Hermite density deconvolution

Ousmane B Sacko

MAP5 UMR 8145, Université Paris Descartes, 45 Rue des Saints-Pères, 75006 Paris, France. Email : ousmane.sacko@parisdescartes.fr

Résumé. Considérons le modèle à bruit additif : $Z = X + \varepsilon$, où X et ε sont indépendantes. Nous construisons un nouvel estimateur de la densité de X à partir d'observations de Z, fondé sur une méthode de projection en base d'Hermite sur $\mathbb{R}$. Nous étudions le risque quadratique intégré de notre estimateur. Nous prouvons qu'il est consistant et atteint les vitesses classiques dans ce contexte. Nous proposons également une sélection de modèle pour choisir la bonne dimension dans l'espace de projection. L'estimateur résultant réalise automatiquement un compromis biais-variance.

Mots-clés. Déconvolution, base d'Hermite, sélection de modèle, estimateur non paramétrique.

Abstract. We consider the additive noise model : $Z = X + \varepsilon$, where X and ε are independent. We build a new estimator of the density of X from observations of Z, based on a projection method in Hermite basis on $\mathbb{R}$. We study the integrated quadratic risk of our estimator. Our estimator is consistent and reaches the classical rate in this context. We also propose a model selection to select the relevant dimension of the projection space. The resulting estimator realizes automatically a bias-variance compromise.

Keywords. Deconvolution, Hermite basis, nonparametric estimator, model selection.

1 Introduction

Considérons le modèle de convolution

$$Z_k = X_k + \varepsilon_k, \quad k = 1, \ldots, n,$$

où

($\mathbf{H}_1$) les $(X_k)_{k\geq 1}$ sont indépendantes et identiquement distribuées (i.i.d.) de densité f inconnue par rapport à la mesure de Lebesgue,

($\mathbf{H}_2$) les $(\varepsilon_k)_{k\geq 1}$ sont i.i.d. de densité f_ε connue, par rapport à la mesure de Lebesgue,

($\mathbf{H}_3$) les suites $(X_k)_{k\geq 1}$ et $(\varepsilon_k)_{k\geq 1}$ sont indépendantes.

On cherche à estimer f à partir des données $Z_1, \ldots, Z_n$. Notons f_Z la densité de Z_1. Sous ($\mathbf{H}_3$) nous avons que $f_Z = f * f_\varepsilon$, où $g * h$ désigne le produit de la convolution entre g et h, $g * h(x) = \int g(x-y)h(y)dy$, c'est ce qui explique le terme de "deconvolution" pour l'estimation de f. Outre la régularité de f, la régularité de f_ε influe sur la vitesse de convergence des estimations. On a les deux hypothèses classiques sur la loi du bruit.

($\mathbf{H}_4$) $\forall u \in \mathbb{R}$, $f_\varepsilon^*(u) \neq 0$, où g^* désigne la transformée de Fourier de g, $g^*(u) = \int e^{ixu} g(x)dx$.
On suppose aussi qu'il existe $c_1 \geq c_1' > 0$, et $\gamma \geq 0, \mu \geq 0, \delta \geq 0$ (avec $\gamma > 0$ si $\delta = 0$) tels que

$$c_1'(1+u^2)^\gamma e^{\mu|u|^\delta} \leq \frac{1}{|f_\varepsilon^*(u)|^2} \leq c_1(1+u^2)^\gamma e^{\mu|u|^\delta}. \tag{1}$$

Si $\delta = 0$ dans (1), la loi du bruit f_ε et les erreurs sont appelées ordinairement régulières ("ordinary smooth"), sinon elles sont dites super régulières ("super smooth").

