Code classique (éléments finis): u est cherché dans un espace de grande dimension, **non adapté** spécifiquement au problème.

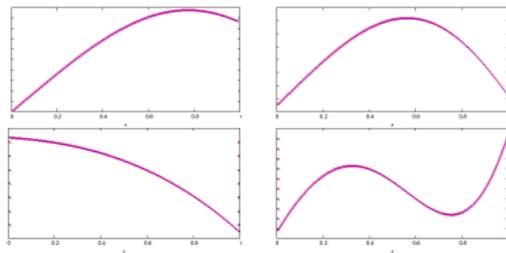


$$u(\mathbf{X}) = \sum_{i=1}^{\mathcal{N}-1} u_i(\mathbf{X}) \phi_i$$

inconnues

Typiquement, en 1D: $\mathcal{N} \simeq 100$, $n \simeq 10$.

Métamodèle: \tilde{u} est cherché dans un **espace de dimension plus petite**, **adapté** au problème.



$$\tilde{u}(\mathbf{X}) = \sum_{i=1}^n \tilde{u}_i(\mathbf{X}) \zeta_i$$

à déterminer
(POD [Sirovitch],
PGD [Nouy],
Greedy [Patera,
Maday, Prud'homme])

inconnues

Borne d'erreur

- ▶ On peut avoir une borne majorant l'erreur entre $u(\mathbf{X})$ et $\tilde{u}(\mathbf{X})$.
- ▶ Borne calculable explicitement, via une procédure offline/online efficace.

- ▶ Se traduit en une borne $\epsilon(\mathbf{X})$ sur l'erreur de la sortie réduite \tilde{Y} :

$$\left| Y(\mathbf{X}) - \tilde{Y}(\mathbf{X}) \right| \leq \epsilon(\mathbf{X}) \quad \forall \mathbf{X}$$

- ▶ Une telle borne d'erreur *ne peut pas* s'obtenir à partir de méthodes non intrusives.

Partie 3: Résumé des contributions

- ▶ 3.a. Métamodèle base réduite pour l'équation de Burgers
- ▶ **3.b.** Estimation d'erreur sortie-dépendante pour la méthode base réduite
- ▶ **3.c.** Intervalles de confiance combinés pour l'estimation des indices de Sobol sur métamodèle
- ▶ **3.d.** Propriétés asymptotiques de l'estimation d'indices de Sobol par Monte-Carlo

Modèle

- Modèle donné par l'équation de Burgers avec viscosité: on cherche u fonction de $t \in [0; T]$ et de $x \in [0; 1]$ telle que:

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{1}{2} \frac{\partial}{\partial x}(u^2) - \nu \frac{\partial^2 u}{\partial x^2} = \psi \\ u(t=0, x) = u_0(x) \quad \forall x \in [0; 1] \\ u(t, x=0) = b_0(t) \\ u(t, x=1) = b_1(t) \end{cases}$$

- Paramétrisation de la condition initiale, du terme source et des conditions de bord:

$$\begin{aligned} b_0(t) &= b_{0m} + \sum_{l=1}^{n(b_0)} A_l^{b_0} \sin(\omega_l^{b_0} t) & b_1(t) &= b_{1m} + \sum_{l=1}^{n(b_1)} A_l^{b_1} \sin(\omega_l^{b_1} t) \\ \psi(t, x) &= f_m + \sum_{l=1}^{n_T(f)} \sum_{p=1}^{n_S(f)} A_{lp}^f \sin(\omega_l^{fT} t) \sin(\omega_p^{fS} x) & u_0(x) &= u_{0m} + \sum_{l=1}^{n(u_0)} A_l^{u_0} \sin(\omega_l^{u_0} x) \end{aligned}$$

Schéma de résolution

- ▶ **Discrétisation en temps:** schéma d'Euler implicite:

$$(\text{Id} - \Delta t \mathcal{F})(u_d^{n+1}) = u_d^n$$

Δt : pas de temps; u_d^1, \dots, u_d^T : solutions approchées à chaque instant discret, \mathcal{F} : opérateur différentiel non linéaire.

