Lundi 13 mai 2024 - Jérémy Bettinger

Audition, Laboratoire du CREST

Sous la direction de François Portier & Adrien Saumard
ENSAI, Rennes.

Sommaire

Mon parcours - Parcours Académique

Parcours Académique

- Classe prépa maths sup/spé 2018-2020
- ENS Rennes 2020-2024
- L3 16.3/20 (2021), M1 16/20 (2022)
- Agrégé de mathématiques (2023)
- M2 Recherche, parcours Aléatoire 17.8/20 (S1) (2024)
- Diplôme de l'ENS Rennes, Mention Très Bien (à venir)

Gratifications

- Bourse de stage, Association des Membres de l'Ordre des Palmes Académiques. Stage de L3 (2021)
- Bourse d'excellence, Labex Centre de recherche de mathématiques Henri Lebesgue. M2 Recherche (2023/2024)

Mon parcours - Responsabilités

Responsabilités

- Membre du Conseil d'Administration du lycée-CPGE (2018)
- Secrétaire de l'Association des CPGE du lycée (2019)
- Représentant de promotion des élèves de l'ENS Rennes, département mathématiques (2020-2024)
- Représentant des étudiants de prépa agreg (2022/2023)
- Représentant des étudiants du M2 Recherche (2023/2024)
- Membre élu au Conseil de l'UFR Mathématiques de Rennes (2021-2024)
- Membre de la commission enseignement de l'UFR Mathématiques de Rennes (2021-2024)

Mon parcours - Enseignements

Enseignements

- Colles en classe prépa (2020-2022)
- Agrégé de mathématiques (2023)
- Environ 190h/TD donnés en 2023/2024 : ENSAI (30h) ; EPITA (70h) ; INSA (90h)
- Colles en prépa en 2023/2024 (160h)
- Jury de concours CPGE (2024)
- Encadrement d'un stagiaire de L2 pendant 3 mois en 2024 (sujet : topologie matricielle)

Promotion des maths

MATh.en.JEANS, Jury du TFJM^2 Rennais, RJMI de Rennes, Tutorat pour des élèves de REP+, Exposés.

Mon parcours - Stages

Stages

- L3 Recherche (2021) : Le théorème des nombres premiers, encadré par Gérald Tenenbaum (IECL).
- M1 Mathfonda (2022) : Le problème des moments et le théorème de Berry-Esseen, encadré par Nicolas Juillet (UHA).
- Séminaire de recherche de M2 Recherche (2024) : Modélisation de la croissance bactérienne par des Processus de Markov déterministes par morceaux et estimation statistique non paramétrique, encadré par Nathalie Krell (IRMAR).

Stage & Thèse

Titre du sujet de thèse

Apprentissage de métrique pour les méthodes statistiques par moyennes locales.

Idée

Calibrer les distances de façon adaptée au problème sous-jacent et ainsi atténuer les défauts liés à l'utilisation de distances usuelles.

Objectif

Obtenir un contrôle stochastique de l'erreur de prédiction à échantillon fini, en faisant apparaître le rôle de la dimension de façon explicite.

Fléau de la dimension

- On a potentiellement de variables inutiles pour prédire la variable d'intérêt.
- Sans structure, les vitesses optimales de convergence dépendent de la dimension ambiante des covariables et se dégradent quand la dimension augmente.
- Le but est de parvenir à améliorer les vitesses de convergence lorsque le modèle sous-jacent présente une structure de dimension réduite (par exemple quand certaines variables sont non informatives)

Fléau de la dimension - approche semi-paramétrique

Le modèle qui nous intéresse est

$$Y = g(X) + \varepsilon$$

On supposera que la fonction g a une forme particulière avec une dimension réduite :

$$Y = \tilde{g}(\beta^T X) + \varepsilon$$

où $X \in \mathbb{R}^d, g : \mathbb{R}^d \mapsto \mathbb{R}, \tilde{g} : \mathbb{R}^{d_0} \mapsto \mathbb{R}, \beta \in \mathbb{R}^{d \times d_0}$ **inconnue**.

- Réduction de la dimension : la dimension effective est $d_0 \ll d$.

Idée

Le but est donc d'estimer β afin de réduire la dimension des variables explicatives via la transformation $X \mapsto \hat{\beta}^T X$.

Une stratégie possible :

- Estimer le gradient de g : ∇g.
- En déduire une estimation de β (en imposant une renormalisation pour que le modèle soit identifiable).
- En déduire les directions peu exploitées.

Axe plus général : apprentissage de métrique

Utilisation de la distance de Mahalanobis :

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$$

où Σ est une matrice symétrique définie positive à estimer.

- Le but est de se libérer des axes conventionnels et de trouver des axes pertinents adaptés aux données.
- Dans l'objectif de réduire la dimension.
- But : avoir de meilleures bornes.

