Adaptation sur des espaces de Besov en estimation de la densité sous contrainte de confidentialité différentielle locale

Amandine Dubois [1], Cristina Butucea [2], Martin Kroll [2], et Adrien Saumard [1]

[1] *CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35712 Bruz cedex, France. amandine.dubois@ensai.fr; adrien.saumard@ensai.fr*

[2] *CREST, ENSAE-ParisTech , 5 avenue Henry Le Chatelier, F-91120 Palaiseau. Butucea.Cristina@ensae.fr; martin.kroll@ensae.fr*

Résumé. Nous nous intéressons à l'estimation non-paramétrique d'une densité de probabilité sous la contrainte supplémentaire que seules des données privatisées sont disponibles. À cette fin, nous adoptons une récente généralisation de la théorie minimax classique au cadre de la confidentialité différentielle locale et nous donnons des bornes inférieures sur la vitesse de convergence sur les espaces de Besov $\mathcal{B}^s_{pq}$ pour le risque L^r. Les vitesses de convergence dans le cas privé sont détériorées par rapport à celles obtenues dans le cadre classique mais révèlent un changement de régime analogue. Afin de répondre à l'exigence de confidentialité, nous suggérons d'ajouter aux coefficients d'ondelettes empiriques un bruit de Laplace correctement calibré. Un estimateur non linéaire adaptatif par ondelettes avec un seuillage correctement choisi atteint la vitesse donnée par la borne inférieure à un facteur logarithmique près.

Mots-clés. Estimation d'une densité, confidentialité différentielle locale, ondelettes, problèmes inverses.

Abstract. We address the problem of non-parametric density estimation under the additional constraint that only privatised data are available for inference. For this purpose, we adopt a recent generalisation of classical minimax theory to the framework of local α-differential privacy and provide lower bounds on the rate of convergence over Besov spaces $\mathcal{B}^s_{pq}$ under mean integrated L^r-risk. The rates of convergence in this case are deteriorated compared to the standard setup without privacy, but reveal an analogous elbow effect. In order to fulfill the privacy requirement, we suggest adding suitably scaled Laplace noise to empirical wavelet coefficients. An adaptive non-linear wavelet estimator with appropriately chosen thresholding is shown to attain the lower bound within a logarithmic factor.

Keywords. Density estimation, local differential privacy, wavelets, inverse problems.

1 Introduction

Problème. De nos jours, une grande quantité de données, telles que les dossiers médicaux, les informations de localisation des téléphones portables, l'historique de navigation sur Internet, sont collectées et stockées. Un enjeu majeur consiste à caractériser et à équilibrer l'utilité statistique de ces données et la protection de la vie privée des personnes auprès desquelles elles sont obtenues.

Nous nous intéressons ici au problème de l'estimation non-paramétrique d'une densité de probabilité sous contrainte de confidentialité différentielle locale : pour $i = 1, \ldots, n$, le i-ème détenteur de données observe une variable aléatoire réelle X_i distribuée selon une densité de probabilité f. L'objectif est que chaque détenteur de données publie une version anonymisée Z_i de X_i de sorte que la notion de confidentialité différentielle locale définie ci-dessous soit satisfaite et que la densité f puisse être estimée à partir des données $Z_1, \ldots, Z_n$ d'une manière optimale.

Confidentialité différentielle locale. La notion de *confidentialité différentielle locale* regroupe deux concepts différents, à savoir la confidentialité *locale* et la confidentialité *différentielle*. La notion qualitative de confidentialité *locale* caractérise la manière dont les différentes entités détenant les données $X_1, \ldots, X_n$ peuvent interagir pour générer un échantillon privé Z qui pourra être communiqué. Elle s'oppose au concept de confidentialité *globale* où les détenteurs de données font confiance à une même autorité qui a accès à l'ensemble des données non masquées $X_1, \ldots, X_n$ et qui génère, à partir de cette information complète, des données privées communicables. Dans la configuration *locale*, une telle autorité, en qui toutes les parties ont confiance, n'existe pas. Cependant un certain degré d'interaction entre les différentes parties est permis. Les données privées $Z_1, \ldots, Z_n$ sont obtenues de la manière suivante : sachant $X_i = x_i$ et $Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1}$, le i-ème détenteur de données génère

$$Z_i \sim Q_i(\cdot \mid X_i = x_i, Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1})$$

pour un noyau de Markov Q_i. Un cas particulier important est celui de la confidentialité locale *non-interactive* où Z_i dépend seulement de X_i. La notion de confidentialité *différentielle* est une notion quantitative. Nous donnons sa définition dans le cas local et renvoyons à l'article de Wasserman et Zhou (2010) pour une définition dans le cas global.