- ▶ **Discrétisation en espace:** \mathcal{F} est discrétisé (en un opérateur non linéaire) grâce à une formulation éléments finis \mathbf{P}^1 (Lagrange) et un traitement de la condition de bord par pénalisation. Ainsi les solutions discrètes u^1, \dots, u^T satisfont l'équation non linéaire:

$$(\text{Id} - \Delta t F)(u^{n+1}) = u^n.$$

- ▶ Cette équation non linéaire est résolue grâce à une **itération de Newton**, qui la transforme en une succession de systèmes d'équations linéaires à résoudre.
- ▶ **C'est la résolution de ces systèmes linéaires qui est accélérée à l'aide de la méthode base réduite, pour obtenir $\tilde{u}^1, \dots, \tilde{u}^T$.**

Borne d'erreur

- ▶ On note $u_e^1, \dots, u_e^{\mathcal{T}}$ la solution de la récurrence non-linéaire:

$$(\text{Id} - \Delta t F_e)(u^{n+1}) = u^n$$

où F_e est l'opérateur \mathcal{F} discrétisé à l'aide d'éléments finis \mathbf{P}^1 et *imposition forte de la condition de bord*.

- ▶ **Théorème:** Pour $\Delta t < \Delta t^*$, on a:

$$\forall n = 1, \dots, \mathcal{T}, \quad \|\tilde{u}^n - u_e^n\| \leq \epsilon^n,$$

où

$$\epsilon^n = \mathcal{B}(\epsilon^{n-1}, \|r_k\|_0, \tilde{u}^n(1) - u_e(1), \tilde{u}^n(0) - u_e(0), \mathcal{Q}).$$

- ▶ $\|\cdot\|$ est la norme $L^2(0, 1)$;
- ▶ \mathcal{B} est une fonction connue;
- ▶ \mathcal{Q} sont des quantités connues;
- ▶ r_k est la forme résidu;
- ▶ $\|\cdot\|_0$ est la norme duale de $H_0^1(0, 1)$.

Borne d'erreur

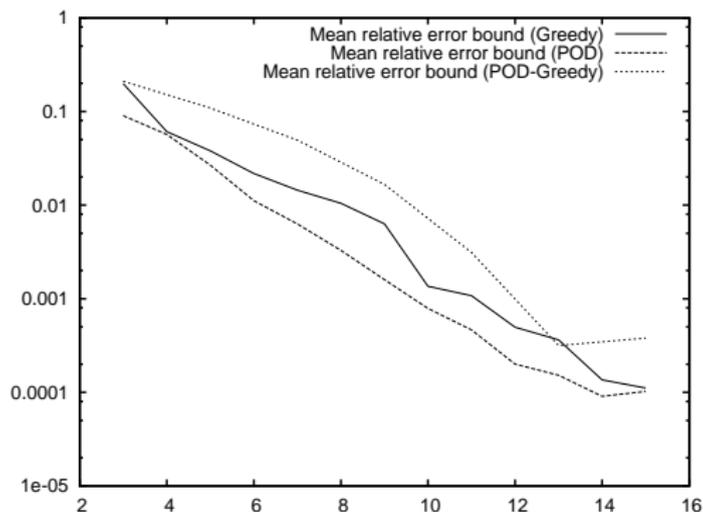
- ▶ *Etapas de preuve:*
 - ▶ estimation d'erreur en norme *énergie*;
 - ▶ passage à la norme $L^2(0, 1)$ (constantes inf-sup);
 - ▶ récurrence en temps par estimation de stabilité du schéma d'Euler.

Notre contribution

- ▶ Nous proposons une **nouvelle borne d'erreur**, reposant sur des majorations plus fines que celles existantes dans la littérature [Veroy-Prud'homme-Patera, Nguyen et al., Jung et al.].
- ▶ Paramétrisation de la condition de bord en utilisant une **forme faible pénalisée**, ce qui diminue la complexité des phases offline et online.

