ENSAI

Quantification d'incertitudes par méthodes à noyaux

Soutenance de fin d'études 2023/2024

Louis Allain

5 Novembre 2024

Tuteurs:
Brian Staber
Sébastien Da Veiga

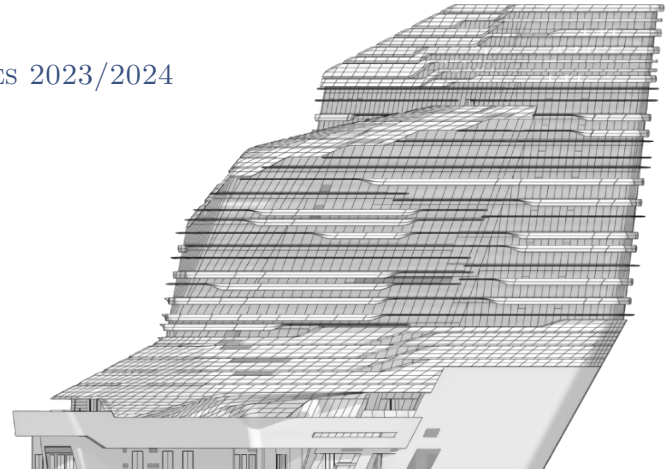


Table des matières

1 Introduction

1. Introduction

2. Méthodes usuelles

3. Prédiction conforme

4. Méthode à noyaux

5. Améliorations des méthodes à noyaux

6. Conclusion

Centre Recherche & Technologie du Groupe Safran.



Safran Tech

Département DSN, équipe FLEX.

Problématique

- Utilisation de simulateurs numériques déterministes pour concevoir des pièces
- Variation des paramètres d'entrée (matériaux, géométries, conditions d'utilisation...)
- Une seule simulation peut prendre plusieurs heures
- Utilisation du machine learning pour lever cette limitation

Construction d'un métamodèle

1 Introduction

Planification d'expériences

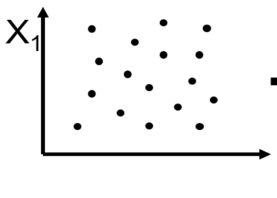
Points où l'on va effectuer les simulations ou les expériences

« Simulation »

On effectue les expériences ou les simulations sur le code

Méta-modélisation

Approximation du phénomène en dehors des points « simulés »



Code de calcul

Expérience

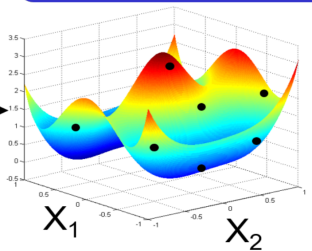


Table des matières

2 Méthodes usuelles

1. Introduction

2. Méthodes usuelles

3. Prédiction conforme

4. Méthode à noyaux

5. Améliorations des méthodes à noyaux

6. Conclusion

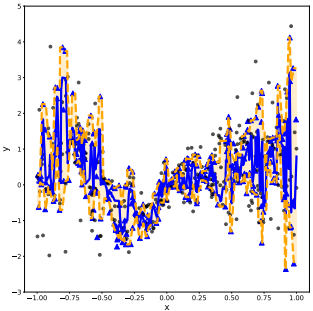
En pratique

2 Méthodes usuelles

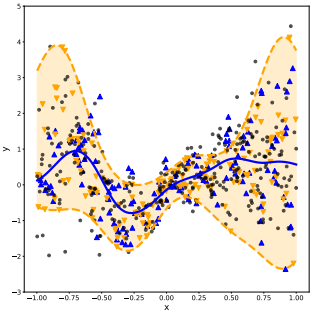
- Méthode largement adoptée dans de nombreuses industries
- Multitude de modèles possibles :
  - Régression linéaire/logistique
  - Forêts aléatoires
  - Processus Gaussiens
  - Réseaux de neurones...
- Pour des applications critiques, on a besoin d'intervalles de confiance autour de ces prédictions
- Quels sont les outils disponibles aujourd'hui ?

De nombreuses méthodes

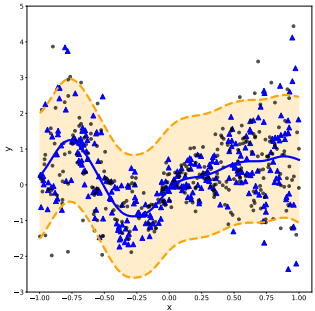
2 Méthodes usuelles



(a) Forêts aléatoires



(b) Régression Quantile



(c) Processus Gaussien

- Des intervalles de confiance très différents suivant les méthodes
- Comment définir un bon intervalle de confiance ?

