Analyse de sensibilité par HSIC pour codes stochastiques

Application à la gestion d'actifs hydrauliques

Julien Pelamatti & Jérôme Lonchampt
Projet AMPH2
EDF R&D

Contexte : parc hydraulique français

Quelques chiffres (2019)

- $\sim$ 60 TWh de production
  - 11 % de la production du pays
- 2300 installations hydroélectriques différentes (433 exploitées par EDF)
  - Barrages
  - Stations de Transfert d'Energie par Pompage (STEP)
  - Turbines *au fil de l'eau*
- Production concentrée en :
  - Auvergne Rhône-Alpes
  - Occitanie
  - Provence-Alpes-Côte d'Azur
  - Grand-Est

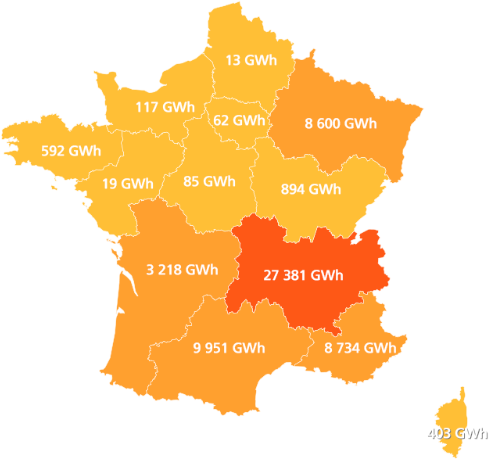


Figure: Source : RTE (2019)

Contexte : maintenance du parc

- Optimisation et justification des stratégies de maintenance du parc hydraulique
- Analyse de la rentabilité technico-économique et planification des opérations de maintenance :
  - Inspection programmée ou déclenchée par des mesures/capteurs
  - Maintenance préventive
  - Maintenance corrective (suite à une défaillance)
  - Approvisionnement de pièces de rechange (préventif ou suite à des maintenances)
- On s'intéressera ici aux équipements suivants :
  - Alternateurs
  - Turbines
  - Transformateurs

Contexte : maintenance du parc

- Difficultés
  - Un parc d'actifs large et varié, en termes de dimension, technologie et âge
    - Difficile de prédire l'usure des différents composants
    - Difficile de prédire les temps et les coûts de maintenance
  - Disponibilité et approvisionnement des pièces de rechange
    - Temps & coûts
  - Disponibilité de main d'œuvre
- Le phénomène est de nature aléatoire
  - Les dates de défaillance des composants ne sont pas connues d'avance
  - L'état des composants après inspection n'est pas connu d'avance
  - Possibilité d'avoir des faux positifs & négatifs sur les inspections

Il n'existe pas une stratégie de maintenance optimale *à priori*

Le coût d'une stratégie de maintenance peut se voir comme une variable aléatoire

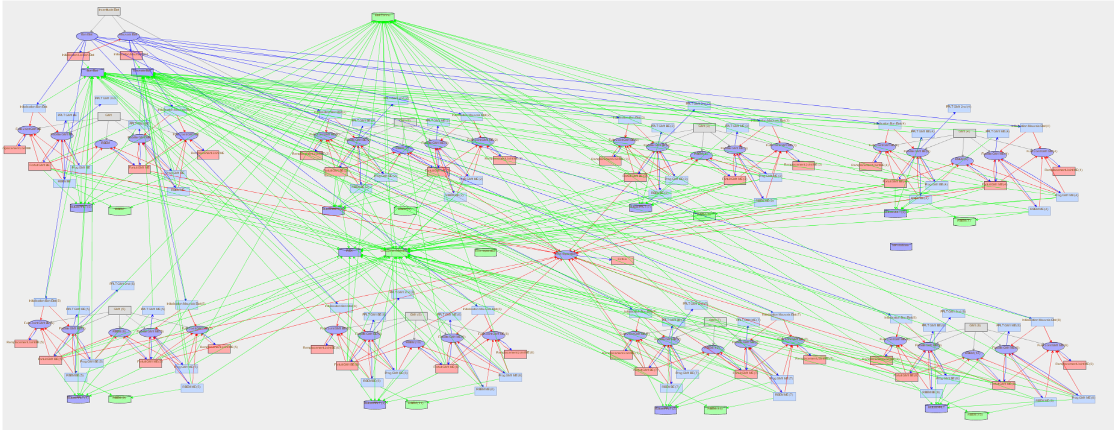
Contexte : simulation de stratégies de maintenance

Simulation des stratégies de maintenance à l'aide du logiciel **VME** (EDF R&D) (Valorisation des Maintenances Exceptionnelles)

- Modélisation d'actifs ou de parcs d'actifs, et de leurs composants
- Modes de défaillances & stratégies de maintenance complexes (maintenance préventive, systématique, conditionnelle, ...)
- Logistique de pièces de rechange (temps et coûts d'approvisionnement), mise en commun de pièces pour des équipements du même type
- Variabilité saisonnière des coûts des pertes associées aux indisponibilités