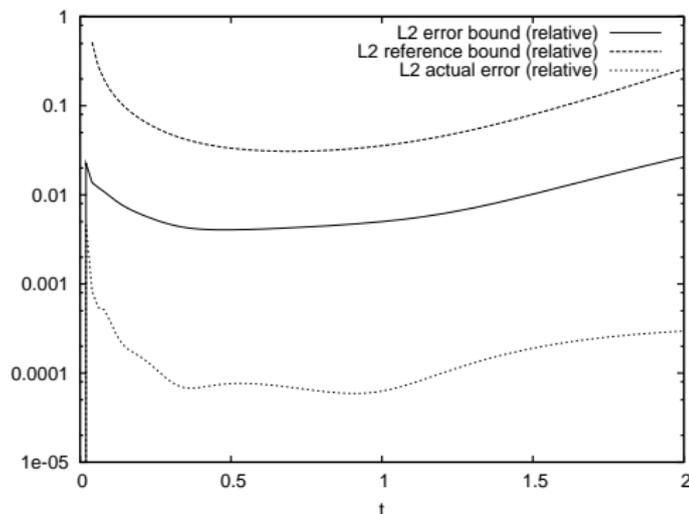
Exemple de résultat obtenu [AJ et al., M2AN, 2012]

Bornes d'erreur relatives moyennes sur un échantillon de paramètres, en fonction de la taille de la base réduite, pour différentes méthodes de choix de base réduite (échelle log.):



Exemple de résultat obtenu [op.cit.]

Erreurs/bornes d'erreurs relatives en fonction du temps (échelle log.):



Points: vraie erreur; plein: borne erreur; pointillés: borne de référence dans la litt.

Partie 3: Résumé des contributions

- ▶ **3.a.** Métamodèle base réduite pour l'équation de Burgers
- ▶ **3.b.** Estimation d'erreur sortie-dépendante pour la méthode base réduite
- ▶ **3.c.** Intervalles de confiance combinés pour l'estimation des indices de Sobol sur métamodèle
- ▶ **3.d.** Propriétés asymptotiques de l'estimation d'indices de Sobol par Monte-Carlo

Estimation d'erreur sur la sortie

► Rappel du contexte:

- u : variable d'état, solution numérique d'une EDP (supposée linéaire);
- \tilde{u} : solution de l'EDP obtenue par base réduite;
- La méthode base réduite donne une borne d'erreur ϵ telle que

$$\|u - \tilde{u}\| \leq \epsilon.$$

► Quantité d'intérêt:

- sur le modèle initial: $f = \ell(u)$ (ℓ : forme linéaire);
- sur le métamodèle: $\tilde{f} = \ell(\tilde{u})$.

► Borne d'erreur sur la quantité d'intérêt:

$$|f - \tilde{f}| \leq L\epsilon,$$

où L : constante de Lipschitz de ℓ .

- Peut-on faire mieux que cette borne sur $|f - \tilde{f}|$?

Estimation d'erreur sur la sortie

- ▶ La borne Lipschitz est optimale parmi les bornes dépendant de $\epsilon \geq \|u - \tilde{u}\|$.
- ▶ Méthode existante (basée sur le problème dual): nécessite de modifier la définition de \tilde{f} et double (en général) les complexités (mémoire et temps) des phases offline et online.
- ▶ Nous proposons une nouvelle borne qui ne présente pas ces deux inconvénients. Elle est basée sur la majoration:

$$\left| f - \tilde{f} \right| \leq \left| \sum_{i=1}^N \underbrace{\langle w, \phi_i \rangle}_{\text{encadrable}} \overbrace{\langle r, \phi_i \rangle}^{\text{calculable en online}} \right| + \underbrace{\left| \sum_{i=N+1}^{\infty} \langle w, \phi_i \rangle \langle r, \phi_i \rangle \right|}_{\text{majorable en probabilité}}$$