Définition 1.1. Une suite de noyaux de Markov Q_i garantit la α-confidentialité différencielle locale si

$$\sup_{A \in \sigma(\mathcal{Z})} \frac{Q_i(A \mid X_i = x, Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1})}{Q_i(A \mid X_i = x', Z_1 = z_1, \ldots, Z_{i-1} = z_{i-1})} \leq \exp(\alpha), \quad \text{for all } x, x' \in \mathcal{X}.$$

Plus $\alpha \in (0, \infty)$ est petit, plus la contrainte de confidentialité ci-dessus est forte. Wasserman et Zhou (2010) proposent une interprétation de la confidentialité différentielle en terme de risque d'identification.

Notons que dans le cadre de l'estimation de densité sous contrainte de confidentialité différentielle locale, le statisticien qui reçoit l'échantillon privatisé non-interactif $Z_1, \ldots, Z_n$ est en présence d'un problème inverse. En effet, il doit estimer la densité de probabilité sous-jacente f à partir d'un échantillon dont la densité de probabilité est le mélange $Kf(z) = \int q^{Z|X=x}(z) f(x) \mathrm{d}x$. Ainsi, dans ce cadre il faut trouver le noyau de Markov qui assure la confidentialité différentielle locale au niveau α et qui rend le problème inverse le moins difficile possible.

Cadre minimax privé. Duchi et al. (2018) ont expliqué comment étendre le cadre minimax classique pour prendre en compte des contraintes de confidentialité différentielle locale. Pour notre problème, on s'intéressera au risque minimax privé suivant :

$$\mathcal{R}_n^\star(\alpha, \|\cdot\|_r^r, \mathcal{D}_{pq}^s(L,T)) = \inf_{\substack{\tilde{f} \\ Q \in Q_\alpha}} \sup_{f \in \mathcal{D}_{pq}^s(L,T)} \mathbb{E}_{f,Q}[\|\tilde{f} - f\|_r^r],$$

où l'infimum est pris sur tous les estimateurs de f et tous les noyaux de Markov garantissant la α-confidentialité différentielle locale, et

$$\mathcal{D}_{pq}^s(L,T) = \left\{ f : f \in \mathcal{B}_{pq}^s(L), f \geq 0, \int_{\mathbb{R}} f(x)\mathrm{d}x = 1 \text{ et } \mathrm{supp}(f) \subset [-T,T] \right\},$$

où $\mathcal{B}_{pq}^s(L)$ est un ellipsoïde de Besov. Pour étudier ce risque minimax, les méthodes par ondelettes s'avèrent particulièrement appropriées, voir Donoho et al. (1996) ou Härdle et al. (1998) pour le cas non privé. Etant donnée une base d'ondelettes

$$\{\varphi_{j_0 k} = 2^{j_0/2}\varphi(2^{j_0}(\cdot) - k) \colon k \in \mathbb{Z}\} \cup \{\psi_{jk} = 2^{j/2}\psi(2^j(\cdot) - k) \colon j \geq j_0, k \in \mathbb{Z}\}, \qquad (1)$$

la densité de probabilité f admet le développement formel suivant (au sens L_2) :

$$f = \sum_{k \in \mathbb{Z}} \alpha_{j_0 k}\varphi_{j_0 k} + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} \beta_{jk}\psi_{jk}, \qquad (2)$$

où les coefficients d'ondelette sont définis par

$$\alpha_{j_0 k} = \int_{\mathbb{R}} f(x)\varphi_{j_0 k}(x)\mathrm{d}x \qquad \text{et} \qquad \beta_{jk} = \int_{\mathbb{R}} f(x)\psi_{jk}(x)\mathrm{d}x.$$

On fera les hypothèses suivantes.

Hypothèse 1.1. Suivant Härdle et al. (1998), on suppose que l'ondelette père φ génère une analyse multi-résolution de $L_2(\mathbb{R})$, qu'elle est $N+1$ fois faiblement dérivable pour un entier N tel que $0 < s < N+1$, et que $\varphi^{(N+1)}$ vérifie $\sup \mathrm{ess}_x \sum_k |\varphi^{(N+1)}(x-k)| < \infty$. On suppose de plus qu'il existe une fonction décroissante bornée Φ telle que $|\varphi(u)| \leq \Phi(|u|)$ p.p., $\int \Phi(|u|)\mathrm{d}u < \infty$ et $\int \Phi(|u|)|u|^N \mathrm{d}u < \infty$.