Qu'est ce qu'un bon intervalle de confiance ?

2 Méthodes usuelles

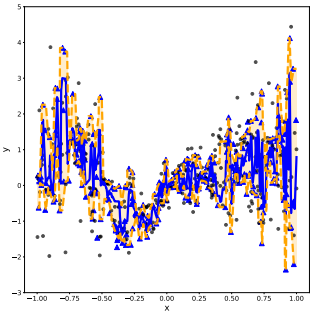
- On considère $\mathcal{D}_n$ un échantillon d'entrainement et un point X_{test} autour duquel on veut construire une bande de prédiction qui contient Y_{test} avec une certaine probabilité
- On se donne un taux d'erreur $\alpha \in (0, 1)$
- Objectif : construire des bandes de prédiction $\widehat{C}_{\mathcal{D}_n}$ avec une garantie de couverture :

$$\mathbb{P}\left(Y_{\text{test}} \in \widehat{C}_{\mathcal{D}_n}(X_{\text{test}})\right) \geq 1 - \alpha \tag{1}$$

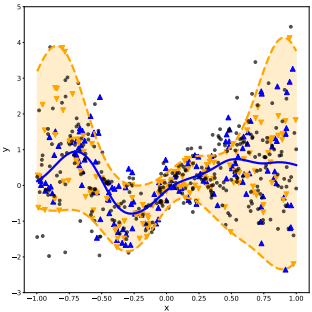
  "En moyenne sur $\mathcal{D}_n$ et X_{test}, $\widehat{C}_{\mathcal{D}_n}$ contient Y_{test} avec probabilité $1 - \alpha$"

- Conditions supplémentaires :
  - Pas d'hypothèses sur la distribution des données
  - Échantillon fini
  - (Agnostique au modèle de machine learning)

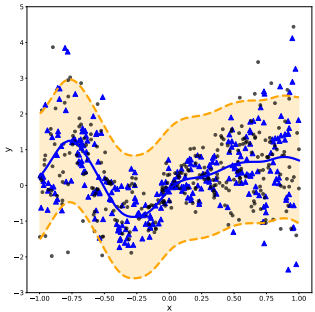
Les problèmes avec les méthodes précédentes



(a) Forêts aléatoires, garantie de couverture lorsque $n \to +\infty$



(b) Régression Quantile, garantie de couverture lorsque $n \to +\infty$



(c) Processus Gaussien, la fonction de régression doit être dans le RKHS

- **Les méthodes usuelles ne fournissent pas d'intervalles de confiance satisfaisant les propriétés recherchées**

Table des matières

3 Prédiction conforme

1. Introduction

2. Méthodes usuelles

3. Prédiction conforme

4. Méthode à noyaux

5. Améliorations des méthodes à noyaux

6. Conclusion

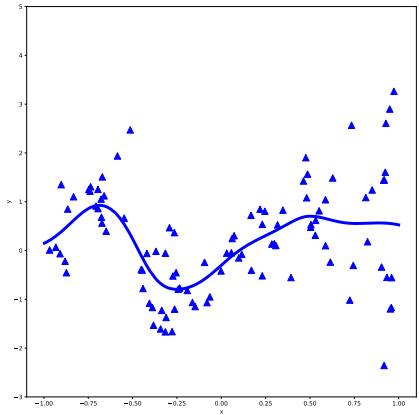
Les objectifs

3 Prédiction conforme

- Méthode récente de construction d'intervalles de confiance qui présentent les propriétés suivantes :
  - Garantie de couverture
  - Échantillon fini
  - Aucune hypothèse de distribution
  - Agnostique au modèle de machine learning

Présentation de la variante split

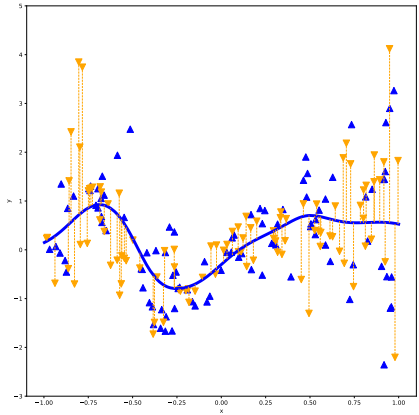
1. Entraînement d'un modèle de machine learning sur un échantillon $\mathcal{D}_p$: $\widehat{m}$