Contexte : simulation de stratégies de maintenance

Exemple de modèle VME : *n* groupes identiques



Contexte : simulation de stratégies de maintenance

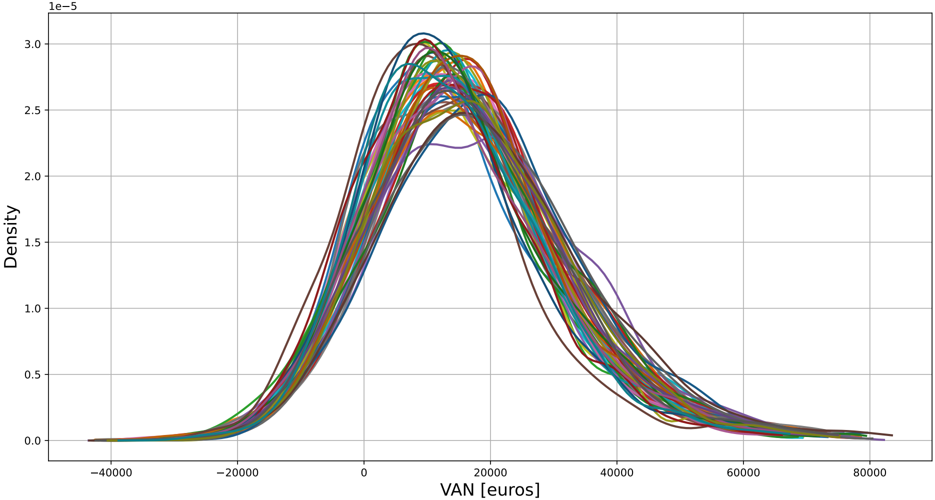
- Les dates de défaillance, les erreurs d'inspections, et autres informations méconnues sont générées aléatoirement
- Deux évaluations d'une simulation avec les mêmes paramètres d'entrée pourront fournir des résultats différents
  - **Code stochastique**
- Comparaison avec une stratégie de référence, typiquement purement corrective
  - Est-il préférable d'effectuer des maintenances préventives, ou bien est-ce moins coûteux d'effectuer les réparations suite aux défaillances ?
  - Est-il convénient d'approvisionner une pièce de rechange avant épuisement du stock ?
- On s'intéresse à la Valeur Actuelle Nette (VAN) d'une stratégie de maintenance
  - Gains (ou pertes) économique par rapport à la stratégie de référence

Contexte : simulation de stratégies de maintenance

- Évaluation des stratégies par échantillonnage Montecarlo
  - Chaque simulation est effectuée une large nombre de fois, afin de pouvoir estimer la distirbution de la VAN
- Comment analyser les résultats?
  - Distribution entière
  - Moyenne / Médiane
  - Quantiles
  - Probabilité de VAN négative (probabilité de regret)
- Coût de calcul potentiellement élevé
  - Large nombre de simulations MC = meilleure estimation de la distribution de la VAN

Contexte : simulation de stratégies de maintenance

Exemple de modèle VME : n groupes identiques (50 calculs, 1000 simulations par calcul)



Analyse de sensibilité pour codes stochastiques

Analyse de sensibilité

- L'évaluation d'une stratégie de maintenance dépend de plusieurs paramètres :
  - Types & paramètres de lois de défaillance des composants
  - Coûts des maintenances et des composants
  - Durées des maintenances & des approvisionnements
  - Pertes de production (indisponibilités)
  - État des composants suite aux maintenances
- Besoin d'une analyse de sensibilité de la distribution de la VAN par rapport aux différents paramètres caractérisant la simulation
  - Identification des paramètres de simulation influents sur la distribution de la VAN, afin de simplifier le modèle
  - Les paramètres les plus influents pourront faire l'objet d'une étude plus approfondie afin d'en valider les valeurs fournies

Rappels : indices HSIC pour l'analyse de sensibilité

Indices HSIC

- Quantification de la dépendance entre deux variables aléatoires X et Y
- *Distance* dans un RKHS entre les *projections* de la loi jointe $\mathcal{P}_{XY}$ et le produit des lois marginales $\mathcal{P}_X\mathcal{P}_Y$

$$HSIC(X,Y) = MMD^2(\mathcal{P}_{XY}, \mathcal{P}_X\mathcal{P}_Y)$$

  - $HSIC(X,Y) = 0 \longleftrightarrow$ indépendance entre X et Y (sous conditions)
- Estimation à l'aide de noyaux de covariance, $k_x(\cdot,\cdot)$, $k_y(\cdot,\cdot)$

$$\widehat{HSIC}(X,Y)_{Vstat} = \frac{1}{n^2}\mathrm{Tr}(K_x H K_Y H)$$

  - K_x, K_y matrices de covariance, H matrice de centrage

Rappels : indices HSIC pour l'analyse de sensibilité

Indices HSIC

- On s'intéresse ici aux HSIC calculés indépendamment pour chaque paramètre d'entrée
  - pas de HSIC-ANOVA

- On utilisera un indice normalisé (entre 0 et 1) :

$$\widehat{R^2}_{HSIC(X,Y)} = \frac{\widehat{HSIC}(X,Y)}{\sqrt{\widehat{HSIC}(X,X)\ \widehat{HSIC}(Y,Y)}}.$$