Hypothèse 1.2. On suppose que φ et ψ sont à support compact inclus dans $[-A, A]$.

L'hypothèse 1.1 permet de caractériser l'appartenance de f à un espace de Besov $\mathcal{B}^s_{pq}$ en fonction de ses coefficients d'ondelettes. Une conséquence de l'hypothèse 1.2 est que si f est à support compact alors pour tout $j_0 \in \mathbb{Z}$ et tout niveau de résolution $j \in \mathbb{Z}$ fixé, les coefficients α_{j_0k} et β_{jk} peuvent être non nuls seulement pour un nombre fini de k. On note $\mathcal{N}_{j_0-1}$ l'ensemble des k vérifiant $\alpha_{j_0k} \neq 0$ et pour $j \geq j_0$ on note $\mathcal{N}_j$ l'ensemble des k vérifiant $\beta_{jk} \neq 0$. A partir d'ici, on travaille avec $j_0 \in \mathbb{N}$.

Notations. Pour deux suites $\{a_n\}_n$ et $\{b_n\}_n$, on note $a_n \lesssim b_n$ si il existe une constante $C > 0$ et un entier naturel N fixé tels que $a_n \leq C \cdot b_n$, pour tout $n \geq N$. On note $a_n \asymp b_n$, si $a_n \lesssim b_n$ et $b_n \lesssim a_n$.

2 Bornes inférieures

Le résultat suivant fournit des bornes inférieures pour le risque minimax privé.

Proposition 2.1. *Soient $\alpha \in]0, \bar{\alpha}]$ pour un $\bar{\alpha} > 0$ fixé, $L > 0$ et $T > 0$. Soient $p \geq 1$ et $s > 1/p$. On a*

$$\mathcal{R}^\star_n(\alpha, \|\cdot\|_r^r, \mathcal{D}^s_{pq}(L,T)) \gtrsim \begin{cases} (n\alpha^2)^{-r\cdot\frac{s}{2s+2}}, & \text{si } p > \frac{r}{s+1}, \\ \left(\frac{n\alpha^2}{\log(n\alpha^2)}\right)^{-r\cdot\frac{(s-1/p+1/r)}{2(s-1/p)+2}}, & \text{si } p \leq \frac{r}{s+1}. \end{cases} \tag{3}$$

Cette proposition met en évidence un changement de régime analogue à celui que Donoho et al. (1996) ont obtenu dans le cadre non privé et connu sous le nom de *"elbow effect"*. La démonstration de ce résultat repose principalement sur un schéma de réduction au problème de test d'un nombre fini d'hypothèses et sur un résultat de contraction prouvé par Duchi et al. (2018). Les bornes inférieures données par (3) sont complétées dans la suite par des bornes supérieures. Dans la partie 4, on construit ainsi un estimateur par seuillage qui est adaptatif et atteint la vitesse optimale, aussi bien pour $p > \frac{r}{s+1}$ que pour $p \leq \frac{r}{s+1}$, à un facteur logarithmique près. Cet estimateur est défini à partir de données privatisées $Z_1, \ldots, Z_n$ dont la construction est donnée dans la partie 3.

3 Mécanisme de privatisation

On introduit ici un mécanisme de privatisation qui vérifie la condition de confidentialité différentielle locale et qui va nous permettre de construire un estimateur optimal à un facteur logarithmique près.

Fixons $\nu > 1$ et définissons

$$\sigma_{j_0-1} = \frac{4c_A\|\varphi\|_\infty}{\alpha} \cdot 2^{j_0/2} \text{ et } \sigma_j = \frac{4(2\nu-1)c_A\|\psi\|_\infty}{(\nu-1)\alpha} \cdot (j \vee 1)^\nu \cdot 2^{j/2}$$

pour $j \in [\![j_0, j_1]\!]$ avec $c_A = 2\lceil A \rceil + 1$. Pour $i \in [\![1, n]\!]$ et $j \in [\![j_0 - 1, j_1]\!]$, on définit

$$Z_{ijk} = \begin{cases} \varphi_{j_0 k}(X_i) + \sigma_{j_0 - 1} W_{i, j_0 - 1, k}, & \text{si } j = j_0 - 1, k \in \mathcal{N}_{j_0 - 1}, \\ \psi_{jk}(X_i) + \sigma_j W_{ijk}, & \text{si } j \in [\![j_0, j_1]\!], k \in \mathcal{N}_j, \end{cases} \tag{4}$$

où les W_{ijk} sont des variables aléatoires indépendantes distribuées selon une loi de Laplace de paramètre 1. Notons que le mécanisme défini par (4) est non-interactif puisque Z_{ijk} ne dépend pas de $X_{i'}$ pour $i' \neq i$.