Présentation de la variante split

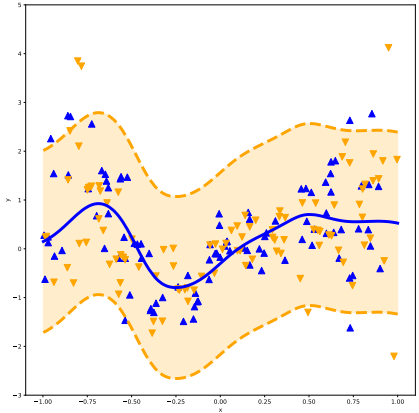
1. Entraînement d'un modèle de machine learning sur un échantillon $\mathcal{D}_p$: $\widehat{m}$
2. Sur un second échantillon de calibration $\mathcal{D}_m$ on calcule les résidus :

$$|y_i - \widehat{m}(x_i)|, \quad i \in \mathcal{D}_m \tag{2}$$

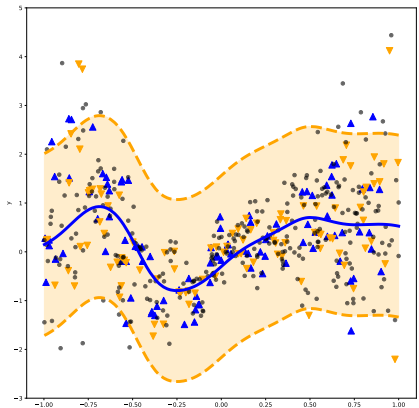


3. Calcul du quantile empirique de ces résidus au taux d'erreur α : $\widehat{q}_\alpha(\text{résidus})$
4. Construction des bandes de prédiction pour un point test X_{test} :

$$\widehat{C}_{\mathcal{D}_n}(X_{\text{test}}) = \widehat{C}_{\mathcal{D}_p \cup \mathcal{D}_m}(X_{\text{test}}) = [\widehat{m}(X_{\text{test}}) \pm \widehat{q}_\alpha(\text{résidus})] \tag{3}$$



5. Vérification de la couverture empirique sur un échantillon de test : $\mathcal{D}_t$



Limitations de la prédiction conforme

- Limitations :
  - Adaptativité des bandes
  - Hypothèse d'échangeabilité
- Avancées récentes :
  - Changer de fonction de score (e.g. résidus normalisés)
  - Régression quantile conformalisée
  - **D'autres propositions originales sous la forme d'apprentissage de la fonction de score**

Table des matières

4 Méthode à noyaux

1. Introduction

2. Méthodes usuelles

3. Prédiction conforme

4. Méthode à noyaux

5. Améliorations des méthodes à noyaux

6. Conclusion

Problème d'apprentissage de la fonction de score

- Lorsque la distribution de $Y|X$ est continue symétrique, alors construire des intervalles de prédiction avec la couverture souhaitée est "simple" :

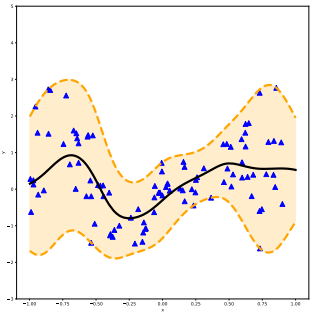
$$m(x) \pm q_\alpha^\star \sqrt{v(x)}$$

- Principe :
  - Estimer la variance conditionnelle $v(x)$ (NB : fonction positive)
  - Comme en CP, estimation du quantile $q_\alpha^\star$ pour renormalisation qui permet de garantir la couverture

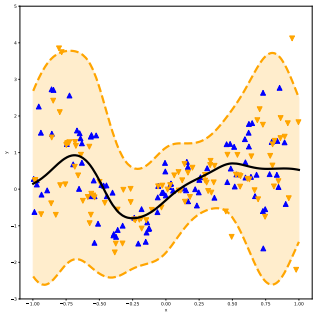
Problème d'apprentissage de la fonction de score

4 Méthode à noyaux

- Liang (2023) and Fan et al. (2024) proposent la solution suivante :
  1. Pré-estimation de v de manière à garantir une couverture de 100% sur l'échantillon d'entraînement
  2. Normalisation de la variance conditionnelle