Axe de recherche : Définir une telle procédure pour les arbres.

Stage - Cadre

Soient $(X_i)_{i=1,\dots,n}$ les variables d'entrée et $(Y_i)_{i=1,\dots,n}$ les variables de sortie.
Nous supposons avoir $(Y_i, X_i)_{i=1,\dots,n}$ i.i.d. selon la distribution $\mathcal{P}$.
De plus, $Y_i \in \mathbb{R}$ et $X \in \mathcal{X} \subset \mathbb{R}^d$.
Nous désignons par $\mathcal{A} = (A_p)_{p=1\dots L}$ une classe d'ensembles qui partitionne l'espace $\mathcal{X}$, i.e. $\mathcal{X} = \sqcup_{p=1}^{L} A_p$.

Nous cherchons à estimer une fonction de régression $g(x) = \mathbb{E}(Y|X = x)$ telle que $Y = g(X) + \varepsilon$.

Nous supposons que $(X_i, \varepsilon_i)_{i=1,\dots,n}$ sont indépendants.
Nous étudions l'estimateur suivant de la fonction g :

$$\hat{g}(A_p) = \frac{\sum_{i=1}^{n} Y_i \, \mathbf{1}_{A_p}(X_i)}{\sum_{j=1}^{n} \mathbf{1}_{A_p}(X_j)},$$

et pour tout $x \in \mathcal{X}$:

$$\hat{g}(x) = \sum_{p=1}^{L} \hat{g}(A_p) \, \mathbf{1}_{A_p}(x).$$

Un résultat

Théorème

Sous les hypothèses suivantes pour $x \in A$:

1. *L'espace $\mathcal{X}$ est borné par $M > 0$,*
2. *Les $(X_i, \varepsilon_i)_{i=1,\dots,n}$ sont indépendants, avec les $(\varepsilon_i | X_i)_{i=1,\dots,n}$ sous-gaussiens de paramètre σ^2,*
3. *Les $(X_i)_{i=1,\dots,n}$ ont une loi à densité bornée inférieurement par une constante $a > 0$,*
4. *L'espace $\mathcal{X}$ est partitionné en rectangles contenant au moins k points et au plus $2k$ points,*
5. *La fonction g a des dérivées partielles au premier ordre au point x et le cardinal de l'ensemble $S(x) = \{k \in [\![1, d]\!],\ \partial_k g(x) \neq 0\}$ est d_0,*
6. *Il existe une constante $c > 0$ telle que le côté le plus long restreint à $S(x)$ du rectangle A, noté h_+, est borné supérieurement par c fois la taille du côté le plus court restreint à $S(x)$ de A, noté h_-, i.e. $\exists\, c > 0 :\ h_+ \leq c\, h_-$,*

Un résultat

Alors, nous avons avec une probabilité au moins $1 - 2\delta$:

$$\begin{aligned} |\hat{g}(A) - g(x)| &\leq 2\sqrt{\frac{2\sigma^2}{k} \ln\left(\frac{(n+1)^{2d}}{\delta}\right)} \\ &+ \psi\left[\frac{k}{n} + \frac{1}{n} \ln\left(\frac{4(2n+1)^{2d}}{\delta}\right)\right]^{1/d_0} \end{aligned}$$

où $\psi = 4c\, d_0\, \sup_{x \in A} \|\nabla g(x)\|_{\infty, S(x)} / (aM^{d-d_0})^{1/d_0}$.

Remarques

- Nécessite de connaître $S(x)$ afin de connaître h_+, h_-, c et d_0. Il faudrait essayer de l'estimer ou de ne pas couper selon des axes (de $S(x)^c$).
- Avec ce choix de considérer $S(x)$, les cellules où des gradients directionnels sont éventuellement nuls peuvent être grandes selon ces axes.
- Quels choix optimal en pratique de k ?
  Piste de recherche : technique de validation croisée.

Un résultat

Si on note $d_{max} = \max\{d_0(A), A \in \mathcal{A}\}$ on a avec probabilité au moins $1 - 2\delta$:

$$\begin{aligned} |\hat{g}(x) - g(x)| &\leq 2\sqrt{\frac{2\sigma^2}{k} \ln\left(\frac{(n+1)^{2d}}{\delta}\right)} \\ &+ C\left[\frac{k}{n} + \frac{1}{n}\ln\left(\frac{4(2n+1)^{2d}}{\delta}\right)\right]^{1/d_{max}} . \end{aligned}$$