Rappels : indices HSIC pour l'analyse de sensibilité

Indices HSIC ciblés

- On peut exploiter les indices HSIC afin de faire de l'analyse de sensibilité **ciblée**
  - Quels sont les paramètres influents pour le franchissement dans un domaine critique $\mathcal{C}$ de la sortie?
  - Exemple : on veut savoir quels sont les paramètres qui peuvent mener à des valeurs moyennes de VAN négatives
- On peut appliquer une transformation sur la variable de sortie, au plus simple :

$$\tilde{Y} = w(Y) = \begin{cases} 0 \text{ } if \text{ } Y \notin \mathcal{C} \\ 1 \text{ } if \text{ } Y \in \mathcal{C} \end{cases}$$

- Des fonctions de poids $w(\cdot)$ plus pertinentes existent
- Nécessite l'utilisation d'un noyau $k_y(\cdot,\cdot)$ adapté
- On calcule ensuite $HSIC(X, \tilde{Y})$ avec des estimateurs standards

Rappels : indices HSIC pour l'analyse de sensibilité

p-valeurs pour les indices HSIC

- On peut exploiter les indices HSIC afin de faire du *criblage*
  - Caractérisation qualitative des variables influentes / non-influentes
- On exploite des tests statistiques sur l'hypothèse d'indépendance $\mathcal{H}_0$ entre X et Y
- On s'intéresse à la statistique $\widehat{S}_{\mathcal{T}} := n \times \widehat{HSIC}(X, Y)$
- On peut calculer la p-valeur associée

$$p_{\text{val}} = \mathbb{P}\left(\widehat{S}_{\mathcal{T}} > \widehat{S}_{\mathcal{T},\text{obs}} \mid \mathcal{H}_0\right)$$

  - Faible valeur de $p_{\text{val}} \to$ on peut sereinement rejeter $\mathcal{H}_0$
- On exploite une approximation asymptotique de la loi de $\widehat{HSIC}(X, Y)$ (loi Gamma) pour calculer p_{val}

Analyse de sensibilité pour codes stochastiques

- On dispose d'un jeu de données contenant entrées et échantillons MC de la sortie

$$\left(\mathbf{x}^{(j)}, \mathbf{y}^{(j)}\right)_{(1\leq j\leq n)} = \left(x_1^{(j)}, x_2^{(j)}, \ldots, x_d^{(j)}; y_1^{(j)}, \ldots, y_p^{(j)}\right)_{(1\leq j\leq n)}$$

Comment utiliser les indices HSIC dans ce contexte ?

- Sur la valeur moyenne de y ?
- Sur la variance de y ?
- HSIC *ciblés* sur des quantiles de y ?

Analyse de sensibilité pour codes stochastiques

Exemple :

$$y(x_1, x_2) = x_1 \sin(x_1) + x_2 \mathcal{N}(0, 1)$$

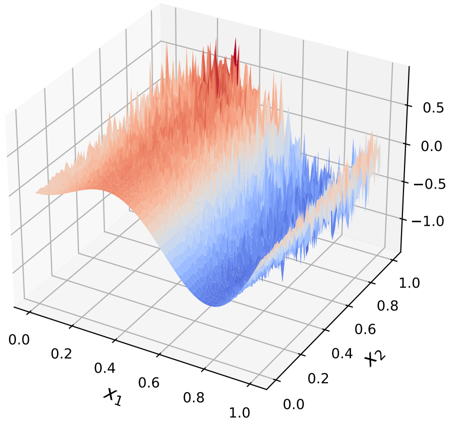
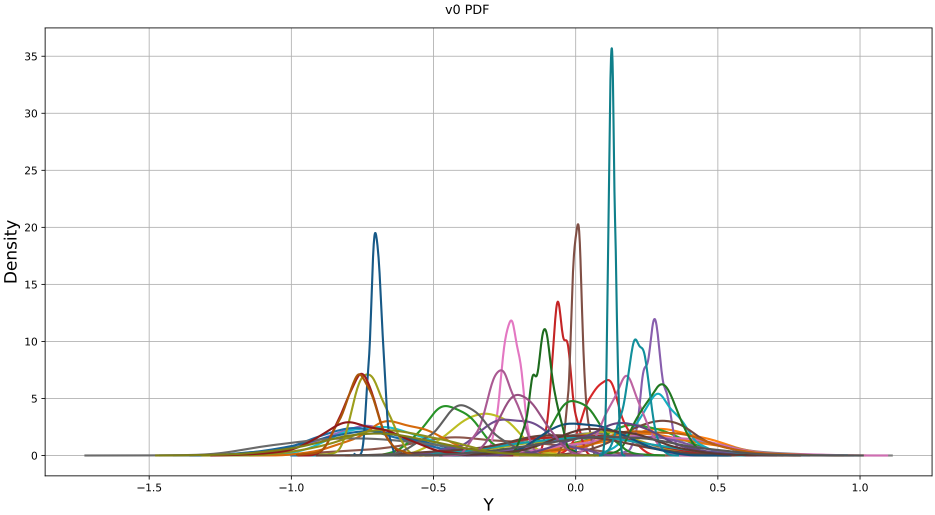


Figure: Réalisation de y

Analyse de sensibilité pour codes stochastiques

Exemple (50 calculs, 300 simulations par calcul)