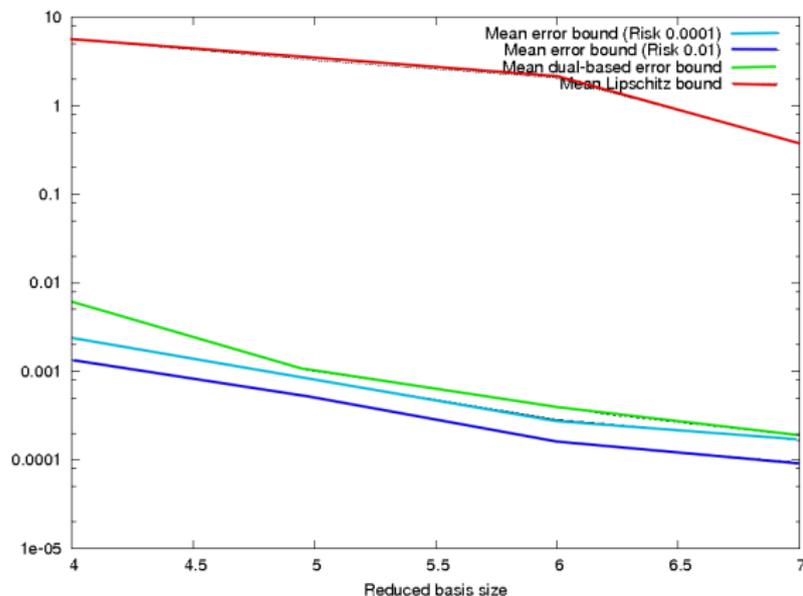
pour toute base orthonormée $\{\phi_i\}$ de l'espace «éléments finis», et tout N . (r : résidu; w : solution du problème dual)

- ▶ On obtient une borne d'erreur avec forte probabilité (au choix de l'utilisateur).

Résultat numérique 1 [AJ et al, soumis, 2012]

EDP sous-jacente: diffusion.

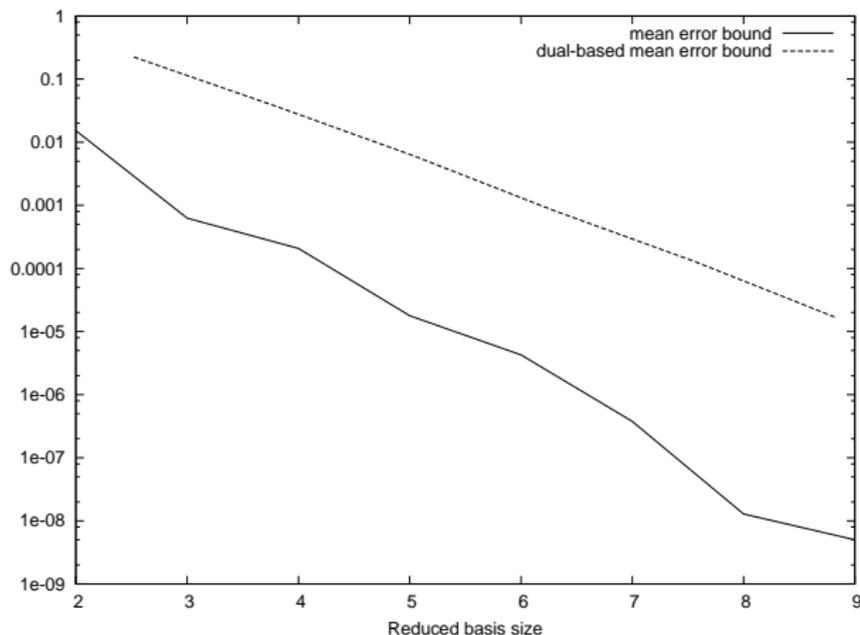
Paramétrisation de la géométrie du domaine (3 paramètres).



Résultat numérique 2

EDP sous-jacente: transport (espace-temps).

1 paramètre (vitesse de transport); risque: 0.0001.



Partie 3: Résumé des contributions

- ▶ **3.a.** Métamodèle base réduite pour l'équation de Burgers
- ▶ **3.b.** Estimation d'erreur sortie-dépendante pour la méthode base réduite
- ▶ **3.c.** Intervalles de confiance combinés pour l'estimation des indices de Sobol sur métamodèle
- ▶ **3.d.** Propriétés asymptotiques de l'estimation d'indices de Sobol par Monte-Carlo