Le problème de déconvolution a été beaucoup étudié dans la littérature. Les premiers travaux ont été menés dans le cas non adaptatif par : Carroll et Hall (1988), Fan (1991), (1993).... La procédure d'estimation adaptative a été proposée par : Pensky et Vidakovic (1999), Comte et Lacour (2011) (cas bruit inconnu), Mabon (2017) dans le cas où $X_k \geq 0$.... Fan (1993), Butucea (2004), Butucea et Tsybakov (2007a, 2007b) établissent l'optimalité des vitesses au sens minimax. Nous présentons ici une nouvelle procédure fondée sur un développement en base d'Hermite, inspiré d'idées développées dans Comte et Genon-Catalot (2017).

2 Base d'Hermite et estimateur par projection

2.1 Base d'Hermite

On définit une base orthonormée dans $\mathbb{L}^2(\mathbb{R})$, la base d'Hermite par :

$$\varphi_j(x) = c_j H_j(x) e^{-x^2/2},\ H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j}(e^{-x^2}),\ c_j = (2^j j! \sqrt{\pi})^{-1/2},\ x \in \mathbb{R},\ j \geq 0$$

où H_j est le polynôme d'Hermite de degré j. La transformée de Fourier de φ_j vérifie :

$$\varphi_j^* = \sqrt{2\pi}(i)^j \varphi_j,\ \text{où } i \text{ est le nombre complexe tel que } i^2 = -1. \tag{2}$$

Nous avons aussi par la formule d'Askey et Wainger (1965) que

$$|\varphi_j(x)| < Ce^{-\xi x^2}, \qquad |x| \geq \sqrt{2j+1}, \qquad C > 0,\ \text{avec } \xi \text{ indépendant de } x. \tag{3}$$

2.2 Estimateur par projection

Soit $S_m = \text{Vect}\{\varphi_0, \ldots, \varphi_{m-1}\}, m \geq 1$, entier, l'espace engendré par $\varphi_0, \ldots, \varphi_{m-1}$. On suppose que $f \in \mathbb{L}^1(\mathbb{R}) \cap \mathbb{L}^2(\mathbb{R})$, donc $f = \sum_{j \geq 0} a_j(f)\varphi_j$ avec $a_j(f) = \langle f, \varphi_j \rangle = \int f(x)\varphi_j(x)dx$ et la projection orthogonale de f est donnée par : $f_m = \sum_{j=0}^{m-1} a_j(f)\varphi_j$. Nous allons estimer f_m. On considère le contraste empirique défini par :

$$\gamma_n(t) = ||t||^2 - \frac{2}{n}\sum_{k=1}^n \phi_t(Z_k), \quad \phi_t(x) = \frac{1}{2\pi}\int \frac{t^*(u)}{f_\varepsilon^*(-u)} e^{-ixu} du, \quad ||t||^2 = \int t^2(x)dx. \tag{4}$$

En supposant que $\varphi_j/f_\varepsilon^*$ est intégrable, on définit notre estimateur par :

$$\widehat{f}_m = \underset{t\in S_m}{\operatorname{argmin}}\ \gamma_n(t) = \sum_{j=0}^{m-1}\widehat{a}_j\varphi_j,\ \widehat{a}_j = \frac{(-i)^j}{\sqrt{2\pi}}\int\frac{\widehat{f}_Z^*(u)}{f_\varepsilon^*(u)}\varphi_j(u)du,\ \widehat{f}_Z^*(t) = \frac{1}{n}\sum_{k=1}^{n}e^{itZ_k}. \quad (5)$$

Nous avons par le théorème de Plancherel-Parseval et d'après (2) que $\mathbb{E}[\widehat{a}_j] = a_j(f)$. Donc $\widehat{f}_m$ est un estimateur sans biais de f_m. Notons que la base d'Hermite a la particularité de rendre la quantité $\varphi_j/f_\varepsilon^*$ intégrable dans beaucoup de cas (voir (3)).

3 Risque de l'estimateur

3.1 Risque pour m fixé

Sous l'hypothèse additionnelle :
$(\mathbf{H}_5)$ $\quad ||f_Z||_\infty = \sup_{x\in\mathbb{R}}|f_Z(x)| < \infty$, on peut majorer le risque de notre estimateur.