Analyse de sensibilité pour codes stochastiques

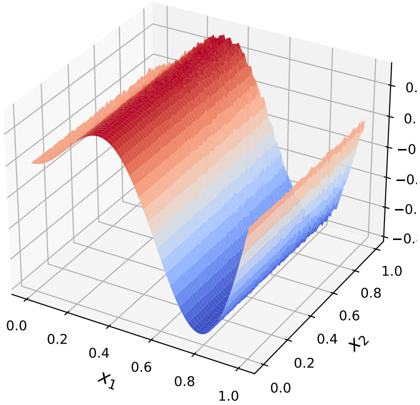


Figure: Moyenne empirique

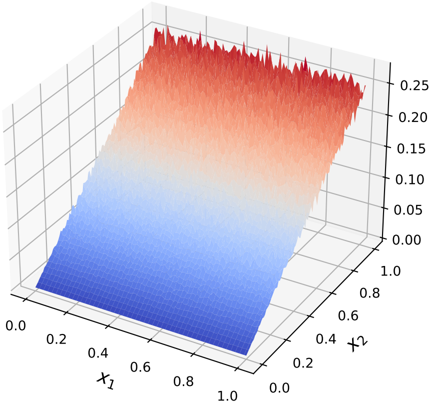
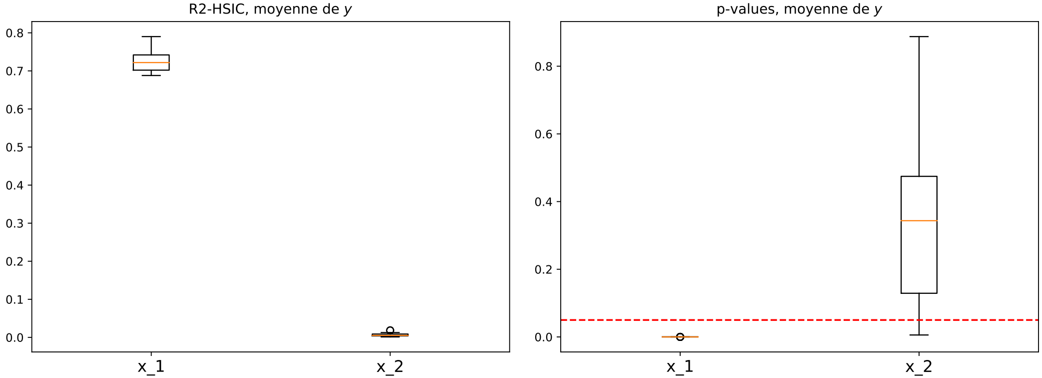


Figure: Variance empirique

Analyse de sensibilité pour codes stochastiques

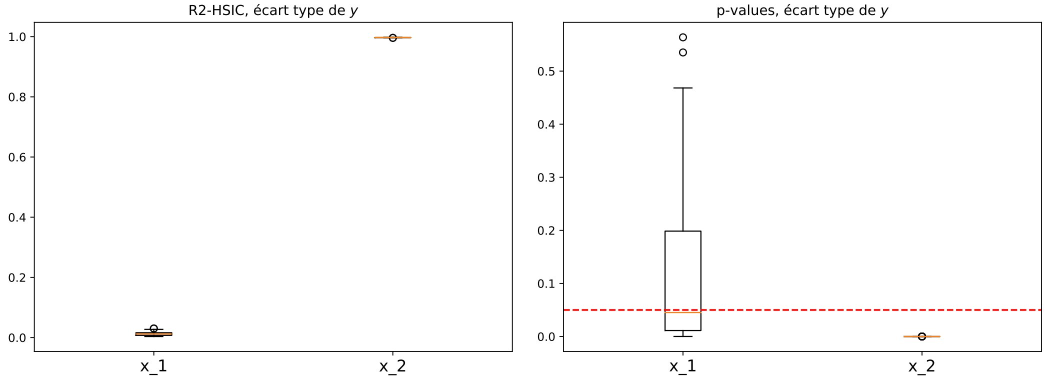
Calcul des indices HSIC entre x_1, x_2 et la moyenne (empirique) de y



On ne capte pas l'effet de x_2 :(

Analyse de sensibilité pour codes stochastiques

Et si on regarde les HSIC entre x_1, x_2 et l'écart type (empirique) de y ?



On ne capte pas l'effet de x_1 :(

Analyse de sensibilité pour codes stochastiques

On aimerait pouvoir analyser la distribution de y dans son entièreté

- Peut-on adapter les indices HSIC au cadre stochastique ?
- Comment définir un noyau de façon à ce qu'il calcule une covariance entre deux échantillons de y ?

$$k(\mathbf{y}, \mathbf{y}') = k(y_1, \dots, y_p \ ; \ y'_1 \dots, y'_{p'})$$

- On peut remplacer $||y - y'||_2^2$ (cas déterministe) par $MMD^2(\mathbf{y}, \mathbf{y}')$, e.g.,

$$k(\mathbf{y}, \mathbf{y}') = \exp\left(\frac{MMD^2(\mathbf{y}, \mathbf{y}')}{\sigma^2}\right)$$

Quelques rappels sur la MMD

- $Y \sim \mathcal{P}_Y$, $Y' \sim \mathcal{P}_{Y'}$ deux variables aléatoires
- kernel embedding de Y, Y' dans un RKHS $\mathcal{H}$:

$$\mu_Y = \mathbb{E}_Y[k_{\mathcal{H}}(Y, \cdot)], \qquad \mu_{Y'} = \mathbb{E}_{Y'}[k_{\mathcal{H}}(Y', \cdot)]$$