Retour à l'estimation des indices de Sobol

- ▶ On désire tenir compte:
 - ▶ de l'erreur d'échantillonnage
 - ▶ et de l'erreur due au métamodèle
- ▶ On remarque que l'estimateur est fonction des sorties du modèle:

$$\hat{S}_i = \Psi (\{y_k\}_{k=1,\dots,N}, \{y'_k\}_{k=1,\dots,N})$$

- ▶ On a, pour tout k : $y_k \in [\tilde{y}_k - \epsilon_k; \tilde{y}_k + \epsilon_k]$ où $\tilde{y}_k = \tilde{Y}(\mathbf{X}_k)$, $\epsilon_k = \epsilon(\mathbf{X}_k)$; idem avec $'$.
- ▶ D'où un encadrement:

$$\hat{S}_i^m \leq \hat{S}_i \leq \hat{S}_i^M$$

bornes calculables à partir d'appels au métamodèle et à sa borne d'erreur

Retour à l'estimation des indices de Sobol

- ▶ Nous proposons **deux encadrements possibles** de \widehat{S}_i :
 - ▶ Encadrement « analytique »: sans a priori sur la fonction f ; fournit des expressions explicites pour S_i^m et S_i^M .
 - ▶ Encadrement « avec lissage »: pénalise les fonctions f peu régulières; meilleurs résultats en pratique mais nécessite optimisation numérique.
- ▶ On se donne un niveau de risque α .
- ▶ Le **bootstrap (classique)** sur \widehat{S}_i^m et \widehat{S}_i^M donne des intervalles de confiance de risque α :

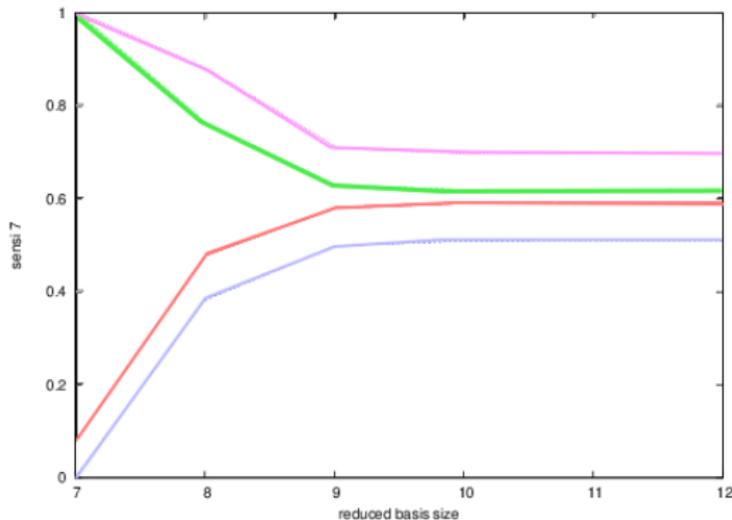
$$[S_{i,\alpha}^{m-}; S_{i,\alpha}^{m+}] \text{ et } [S_{i,\alpha}^{M-}; S_{i,\alpha}^{M+}]$$

- ▶ On montre que:

$$P([S_{i,\alpha}^{m-}; S_{i,\alpha}^{M+}] \ni S_i) \geq 1 - \alpha.$$

Résultats numériques [AJ et al., IJ4UQ, 2012]

- ▶ Modèle donné par l'équation de Burgers, réduite à l'aide de l'approche base réduite que nous avons proposée.
- ▶ Intervalles de confiance obtenus pour un indice de Sobol, pour différentes tailles de bases réduites, à taille d'échantillon fixée :



Légende: $\hat{S}_i^{m-} < \hat{S}_i^m < \hat{S}_i^M < \hat{S}_i^{M+}$

Gains en temps de calcul

- ▶ Implémentation C++; Contribution au package R *sensitivity*.
- ▶ **Comparaison avec l'estimation avec le code classique:**
gain d'un **facteur 5 à 6** en temps de calcul, à précision égale.
- ▶ **Comparaison avec une approche non intrusive:**
résultat **plus précis**, en un **temps plus court**.
- ▶ **On a tiré parti:**
 - ▶ des **propriétés du modèle**
 - ▶ du **travail théorique** nécessaire à la construction du métamodèle et de la borne d'erreur

Partie 3: Résumé des contributions

- ▶ **3.a.** Métamodèle base réduite pour l'équation de Burgers
- ▶ **3.b.** Estimation d'erreur sortie-dépendante pour la méthode base réduite
- ▶ **3.c.** Intervalles de confiance combinés pour l'estimation des indices de Sobol sur métamodèle
- ▶ **3.d.** Propriétés asymptotiques de l'estimation d'indices de Sobol par Monte-Carlo