Proposition 3.1. *Le mécanisme défini par* (4) *vérifie la condition de α-confidentialité différentielle locale.*

4 Bornes supérieures pour les estimateurs adaptatifs par seuillage

Dans cette partie, le mécanisme de privatisation est donné par (4). On étudie les propriétés de l'estimateur par ondelettes non linéaire suivant :

$$\widetilde{f}_n(x) = \sum_k \widehat{\alpha}_{j_0 k} \varphi_{j_0 k}(x) + \sum_{j=j_0}^{j_1} \sum_k \widetilde{\beta}_{jk} \psi_{jk}(x) \tag{5}$$

où

$$\widehat{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n Z_{i, j_0 - 1, k} \quad \text{et} \quad \tilde{\beta}_{jk} = \widehat{\beta}_{jk} \cdot \mathbb{1}_{\{|\widehat{\beta}_{jk}| \geq Kt\}},$$

et $\widehat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^n Z_{ijk}$ (le choix de t et la valeur de la constante K sont donnés dans le Théorème 4.1 ci-dessous).

Théorème 4.1. *Soit $\alpha \in]0, \bar{\alpha}]$ pour un $\bar{\alpha} > 0$ fixé. On suppose que l'ondelette père φ vérifie l'Hypothèse 1.1 pour un certain entier $N > 0$. L'estimateur $\widetilde{f}_n$ defini par* (5) *avec j_0 et j_1 choisis de sorte que*

$$2^{j_0} \asymp (n\alpha^2)^{\frac{1}{2N+2}} \quad \textit{et} \quad 2^{2j_1} \asymp \frac{n\alpha^2}{\log(n\alpha^2)},$$

avec $K = 4c_A \|\psi\|_\infty (2\nu - 1)/(\nu - 1)$ et seuil

$$t = t_{j,n,\alpha} = \gamma j^{\nu + \frac{1}{2}} \frac{2^{j/2}}{\sqrt{n\alpha^2}} \quad \textit{pour } \gamma \textit{ assez grand,}$$

ainsi que le mécanisme (4) ne dépendent pas des paramètres s, p, q et L de la classe de Besov, et on a les bornes supérieures suivantes :

$$\sup_{(s,p,q,L) \in \Theta} \sup_{f \in \mathcal{D}^s_{pq}(L)} \mathbb{E} \|\widetilde{f}_n - f\|_r^r \lesssim (\log n)^C \begin{cases} (n\alpha^2)^{-\frac{rs}{2s+2}}, & \textit{si } p > \frac{r}{s+1}, \\ \left(\frac{n\alpha^2}{\log(n\alpha^2)}\right)^{-\frac{r(s-1/p+1/r)}{2(s-1/p)+2}}, & \textit{si } p \leq \frac{r}{s+1}, \end{cases}$$

où $\Theta = (1/p, N) \times [1, \infty) \times [1, \infty) \times [\underline{L}, \overline{L}]$ *pour des réels* $0 < \underline{L} \leq \overline{L} < \infty$.

On a donc construit, à partir du mécanisme de privatisation défini par (4), un estimateur par seuillage adaptatif et optimal à un facteur logarithmique près.

5 Bibliographie

Donoho, D. L., Johnstone, L. M., Kerchyacharian, G. et Picard, D. (1996). Density estimation by wavelet thresholding, *The Annals of Statistics*, Vol.24, 508-539.

Duchi, J. C., Jordan, M.I. et Wainwright, M.J. (2018). Minimax optimal procedures for locally private estimation, *Journal of American Statistical Association*, 113(521):182-201.

Härdle, W., Kerkyacharian, G., Picard, D. et Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications.* Vol. 129. Lecture Notes in Statistics. Springer-Verlag, New-York, pp xviii+265.

Tsybakov, A. *Introduction to nonparametric estimation.* Springer Series in Statistics. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. Springer, New York, 2009, pp. xii+214.

Wasserman, L, et Zhou, S. (2010). A statistical framework for differential privacy, *Journal of American Statistical Association* 105, 375-389.