(a) Construction à 100%



(b) Normalisation

Écriture du problème d'apprentissage

- Liang (2023) and Fan et al. (2024) proposent la solution suivante :
  1a. Pré-estimation de v de manière à garantir une couverture de 100% sur l'échantillon d'entraînement

$$\begin{aligned} &\min_{v \in \mathcal{V}} \\ &\text{t.q. } v(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2, \ i \in [\![1, n]\!] \end{aligned} \tag{4}$$

Écriture du problème d'apprentissage

- Liang (2023) and Fan et al. (2024) proposent la solution suivante :
  - 1a. Pré-estimation de v de manière à garantir une couverture de 100%
  - 1b. Faire cette estimation en contrôlant la régularité $\|v\|$ de la fonction v

Contrôle explicite de la régularité de v :

$$\begin{aligned} \min_{v \in \mathcal{V}} \ & \|v\| \\ \text{t.q. } & v(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2, \ i \in [\![1, n]\!] \end{aligned}$$

Contrôle implicite de la régularité de v :

$$\begin{aligned} \min_{v \in \mathcal{V}} \ & \\ \text{t.q. } & v(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2, \ i \in [\![1, n]\!] \\ & \|v\| \leq r \end{aligned}$$

Écriture du problème d'apprentissage

- Liang (2023) and Fan et al. (2024) proposent la solution suivante :
  - 1a. Pré-estimation de v de manière à garantir une couverture de 100%
  - 1b. Faire cette estimation en contrôlant la régularité de la fonction v
  - 1c. Optimisation de certaines caractéristiques de la fonction v, comme la largeur moyenne des bandes

$$\begin{aligned} &\min_{v \in \mathcal{V}} \ \frac{1}{n} \sum_{i=1}^{n} v(x_i) \\ &\text{t.q. } \ v(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2, \ i \in [\![1, n]\!] \end{aligned} \tag{5}$$

Écriture du problème d'apprentissage

- Écriture générale :

$$\min_{v \in \mathcal{V}} \; L\left(x_1, \ldots, x_n\right) + \lambda \|v\| \tag{6}$$

où la fonction de perte prend la forme suivante :

$$L\left(x_1, \ldots, x_n\right) = \begin{cases} 0 \,/\, \frac{1}{n} \sum_{i=1}^{n} v(x_i) & \text{si } v(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2, \; i \in [\![1, n]\!] \\ +\infty & \text{sinon.} \end{cases} \tag{7}$$

Difficultés du problème d'apprentissage

- C'est un problème d'optimisation dont l'inconnue est une fonction : $\mathcal{V}$ est de dimension infinie
- La théorie RKHS permet de passer d'un problème de dimension infinie à un problème de dimension finie tractable, grâce au théorème du représentant
- Une contrainte supplémentaire est que la fonction recherchée doit être **positive**
- Marteau-Ferey et al. (2020) a étudié l'estimation d'une fonction positive dans un RKHS

Définition du problème dans un RKHS

- Soit $\mathcal{X}$ l'espace d'entrée, $\mathcal{H}$ un RKHS (feature map ϕ) et $\mathcal{S}(\mathcal{H})$ l'ensemble des opérateurs Hermitiens linéaires de $\mathcal{H}$
- Si $\mathcal{H}$ est de dimension finie, on sait que :

$$\langle u, Mu \rangle_{\mathcal{H}} = u^\top M u \tag{8}$$

  est positif $\forall u \in \mathcal{H}$ ssi la matrice $M \succeq 0$
- Si $\mathcal{H}$ est de dimension infinie, on considère de manière similaire :

$$\langle \phi(x), \mathcal{A}[\phi](x) \rangle_{\mathcal{H}} = v_{\mathcal{A}}(x) \tag{9}$$

  est positif $\forall x \in \mathcal{X}$ ssi l'opérateur $\mathcal{A} \in \mathcal{S}(\mathcal{H}) \succeq 0$

Définition du problème dans un RKHS

$$\min_{v \in \mathcal{V}} \; L\left(x_1, \ldots, x_n\right) + \lambda \|v\| \tag{10}$$

$$\Downarrow \text{ RKHS de fonctions positives}$$

$$\min_{\mathcal{A} \succeq 0} \; L\left(v_{\mathcal{A}}(x_1), \ldots, v_{\mathcal{A}}(x_n)\right) + \lambda_1 \|\mathcal{A}\|_\star + \lambda_2 \|\mathcal{A}\|_F^2 \tag{11}$$

où $\|\mathcal{A}\|_\star$ et $\|\mathcal{A}\|_F^2$ sont des mesures de la régularité de $v_{\mathcal{A}}$