Propriétés asymptotiques

- ▶ On souhaite étudier le comportement de \widehat{S}_i lorsque $N \rightarrow +\infty$.
- ▶ Dans un premier temps: utilisation du modèle exact.
 - ▶ **Thm.** Si $Y \in L^4$, alors il existe $\sigma_S^2 \geq 0$ tel que:

$$\sqrt{N}(\widehat{S}_i - S_i) \underset{N \rightarrow \infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \sigma_S^2).$$

- ▶ *Preuve.* Utilisation de la δ -méthode.
- ▶ σ_S^2 est estimable (connu en fonction des moments de Y et de Y').
- ▶ Ceci permet de construire (lorsque le vrai modèle est peu coûteux), de manière justifiée, des intervalles de confiance asymptotiques pour S_i .
- ▶ Peut-on construire un estimateur "meilleur" que \widehat{S}_i en termes de variance asymptotique ?

Un deuxième estimateur de S_j

- ▶ Idée: utiliser les deux échantillons de sorties pour estimer $\mathbf{E}(Y)$ et $\mathbf{E}(Y^2)$.
- ▶ Ceci donne lieu à l'estimateur suivant:

$$\widehat{T}_i = \frac{\frac{1}{N} \sum_{k=1}^N Y_k Y'_k - \left(\frac{1}{N} \sum_{k=1}^N \left[\frac{Y_k + Y'_k}{2} \right] \right)^2}{\frac{1}{N} \sum_{k=1}^N \left[\frac{Y_k^2 + (Y'_k)^2}{2} \right] - \left(\frac{1}{N} \sum_{k=1}^N \left[\frac{Y_k + Y'_k}{2} \right] \right)^2}.$$

- ▶ **Thm.** Si $Y \in L^4$, il existe $\sigma_T^2 \geq \sigma_S^2$ tel que:

$$\sqrt{N} \left(\widehat{T}_i - S_i \right) \xrightarrow[N \rightarrow \infty]{\mathcal{L}} \mathcal{N} \left(0, \sigma_T^2 \right)$$

avec σ_T^2 exprimable en fonction des moments de Y et de Y' ,

- ▶ Remarque: on a $\sigma_S^2 = \sigma_T^2$ ssi $S_i \in \{0, 1\}$.

Efficacité asymptotique de \widehat{T}_j

- ▶ **Thm.** \widehat{T}_j est *asymptotiquement efficace*, i.e. est de variance asymptotique minimale parmi les estimateurs "réguliers" de S_j .
- ▶ *Schéma de preuve.*
 - ▶ On identifie un estimateur asymptotiquement efficace de $(\mathbf{E}(Y), \mathbf{E}(Y^2), \mathbf{E}(YY'))$ fonction des vecteurs observés $(Y_k, Y_k')_{k=1, \dots, N}$ indépendants échangeables.
 - ▶ On applique une fonction différentiable, qui préserve l'efficacité asymptotique.

Etude asymptotique: utilisation d'un métamodèle

- ▶ On considère les estimateurs \widehat{S}_i et \widehat{T}_i .
- ▶ Convergence de $\widehat{S}_i - S_i$ (ou $\widehat{T}_i - S_i$) lorsque $N \rightarrow +\infty$?
- ▶ Comme \widehat{S}_i (et \widehat{T}_i) tendent vers \widetilde{S}_i , il n'y a pas consistance (sauf si $\widetilde{S}_i = S_i \dots$).
- ▶ Il faut donc supposer que *le métamodèle est construit en fonction de N et que sa précision augmente avec N .*

Etude asymptotique: utilisation d'un métamodèle

- ▶ On suppose que:

$$\tilde{Y} = Y + \epsilon_N$$

- ▶ **Thm.** Sous des hypothèses techniques de domination,

$$\sqrt{N}(\widehat{S}_i - S_i) \underset{N \rightarrow \infty}{\overset{\mathcal{L}}{\rightarrow}} \mathcal{N}(0, \sigma_S^2) \quad \text{et} \quad \sqrt{N}(\widehat{T}_i - S_i) \underset{N \rightarrow \infty}{\overset{\mathcal{L}}{\rightarrow}} \mathcal{N}(0, \sigma_T^2)$$

si et seulement si:

$$\mathbf{Var} \epsilon_N = o(1/N).$$

- ▶ *Outil de preuve:* TCL triangulaire (Lindeberg-Feller).
- ▶ Sous d'autres hypothèses techniques, \widehat{T}_i est asymptotiquement efficace pour S_i sous la même condition.

