Estimation adaptative de densité dans des problèmes de déconvolution avec une loi des erreurs inconnue

Gwennaëlle Mabon [1,2] & Johanna Kappus [3]

[1] *CREST*
3 avenue Pierre Larousse
92245 Malakoff, France
gwennaelle.mabon@ensae.fr
[2] *MAP5, Université Paris Descartes*
45 rue des Saints-Pères
75006 Paris, France
[3] *Institut für Mathematik, Universität Rostock*
18051 Rostock, Germany
johanna.kappus@uni-rostock.de

Résumé. Dans ce travail, nous nous intéressons au problème de déconvolution de densité lorsque la loi de l'erreur est inconnue : $Y = X + \varepsilon$. Afin d'estimer la densité f de X, il faut tout de même avoir accès à une information sur le bruit tel un échantillon préliminaire $(\varepsilon_{-j})_{1\leq j\leq M}$ observé indépendamment des Y_j. Ainsi nous supposons que nous avons des observations *i.i.d.* de deux échantillons : $Y_1, \dots, Y_n$ et $\varepsilon_{-M}, \dots, \varepsilon_{-1}$ respectivement de densité f_Y et f_ε. Le but est d'estimer non-paramétriquement f à partir des observations Y_j. Pour cela, nous introduisons des estimateurs de type noyau basé sur une approche Fourier du problème ainsi qu'un nouvel estimateur de f_ε. Nous présentons aussi une majoration du risque $\mathbb{L}^2$. L'intérêt de ce travail repose sur le choix adaptatif d'un paramètre de lissage déterminé par un critère de pénalisation. Nous proposons ainsi une procédure adaptative pour n et M quelconques. Ce dernier résultat est nouveau dans la littérature.

Mots-clés. Estimation adaptative. Déconvolution. Estimation de densité. Risque quadratique. Méthodes non-paramétrique.

Abstract. A density deconvolution problem with unknown distribution of the errors is considered. To make the target density identifiable, one has to assume that some additional information on the noise is available. We consider the framework where some additional sample of the pure noise is available. We introduce kernel estimators and present upper risk bounds. The focus of this work lies on the optimal data driven choice of the smoothing parameter using a penalization strategy.

Keywords. Adaptive estimation. Deconvolution. Density estimation. Mean square risk. Nonparametric methods.

1 Introduction

Nous considérons le modèle classique de déconvolution de densité

$$Y_j = X_j + \varepsilon_j, \quad j = 1, \dots, n. \tag{1}$$

Nous supposons que f la densité des X_j par rapport à la mesure de Lebesgue appartient à $\mathbb{L}^2(\mathbb{R})$. Le but est d'estimer f à partir des données Y_j.

Le problème de déconvolution de densité a été largement étudié dans la littérature. Les vitesses de convergence et leur optimalité ont été étudiées, par exemple, par Caroll et Hall (1988), Stefanski et Carroll (1990), Fan (1991), Efromovich (1997) pour des estimateurs à noyaux, Butucea (2004), Butucea et Tsybakov (2008a) pour l'optimalité au sens minimax. Néanmoins, ces auteurs travaillent sous l'hypothèse que la densité du bruit est connue, ce qui ne constitue pas toujours un cadre réaliste.

Dans le cas présent, nous supposons que la densité du bruit est inconnue. Il faut cependant avoir une information sur le bruit. Nous nous plaçons donc dans le cas le plus communément étudié dans la littérature lorsqu'il existe un échantillon préliminaire du bruit $(\varepsilon_{-j})_{1\leq j\leq M}$ indépendant de $(Y_j)_{1\leq j\leq n}$. En pratique, nous pouvons disposer de tels échantillons en astrophysique ou en spectrométrie. Dans ce contexte, nous nous référons aux travaux de Comte et Lacour (2011), Schwarz et Johannes (2012), Kappus (2014).

Dans ce papier, nous introduisons un estimateur à noyau basé sur une approche Fourier du problème ainsi qu'un estimateur alternatif de f_ε nécessaire à notre procédure adaptative. Nous présentons aussi une majoration du risque $\mathbb{L}^2$. L'intérêt de ce travail repose donc sur le choix adaptatif d'un paramètre de lissage déterminé par un critère de pénalisation. Nous proposons ainsi une procédure adaptative pour n et M quelconques. Ce dernier résultat est nouveau dans la littérature. Il est à noter que la méthode proposée s'applique encore dans le cas où l'on peut estimer la transformée de Fourier à partir de données répétées. La méthodologie complète est présentée dans Kappus et Mabon (2013) ainsi qu'une illustration numérique des résultats.