- La MMD entre Y et Y' est définie comme :

$$MMD(Y, Y') = ||\mu_Y - \mu_{Y'}||_{\mathcal{H}}$$

- Estimation :

$$\begin{aligned} MMD^2(Y, Y') &= \mathbb{E}[k_{\mathcal{H}}(Y, \tilde{Y})] + \mathbb{E}[k_{\mathcal{H}}(Y', \tilde{Y}')] - 2\mathbb{E}[k_{\mathcal{H}}(Y, Y')] \\ \widehat{MMD^2}(Y, Y') &= \frac{1}{p(p-1)} \sum_{i=1}^{p} \sum_{j \neq 1}^{p} k_{\mathcal{H}}(Y_i, Y_j) + \frac{1}{p'(p'-1)} \sum_{i=1}^{p'} \sum_{j \neq 1}^{p'} k_{\mathcal{H}}(Y'_i, Y'_j) - \\ & \quad \frac{2}{pp'} \sum_{i=1}^{p} \sum_{j=1}^{p'} k_{\mathcal{H}}(Y_i, Y'_j) \end{aligned}$$

Noyau de covariance *energy distance*

Afin de calculer la MMD entre deux échantillons de Y et Y', on utilisera ici le noyau dit *energy distance* :

$$k(y, y') = |y| + |y'| - 2|y - y'|$$

A noter :

- Non-paramétré
- Non stationnaire
- Caractéristique

NB : on normalise les données entre 0 et 1

Noyau de covariance *energy distance*

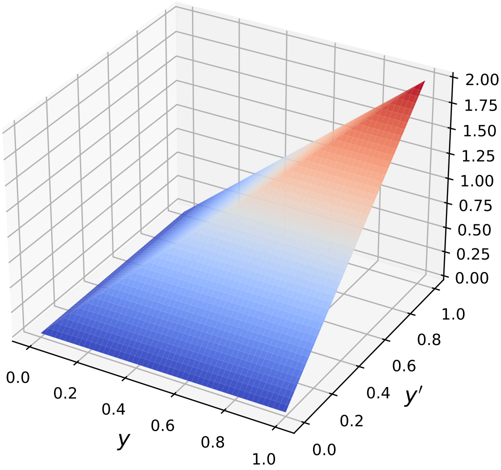


Figure: Noyau energy distance

Analyse de sensibilité pour codes stochastiques

- Une fois défini le noyau entre échantillons de y, il est possible de calculer et exploiter les indices HSIC entre chaque paramètre et la sortie stochastique de façon standard
  - Estimateurs Vstat & Ustat
  - P-valeurs (estimateurs asymptotiques et par permutation)
- NB : il est possible de calculer la MMD entre deux échantillons de tailles p et p' différentes
- Coût de calcul considérablement plus large que pour les analyses sur des sorties déterministes

Analyse de sensibilité pour codes stochastiques

On reprend l'exemple précédent :

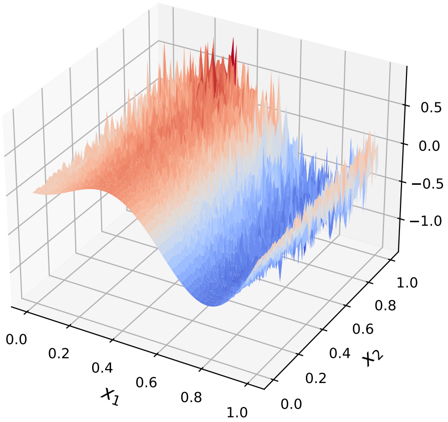
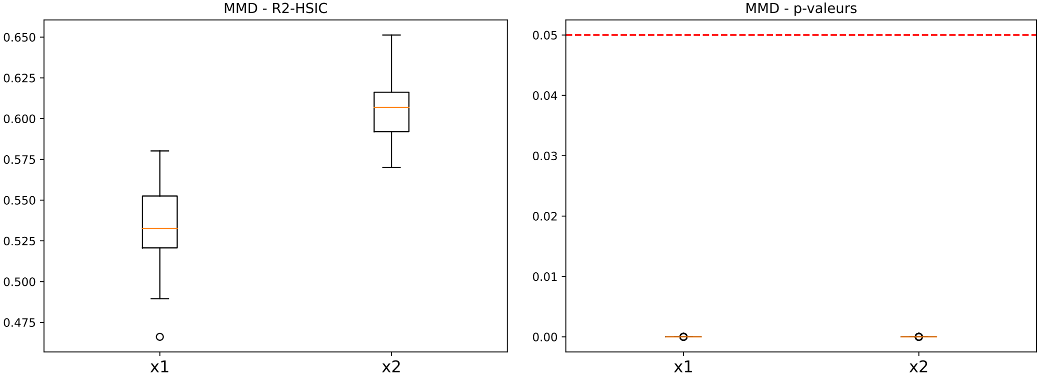


Figure: Réalisation de y

Analyse de sensibilité pour codes stochastiques

Et calcule les HSIC entre x_1, x_2 et les échantillons des y correspondants



On capte les effets des deux variables :)

Analyse de sensibilité, application aux modèles VME

Utilisation des indices HSIC dans le cadre de VME

On compare ici trois types d'analyses différentes :