Théorème du représentant

- L'équation précédente admet une solution :

$$A_\star = \sum_{i,j=1}^{n} B_{i,j}\phi(x_i)\phi(x_j)^\top, \quad \text{pour une matrice } B \in \mathbb{R}^{n\times n},\ B \succeq 0 \tag{12}$$

- La fonction positive associée s'écrit :

$$\forall x \in \mathcal{X},\ v_{A_\star}(x) = \sum_{i,j=1}^{n} B_{i,j}k(x,x_i)k(x,x_j) \tag{13}$$

- On a une représentation en dimension finie

Problème d'apprentissage en dimension finie

- Le problème d'apprentissage s'écrit à l'aide du théorème du représentant :

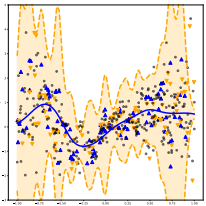
En dimension infinie :

$$\begin{aligned} \min_{\mathcal{A}\in\mathcal{S}(\mathcal{H})} \quad & \|\mathcal{A}\|_\star \\ \text{t.q.} \quad & v_{\mathcal{A}}(x_i) \geq (y_i - \widehat{m}(x_i))^2 \\ & \mathcal{A} \succeq 0 \end{aligned}$$

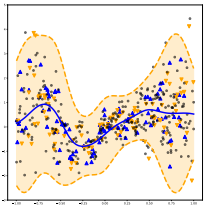
En dimension finie :

$$\begin{aligned} \min_{\boldsymbol{B}\in\mathbb{S}^n} \quad & \mathrm{Tr}(\boldsymbol{K}\boldsymbol{B}) \\ \text{t.q.} \quad & \langle K_{x_i}, \boldsymbol{B}K_{x_i}\rangle \geq (y_i - \widehat{m}(x_i))^2 \\ & \boldsymbol{B} \succeq 0 \end{aligned}$$

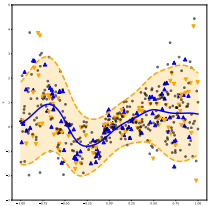
Illustration et point faible



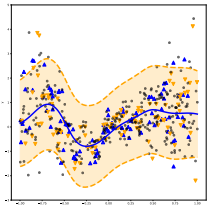
(a) Longueur de corrélation = 0.05



(b) Longueur de corrélation = 0.2



(c) Longueur de corrélation = 0.4



(d) Longueur de corrélation = 0.8

- Plusieurs difficultés :
  - Temps de calcul, complexité en $\mathcal{O}(n^3)$ car résolution problème SDP
  - Choix de la longueur de corrélation du noyau
  - (Procédure de calibration)

Table des matières

5 Améliorations des méthodes à noyaux

1. Introduction

2. Méthodes usuelles

3. Prédiction conforme

4. Méthode à noyaux

5. Améliorations des méthodes à noyaux

6. Conclusion

Améliorer le temps de calcul

5 Améliorations des méthodes à noyaux

1. Reformuler le théorème du représentant

Formulation B :

$$\begin{aligned} \min_{\boldsymbol{B}\in\mathbb{S}^n} \quad & \mathrm{Tr}(\boldsymbol{K}\boldsymbol{B}) \\ \text{t.q.} \quad & \langle K_{x_i}, \boldsymbol{B}K_{x_i}\rangle \geq (y_i - \widehat{m}(x_i))^2 \\ & \boldsymbol{B} \succeq 0 \end{aligned}$$

Formulation A :

$$\begin{aligned} \min_{\boldsymbol{A}\in\mathbb{S}^n} \quad & \mathrm{Tr}(\boldsymbol{A}) \\ \text{t.q.} \quad & \tilde{v}_{\boldsymbol{A}}(x_i) \geq (y_i - \widehat{m}(x_i))^2 \\ & \boldsymbol{A} \succeq 0 \end{aligned}$$

Améliorer le temps de calcul

1. Reformuler le théorème du représentant
2. Utiliser la régularisation de Frobenius
   — Pour une fonction de perte convexe et $\lambda_2 > 0$, le problème SDP devient strictement convexe

Sans régularisation de Frobenius :