Proposition 3.1. *(i) Sous* $(\mathbf{H}_1)$,...,$(\mathbf{H}_5)$ *et pour* $\widehat{f}_m$ *donné par (5), nous avons*

$$\mathbb{E}[||\widehat{f}_m - f||^2] \leq ||f-f_m||^2 + \frac{1}{\pi n}\int_{|u|\leq\sqrt{lm}}\frac{du}{|f_\varepsilon^*(u)|^2} + \frac{2}{n}||f_Z||_\infty\sum_{j=0}^{m-1}\int_{|u|>\sqrt{lm}}\frac{|\varphi_j(u)|^2}{|f_\varepsilon^*(u)|^2}du, \quad (6)$$

où $l > 0$ *est une constante strictement positive.*

(ii) Si de plus $l \geq 2$ *et* f_ε *satisfait (1) avec* $0 < \delta < 2$ *ou* ($\delta = 2$, *avec* $\mu < \xi$), ξ *défini en (3), alors*

$$\frac{2}{n}||f_Z||_\infty\sum_{j=0}^{m-1}\int_{|u|>\sqrt{lm}}\frac{|\varphi_j(u)|^2}{|f_\varepsilon^*(u)|^2}du = \mathcal{O}(\frac{1}{n}).$$

Le premier terme à droite de l'inégalité (6) ($||f_m - f||^2 = \sum_{j\geq m}a_j(f)^2$) est le terme de biais, il mesure la distance entre f et f_m au sens de $\mathbb{L}^2$. C'est un terme décroissant en m. Le deuxième terme est le terme principal de la variance, il croît clairement avec m. Le dernier terme vient aussi du calcul de variance, qui est négligeable d'après le point *(ii)* de la Proposition 3.1. Par conséquent le risque minimal s'obtient en faisant un compromis biais-variance.

Notons que dans le cas où $\delta = 2$ et $\mu \geq \xi$, on peut transformer les variables de façon à se ramener à $\mu < \xi$ (voir Sacko (2019)).

On peut regrouper les deux points de la Proposition 3.1 en un seul, c'est-à-dire :

$$\mathbb{E}[||\widehat{f}_m - f||^2] \leq ||f-f_m||^2 + \frac{\Delta(m)}{n} + \frac{c}{n}, \quad \Delta(m) = \frac{1}{\pi}\int_{|u|\leq\sqrt{lm}}\frac{du}{|f_\varepsilon^*(u)|^2},\ c>0,\ l\geq 2. \quad (7)$$

3.2 Vitesse de convergence

3.2.1 Vitesse sur la boule de Sobolev-Hermite

Soit $s > 0$, la boule de Sobolev-Hermite (voir Bongioanni et Torrea (2006)) de régularité s est définie par :

$$W_H^s(D) = \{\theta \in \mathbb{L}^2(\mathbb{R}), \sum_{k\geq 0} k^s a_k^2(\theta) \leq D\}, \quad a_k(\theta) = \int \theta(x)\varphi_k(x)dx, \quad D > 0.$$

Donc, si $f \in W_H^s(D)$ et au vu de (7), on a :

$$\mathbb{E}[||\widehat{f}_m - f||^2] \lesssim Dm^{-s} + \frac{\Delta(m)}{n}, \text{ où } l \geq 2 \text{ et } \Delta(m) \text{ défini par (7).}$$

Cette inégalité est la même que celle obtenue dans Comte et Lacour (2011), où m est remplacé ici par $\sqrt{m}$. Si on note m_{opt} le m qui fait le compromis biais-variance, les mêmes calculs que dans Comte et Lacour (2011) (cf. aussi Lemme 1) donnent les vitesses et les dimensions m_{opt} présentées dans le tableau suivant :

	$\delta = 0$	$0 < \delta < 2$ or $\delta = 2,\ \mu < \xi$
m_{opt}	$[n^{\frac{2}{2s+2\gamma+1}}]$	$\left[\left(\frac{\log n}{2\mu l^\delta}\right)^{\frac{2}{\delta}}\right]$
Vitesse	$n^{-\frac{2s}{2s+2\gamma+1}}$	$(\log n)^{-\frac{2s}{\delta}}$

Table 1 – Vitesse de convergence du MISE si $f \in W_H^s(D)$.