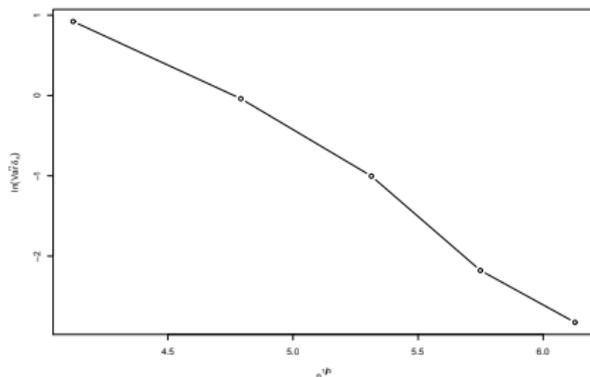
Exemple numérique [AJ et al, soumis, 2012]

- ▶ *Vrai modèle*: fonction d'Ishigami: (X_1, X_2, X_3) iid uniformes sur $[-\pi; \pi]^3$, $p = 3$,

$$f(X_1, X_2, X_3) = \sin X_1 + 7 \sin^2 X_2 + 0.1 X_3^4 \sin X_1,$$

indices connus analytiquement.

- ▶ *Métamodèle*: RKHS; taille d'échantillon d'apprentissage n .
- ▶ Estimation de $\ln \mathbf{Var}(\epsilon)$ en fonction de $n^{1/3}$:



Exemple numérique

- ▶ On prend $n = (a \log N)^3$ avec a paramètre libre.
- ▶ Prédiction: normalité asymptotique ssi $a > 0.52$.
- ▶ Estimations des coverages des IC asymptotiques (sur 100 réplifications):

a	N	n	Coverage S_1	Cov. S_2	Cov. S_3
.4	3000	33	0.1	0	0.7
.4	4000	37	0.08	0	0.78
.4	6000	43	0.26	0.3	0.88
.4	10000	51	0.28	0.18	0.78
.4	20000	77	0.28	0.1	0.59
.6	3000	111	0.79	0.37	0.9
.6	4000	124	0.8	0.7	0.94
.6	10000	169	0.92	0.82	0.94
.6	20000	210	0.93	0.85	0.95

(niveau théorique 0.95)

Conclusion et perspectives



« Bien joué Peter, l'enquête touche à sa fin ! »

Conclusion et perspectives



« Bien joué Peter, l'enquête touche à sa fin ! »

La Classe Américaine

Conclusion

- ▶ La quantification d'incertitude et l'analyse de sensibilité requièrent un **grand nombre** d'appels au code à analyser.
- ▶ L'utilisation d'un **métamodèle** permet de diminuer le temps de calcul nécessaire à chaque appel, au prix d'une certaine approximation.
- ▶ Nous disposons d'une approche permettant de **quantifier** précisément cette approximation afin de:
 - ▶ **garantir** l'estimation numérique fournie
 - ▶ **choisir** de manière optimale les **paramètres** de l'estimation
- ▶ Indépendamment de cette quantification, nous disposons de **conditions nécessaires et suffisantes** pour que l'utilisation d'un métamodèle en analyse de sensibilité soit **justifiée asymptotiquement**.

Travaux en cours et perspectives

- ▶ Appliquer la méthodologie sur d'**autres modèles** et/ou d'autres types de métamodèles (Shallow-Water par ex.).
- ▶ Choisir la base réduite en fonction de la quantité d'intérêt.
- ▶ Poursuivre l'étude asymptotique pour les indices de Sobol généralisés portant sur plusieurs variables d'entrée.
- ▶ Cadre multifidélité: utiliser plusieurs métamodèles (de précisions et de temps de calcul différents) pour améliorer l'efficacité de l'estimation des indices de Sobol.