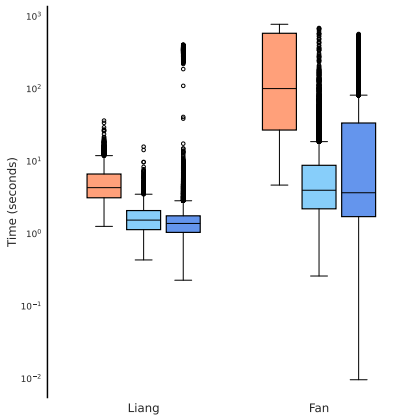
$$\begin{aligned} \min_{\boldsymbol{A} \in \mathbb{S}^n} \quad & \operatorname{Tr}(\boldsymbol{A}) \\ \text{t.q.} \quad & \tilde{v}_{\boldsymbol{A}}(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2 \\ & \boldsymbol{A} \succeq 0 \end{aligned}$$

Avec régularisation de Frobenius :

$$\begin{aligned} \min_{\boldsymbol{A} \in \mathbb{S}^n} \quad & \operatorname{Tr}(\boldsymbol{A}) + \lambda_2 \operatorname{Tr}(\boldsymbol{A}^2) \\ \text{t.q.} \quad & \tilde{v}_{\boldsymbol{A}}(x_i) \geq \left(y_i - \widehat{m}(x_i)\right)^2 \\ & \boldsymbol{A} \succeq 0 \end{aligned}$$

Gains obtenus 1/2

5 Améliorations des méthodes à noyaux

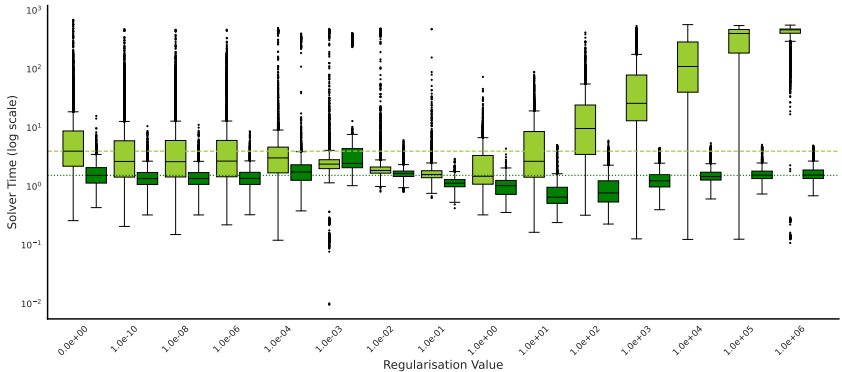


- Un gain de temps significatif en passant de la formulation B à A
- En moyenne, la régularisation de Frobenius n'est pas aussi effective qu'espérée :
  - Est-ce que la valeur du coefficient de régularisation peut avoir un impact sur le temps de calcul ?

Figure 5: B, $\lambda_2 = 0$ - A, $\lambda_2 = 0$ - A, $\lambda_2 > 0$.

Gains obtenus 2/2

5 Améliorations des méthodes à noyaux



- Pour une faible valeur de régularisation les gains de temps sont négligeables
- Pour Fan, une grande valeur de régularisation augmente les temps de calculs
- Pour Liang, une grande valeur de régularisation stabilise les temps de calculs
- **Il existe des valeurs qui permettent de réduire les temps de calculs**

Choix de la longueur de corrélation

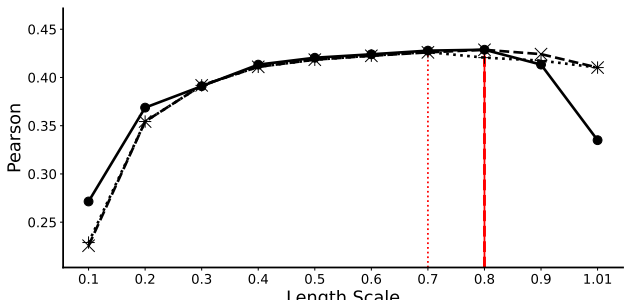
- La longueur de corrélation ls du noyau a un impact très fort sur la forme des bandes de prédiction :
  - $ls \to \infty$, les bandes de prédiction deviennent de largeur constante
  - $ls \to 0$, les bandes de prédiction ne sont plus assez lisses
- Il faut faire un compromis pour trouver la longueur de corrélation qui correspond aux bandes de prédiction les plus adaptatives