Ces vitesses sont connues pour être optimales. Elles coïncident avec celles obtenues par Fan (1993), Pensky et Vidakovic (1999), respectivement calculées sur les classes de Hölder et de Sobolev.

3.2.2 Vitesses de convergence pour des classes de fonctions spécifiques

Pour certaines classes de fonctions, on peut calculer l'ordre exact du terme de biais, qui peut être à décroissance beaucoup plus rapide que polynomiale. C'est le cas par exemple de la densité gaussienne, des mélanges de gaussiennes. L'évaluation du terme de biais s'appuie sur les calculs déjà effectués dans Belomestny *et al.* (2017). Il en ressort que la vitesse est donnée par le terme de variance. Soit

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad g_{p,\sigma}(x) = \frac{x^{2p}}{\sigma^{2p}C_{2p}} f_{0,\sigma}(x), \quad C_{2p} = \mathbb{E}\left[X^{2p}\right],$$

pour X une variable Gaussienne standard.

Proposition 3.2. *Supposons que les hypothèses* ($\mathbf{H}_1$)*,...,*($\mathbf{H}_5$) *sont vérifiées et que* f_ε *est "ordinary smooth". Pour le choix* $m_{opt} = [\log(n)/C_1]$*, avec* $C_1 = \log(2) + e\mu^2$ *si* $f = f_{\mu,1}$*,* $C_1 = \log\left(\frac{\sigma^2+1}{\sigma^2-1}\right)^2$ *si* $f = f_{0,\sigma}$*, nous avons*

$$\mathbb{E}\left[||\widehat{f}_{m_{opt}} - f||^2\right] \lesssim \frac{(\log n)^{\gamma+\frac{1}{2}}}{n},$$

où la constante γ *est donnée en (1).*

Notons que nous avons le même résultat pour $f = g_{p,\sigma}$ et pour des classes de mélanges gaussiens en moyenne ou en variance définies dans Belomestny *et al.* (2017). Cette vitesse est la même que celle obtenue par Butucea (2004) pour des fonctions "super-smooth".

Cependant dans tous les cas précédents le choix $m = m_{opt}$ dépend de la régularité de f qui est inconnue, donc ne peut être utilisé en pratique. C'est pourquoi nous cherchons un autre moyen pour faire le compromis biais-variance de manière automatique.

4 Estimateur adaptatif et sélection de modèle

Dans cette section, notre objectif est de proposer un choix automatique de la dimension m qui réalise le compromis biais-variance dans (7). La procédure que nous allons décrire ne dépend pas de la régularité de f, mais uniquement des données $Z_1, \ldots, Z_n$. Soit $\mathcal{M}_n$ la collection de modèles telle que la variance est bornée :

$$\mathcal{M}_n = \{m \in \mathbb{N}\backslash\{0\}, \Delta(m) \leq n\}.$$

La procédure adaptative cherche à réaliser le compromis biais-variance en remplaçant chacun par une quantité calculable. A la variance, on substitue sa borne $\Delta(m)/n$. Pour le terme de biais, comme $||f - f_m||^2 = ||f||^2 - ||f_m||^2$, on estime seulement -$||f_m||^2$ ($||f||^2$ est une constante) : on lui substitue donc $-||\widehat{f}_m||^2 = \gamma_n(\widehat{f}_m)$. On pose

$$\widehat{m} = \underset{m \in \mathcal{M}_n}{\operatorname{argmin}}\{\gamma_n(\widehat{f}_m) + \operatorname{pen}(m)\}, \tag{8}$$

où γ_n est défini par (4) et pen(m) est une fonction croissante définie par :