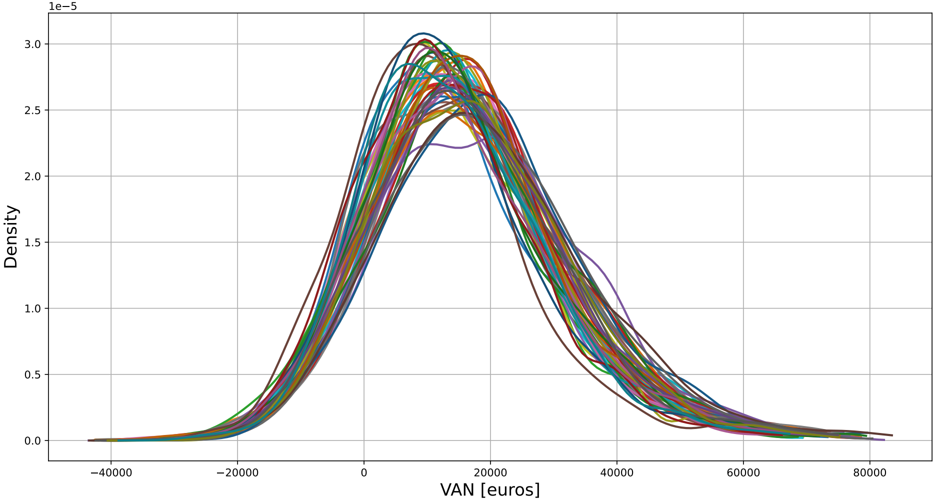
1. Analyse de sensibilité globale sur un scalaire représentatif
   - VAN moyenne
2. Analyse de sensibilité ciblée sur un seuil critique
   - e.g., VAN moyenne inférieure à 0 (investissement à perte)
3. Analyse de sensibilité globale sur la distribution de la VAN
   - Utilisation du noyaux basé sur l'utilisation de la MMD entre échantillons

Cas d'application : n groupes identiques

- On considère n groupes identiques
- 4 co-variables, distributions uniformes
  - Paramètres λ et β de la loi de Weibull de défaillance des groupes
  - Incertitude sur les coûts de maintenance
  - Incertitude sur les temps de maintenance
- 3 analyses de sensibilité :
  - Analyse de sensibilité globale sur la VAN moyenne
  - Analyse de sensibilité ciblée sur la probabilité d'une VAN négative : $P(VAN < 0.) < 0.15$
  - Analyse de sensibilité sur la distribution de la VAN
- 1000 simulations Montecarlo par jeu de paramètres
- 500 jeux de paramètres considérés
- Intervalle de confiance calculé sur 50 re-échantillonnages bootstrap

Contexte : simulation de stratégies de maintenance

Cas d'application : n groupes identiques (50 calculs, 1000 simulations par calcul)



Cas d'application : n groupes identiques

Indices R2-HSIC

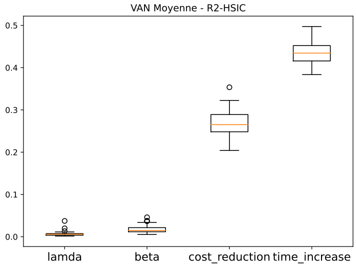


Figure: Global

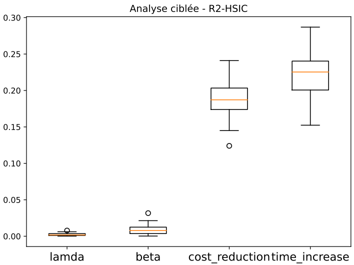


Figure: Target

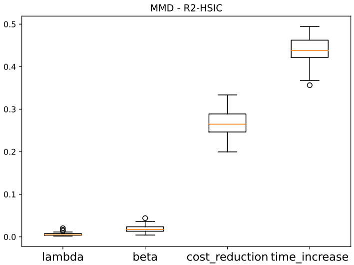


Figure: MMD

Cas d'application : n groupes identiques

P-valeurs

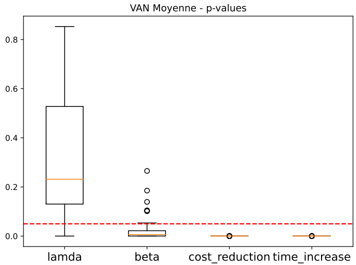


Figure: Global

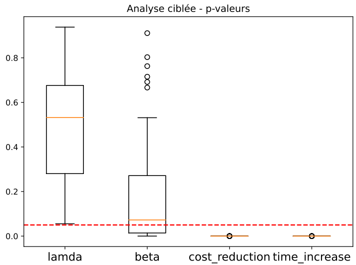


Figure: Target

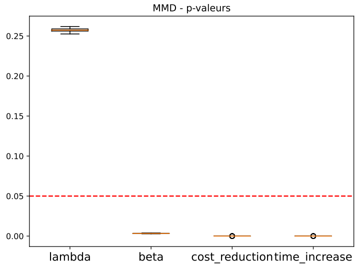
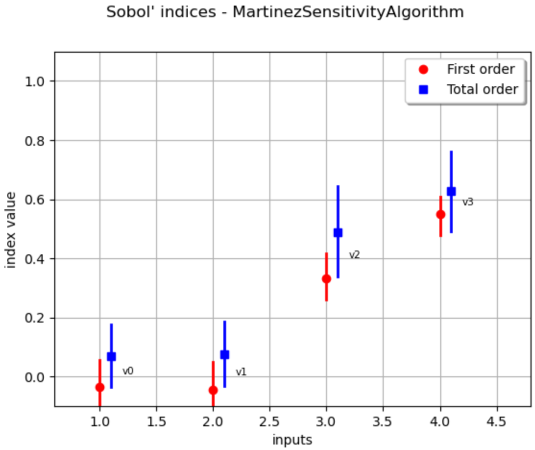