Quantification de l'adaptivité des bandes de prédictions

- Qu'est ce que l'adaptativité :
  - Une grande largeur, lorsque le modèle fait des erreurs et/ou que le bruit est important
  - Une faible largeur, lorsque le modèle est performant et/ou que le bruit est faible
- L'adaptativité des bandes de prédiction se traduit par une forte dépendance entre les résidus et la largeur des bandes de prédiction
- Notre proposition : choisir la longueur de corrélation qui maximise cette dépendance
- Question : choix de la mesure de dépendance

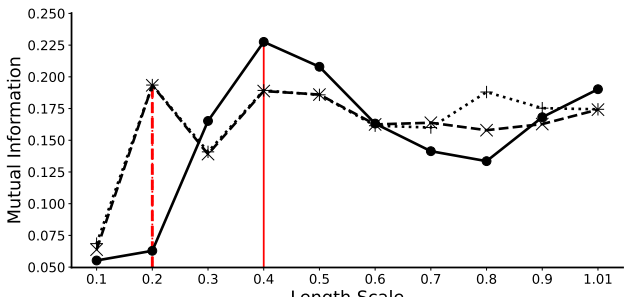
Différentes mesures de dépendance

- Mesure de corrélation : ne détecte pas les dépendances générales et pas suffisamment discriminante



Différentes mesures

- Mesure de corrélation
- Information mutuelle : peut détecter n'importe quelle dépendance mais estimation numérique instable



Différentes mesures

5 Améliorations des méthodes à noyaux

- Mesure de corrélation
- Information mutuelle
- **Critère HSIC :** peut détecter n'importe quelle dépendance et estimation robuste

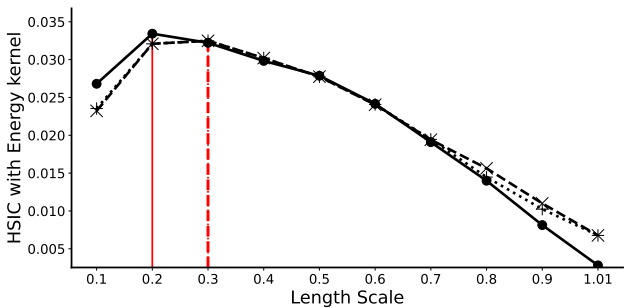
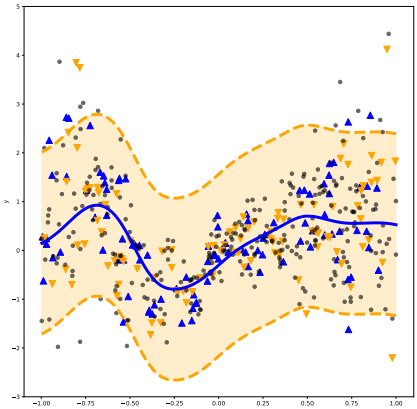
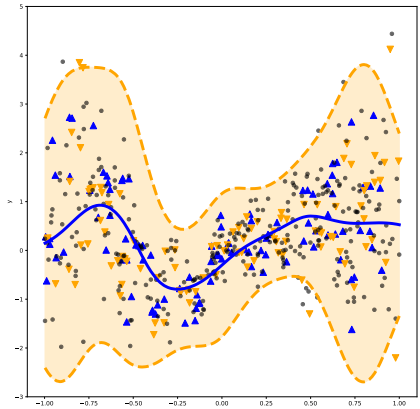


Illustration finale

5 Améliorations des méthodes à noyaux



(a) Prédiction conforme split



(b) Notre proposition

Table des matières

6 Conclusion

1. Introduction

2. Méthodes usuelles

3. Prédiction conforme

4. Méthode à noyaux

5. Améliorations des méthodes à noyaux

6. Conclusion

Conclusion

6 Conclusion

- Étude d'une nouvelle méthode à noyaux :
  - Adaptative
  - Basée sur l'apprentissage d'une fonction de score
  - Reformulation du problème dans un RKHS de fonctions positives
- Nos contributions durant le stage :
  - Accélération drastique du temps de calcul
  - Procédure originale pour le choix des hyperparamètres du noyau
  - Procédure de calibration
- Pour la suite :
  - Augmentation du nombre de données et de la dimension du problème
  - Investiguer d'autres algorithmes d'optimisation et l'impact de λ_2
  - Apport de ces méthodes lorsqu'il n'y a pas d'échangeabilité

ENSAI

Q&R

Merci pour votre attention !

Vos suggestions, remarques et questions sont les bienvenues !