$$\operatorname{pen}(m) = \begin{cases} \kappa\frac{\Delta(m)}{n}, & \text{si } f_\varepsilon \text{ est "ordinary smooth" ou "super smooth" avec } \delta < \frac{1}{2}, \\ 2\kappa\left(1 + 24\mu l^{\delta/2} m^{\delta-\frac{1}{2}}\right)\frac{\Delta(m)}{n} & \text{si } f_\varepsilon \text{ est "super smooth" avec } \delta \geq \frac{1}{2}, \end{cases} \tag{9}$$

où $\kappa > 0$, une constante à calibrer, μ, δ donnés par (1) et $l \geq 2$ dans la Proposition 3.1. La théorie nous dit que $\kappa > 17$ convient. Donc la procédure peut s'implémenter dès qu'on donne une valeur à κ : en pratique cette valeur est calibrée via des simulations préliminaires.

On peut montrer le théorème suivant.

Théorème 4.1. *Supposons* ($\mathbf{H}_1$),...,($\mathbf{H}_5$) *vérifiées et* f_ε *de carré intégrable. Soit* $\text{pen}(m)$ *définie par (9),* $\widehat{f}_m$ *défini par (5) et* $\widehat{m}$ *choisi selon (8). Alors il existe une constante* κ_0*, telle que* $\forall\kappa \geq \kappa_0$*, l'estimateur* $\widehat{f}_{\widehat{m}}$ *satisfait*

$$\mathbb{E}\left[||\widehat{f}_{\widehat{m}} - f||^2\right] \leq C \inf_{m\in\mathcal{M}_n} \left(||f - f_m||^2 + \text{pen}(m)\right) + \frac{C'}{n}, \tag{10}$$

où C *est une constante (C=4 convient) et* C' *une constante qui dépend de* f_ε*.*

Remarque 4.2. *L'estimateur* $\widehat{f}_{\widehat{m}}$ *réalise automatiquement un compromis biais-variance, l'inégalité (10) nous dit que* $\widehat{f}_{\widehat{m}}$ *est aussi performant que le meilleur modèle dans la collection, à une constante multiplicative près.*

Dans l'article (Sacko (2019)) des simulations illustrent la performance de la méthode, et montrent qu'elle est très compétitive.

Bibliographie

Abramowitz, M. and Stegun, I.A. (1965). Handbook of mathematical functions : with formulas, graphs, and mathematical tables. Courier Corporation 55.

Askey, R. and Wainger, S. (1965). Mean convergence of expansion in Laguerre and Hermite Series, *Amer. J. Math* 87, 695–708.

Belomestny, D., Comte, F. and Genon-Catalot, V. (2017). Sobolev-Hermite versus Sobolev nonparametric density estimation on $\mathbb{R}$, *Annals of the Institute of Statistical Mathematics*, 71(1), 29–62.

Bongioanni, B. and Torrea, J.L. (2006). Sobolev spaces associated to the harmonic oscillator *Proc. Indian. Acad. Sci (math. Sci.)*, 116, 337–360.

Butucea, C. (2004). Deconvolution of supersmooth densities with smooth noise. *Canad. J. Statist* 32(2), 181–192.

Butucea, C. and Tsybakov, A.B. (2007). Sharp optimality in density deconvolution with dominating bias. I, *Teor. Veroyatn*, Primen 52(1), 111–128.

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. J. Amer. Statist. Assoc. 83(404), 1184–1186.

Comte, F. and Genon-Catalot, V. (2018). Laguerre and Hermite bases for inverse problems. *J. Korean Statist. Soc.* 47(3), 273–296.

Comte, F. and Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *J. R. Stat. Soc. Ser B Stat. Methodol* 73(4), 601–627.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems, *Ann. Statist*,19(3), 1257–1272.

Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem, *Ann. Statist.* 21(2), 600–610.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution, *Ann. Statist.* 27(6), 2033–2053.

Sacko, O.B. (2019). Hermite density deconvolution, *Preprint hal-01978591.*