Figure: MMD

Cas d'application : n groupes identiques

Indices de Sobol (1er ordre et total) sur la VAN moyenne

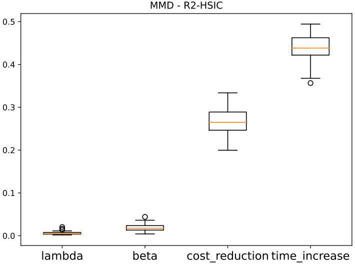


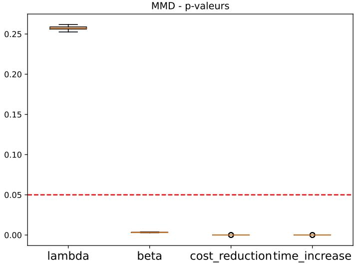
- Nécessite considérablement plus de données ($\sim 10\times$)

Cas d'application : n groupes identiques

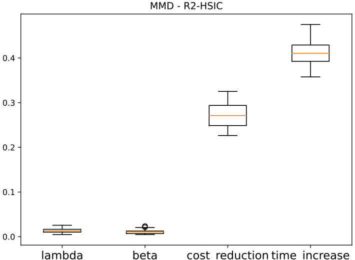
Quelle est l'influence de la distribution associée à chaque paramètre ?

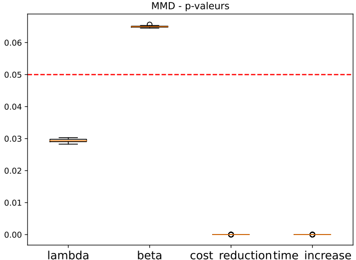
Distribution uniforme





Distribution triangulaire

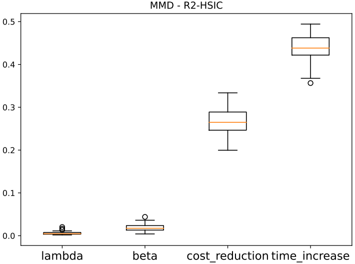


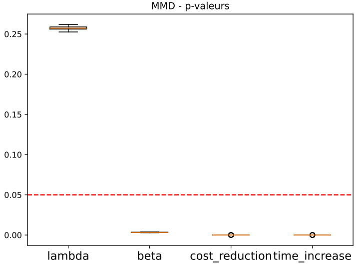


Cas d'application : n groupes identiques

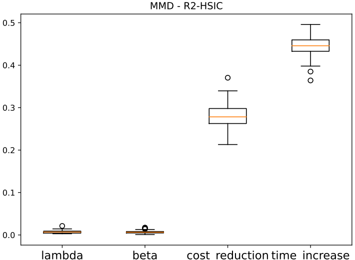
Quelle est l'influence de la distribution associée à chaque paramètre ?

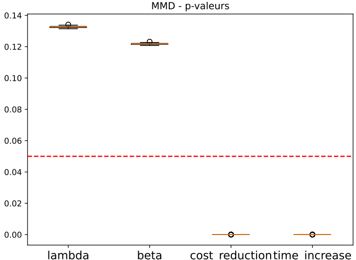
Distribution uniforme





Distribution uniforme avec support élargi ($\times 2$)



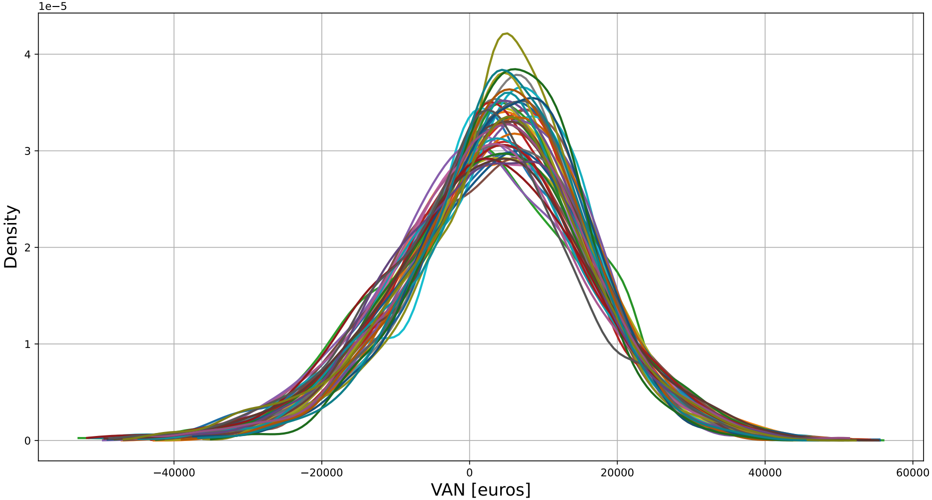


Cas d'application : installation hydroélectrique

- 6 co-variables
- 3 analyses de sensibilité :
  - Analyse de sensibilité globale sur la VAN moyenne
  - Analyse de sensibilité ciblée sur la probabilité d'une VAN négative : $P(VAN < 0.) < 0.15$
  - Analyse de sensibilité sur la distribution de la VAN
- 500 simulations Montecarlo par par évaluation de distribution
- 500 évaluations de distribution de la VAN
- Intervalle de confiance calculé sur 50 re-échantillonnages avec remise