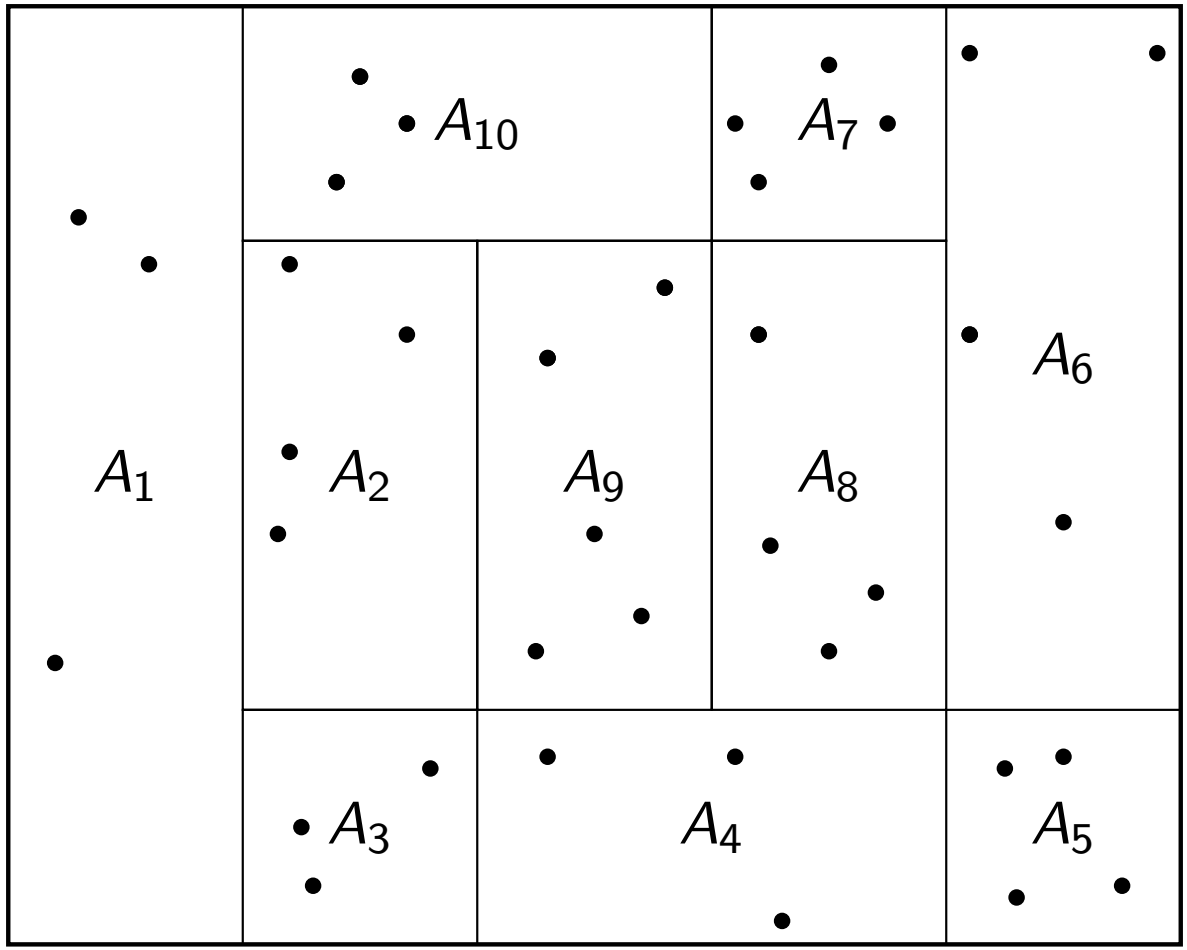
En égalisant les termes à droite de l'inégalité, on choisit $k = n^{2/(2+d_{max})}$.

On obtient pour vitesse finale : pour tout $x \in \mathcal{X}$, on a avec probabilité au moins $1 - 2\delta$:

$$|\hat{g}(x) - g(x)| \leq \tilde{C}(d, \sigma^2) \sqrt{\ln\left(\frac{n^{2d}}{\delta}\right)} \left(\frac{1}{n}\right)^{1/(2+d_{max})} .$$

Exemple

$n = 37$ données ; $\mathbb{R}^{d=2}$; au moins $k = 3$ points ; au plus $2k$ points.



Merci de votre attention

Références :

[1] Vapnik, V. N. and A. Y. Chervonenkis (2015). On the uniform convergence of relative frequencies of events to their probabilities. In Measures of complexity, pp. 11–30. Springer, Cham. Reprint of Theor. Probability Appl. 16 (1971), 264–280.

[2] Van Der Vaart, A. W. and J. A. Wellner (1996). Weak Convergence and Empirical Processes. With Applications to Statistics. Springer Series in Statistics. New York : Springer-Verlag.

[3] Portier, F. (2023). Nearest neighbor process : weak convergence and non-asymptotic bound.