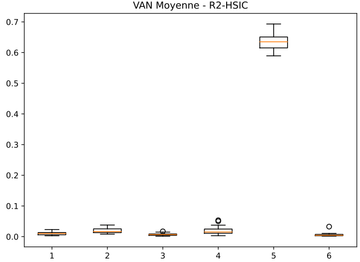
Contexte : simulation de stratégies de maintenance

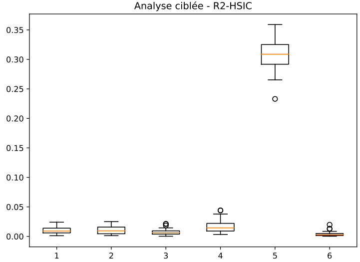
Cas d'application : Installation hydroélectrique (50 calculs, 1000 simulations par calcul)



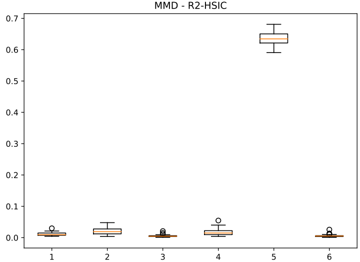
Cas d'application : installation hydroélectrique

R2-HSIC



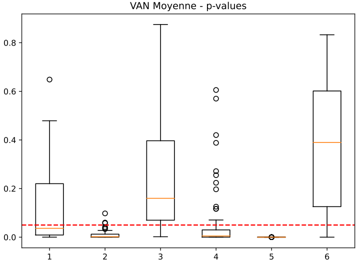


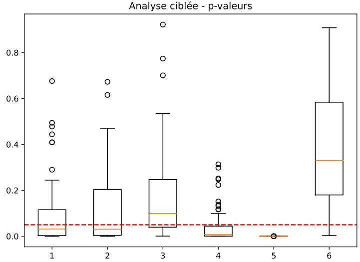
1. Dispo TRSN
2. Pb Defaillance Cables
3. Proba Avarie new T7
4. Proba Avarie old T7
5. Facteur fiab
6. Facteur coût



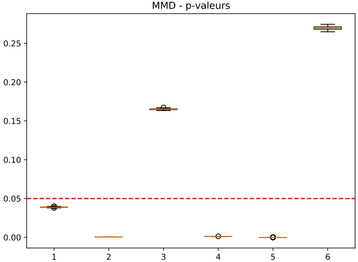
Cas d'application : installation hydroélectrique

p-valeurs





1. Disponibilité TRSN
2. Probabilité defaillance cables
3. Probabilité Avarie new T7
4. Probabilité Avarie old T7
5. Facteur fiabilité
6. Facteur coût



Conclusions

- Méthodologie d'analyse de sensibilité sur les distributions de valorisation d'investissements
- 3 types d'analyse de sensibilité basées sur les indices HSIC (et p-valeurs associées) sont mises en place
  - Analyse globale d'une valeur scalaire représentative
  - Analyse ciblée sur un domaine critique
  - Analyse sur la distribution de la sortie
- Interprétabilité potentiellement difficile
- Les 3 méthodes fournissent des résultats semblables concernant l'importance relative entre co-variables et l'identification de variables non-influentes
  - Modèles relativement simples
- Variabilité des résultats considérablement plus faible pour les analyses de la distribution entière mais coûts de calcul additionnels considérables
  - Analyse du compromis nécessaire

Perspectives & travaux futurs

- Identification de cas d'applications caractérisés par des distributions en sortie plus complexes
  - Multi-modales
  - Queues de distribution importantes
- Prise en considération de co-variables non continues (e.g., loi de défaillance)
- Analyse de l'influence des choix de noyaux
- Extension aux HSIC-ANOVA (meilleure gestion des co-variables dépendantes)
- Optimisation des temps de calcul (parallélisation et exploitation de GPU)
- Applications aux parcs éoliens


"That's all Folks!"

Quelques références

- Da Veiga, Sébastien. "Kernel-based ANOVA decomposition and Shapley effects–Application to global sensitivity analysis." arXiv preprint arXiv:2101.05487 (2021).
- Gretton, Arthur, et al. "A kernel statistical test of independence." Advances in neural information processing systems 20 (2007).
- Gretton, Arthur, et al. "A kernel two-sample test." The Journal of Machine Learning Research 13.1 (2012): 723-773.
- Marrel, Amandine, and Vincent Chabridon. "Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor." Reliability Engineering & System Safety 214 (2021): 107711.
- Székely, Gábor J., and Maria L. Rizzo. "Energy statistics: A class of statistics based on distances." Journal of statistical planning and inference 143.8 (2013): 1249-1272.