2 Procédure d'estimation

Notations. Pour deux réels a et b, on note $a \vee b = \max(a,b)$ et $a \wedge b = \min(a,b)$. Pour deux fonctions φ, $\psi : \mathbb{R} \to \mathbb{C}$ dans $\mathbb{L}^1(\mathbb{R}) \cap \mathbb{L}^2(\mathbb{R})$, on note $\|\varphi\|$ la norme $\mathbb{L}^2$ de φ définie par $\|\varphi\|^2 = \int_{\mathbb{R}} |\varphi(x)|^2 \,\mathrm{d}x$, $\langle \varphi, \psi \rangle$ le produit scalaire entre φ et ψ défini par $\langle \varphi, \psi \rangle = \int_{\mathbb{R}} \varphi(x)\overline{\psi(x)} \,\mathrm{d}x$. La transformée de Fourier φ^* est définie par $\varphi^*(x) = \int e^{ixu}\varphi(u) \,\mathrm{d}u$. De plus, si φ^* appartient à $\mathbb{L}^1(\mathbb{R}) \cap \mathbb{L}^2(\mathbb{R})$, alors la fonction φ est la transformée de Fourier inverse de φ^* et s'écrit $\varphi(x) = 1/(2\pi) \int e^{-ixu}\varphi^*(u) \,\mathrm{d}u$. Enfin le produit de convolution $*$ est définie comme $(\varphi * \psi)(x) = \int \varphi(x-u)\psi(u) \,\mathrm{d}u$.

Dans le Modèle (1) sous hypothèse d'indépendance, il est clair que $f_Y = f * f_\varepsilon$ ce qui implique $f_Y^* = f^* f_\varepsilon^*$. Nous faisons l'hypothèse suivante

(A1) $\forall x \in \mathbb{R}$, $f_\varepsilon^*(x) \neq 0$ et $f_\varepsilon \in \mathbb{L}^2(\mathbb{R})$.

Sous **(A1)**, nous avons l'égalité $f^* = f_Y^*/f_\varepsilon^*$. Si f_ε^* est connue, nous pouvons estimer f^* avec $\hat{f}_Y^*/f_\varepsilon^*$ où $\hat{f}_Y^*$ est un simple estimateur plug-in. Nous n'avons qu'à appliquer la transformée de Fourier inverse pour obtenir un estimateur de f. Cependant, $1/f_\varepsilon^*$ n'est pas intégrable sur tout $\mathbb{R}$. Il faut alors appliquer une régularisation du problème, par exemple, en introduisant un cutoff spectral comme suit : $\hat{f}_m(x) = 1/(2\pi)\int_{|u|\leq\pi m} e^{-iux}\hat{f}_Y^*(u)/f_\varepsilon^*(u)\,\mathrm{d}u$. Cependant, nous supposons dans ce travail que la densité de l'erreur est inconnue. Nous nous plaçons alors dans le cas où il existe un échantillon préliminaire du bruit $(\varepsilon_{-j})_{1\leq j\leq M}$ indépendant de $(Y_j)_{1\leq j\leq n}$. Nous définissons alors les estimateurs empiriques de f_ε^* et f_Y^*

$$\forall x \in \mathbb{R},\ \hat{f}_\varepsilon^*(x) = \frac{1}{M}\sum_{j=1}^{M} e^{ix\varepsilon_{-j}} \quad \text{et} \quad \hat{f}_Y^*(x) = \frac{1}{n}\sum_{j=1}^{n} e^{ixY_j}. \tag{2}$$

3 Majoration du risque $\mathbb{L}^2$

Ainsi nous pouvons proposer un estimateur de f_ε^* grâce à un échantillon préliminaire du bruit. Cependant, $\hat{f}_\varepsilon^*$ étant au dénominateur, il faut éviter qu'il prenne de trop petites valeurs car cela entraîne une instabilité dans l'estimation de $\hat{f}_m^*$. Pour cela, nous suivons la méthode initiée par Neumann (1997) et reprise par Comte et Lacour (2011) en introduisant l'estimateur tronqué

$$\frac{1}{\tilde{\varphi}(x)} = \frac{\mathbb{I}\left\{|\hat{f}_\varepsilon^*(x)| \geq M^{-1/2}\right\}}{\hat{f}_\varepsilon^*(x)}. \tag{3}$$

En utilisant la transformée de Fourier inverse avec un cutoff spectral m, on a alors

$$\hat{f}_{m,M,n}(x) = \frac{1}{2\pi}\int_{-\pi m}^{\pi m} e^{-iux}\frac{\hat{f}_Y^*(u)}{\tilde{\varphi}(u)}\,\mathrm{d}u. \tag{4}$$

On obtient alors la majoration suivante du risque intégré :

Proposition 3.1 (Comte et Lacour (2011)) *Si l'hypothèse* ***(A1)*** *est vérifiée, alors pour $\hat{f}_{m,M,n}$ défini par (4), il existe une constante positive C telle que*

$$\mathbb{E}\left\|f - \hat{f}_{m,M,n}\right\|^2 \leq \|f - f_m\|^2 + C\left(\frac{1}{n}\int_{-\pi m}^{\pi m}\frac{\mathrm{d}u}{|f_\varepsilon^*(u)|^2} + \frac{1}{M}\int_{-\pi m}^{\pi m}\frac{|f^*(u)|^2}{|f_\varepsilon^*(u)|^2}\,\mathrm{d}u\right). \tag{5}$$

Les deux premiers termes de l'Équation (5) correspondent aux termes usuels apparaissant dans le problème de déconvolution de densité à bruit connu (cf. Comte et al. (2006)) : un terme de biais et un majorant de la variance. Le dernier terme en $1/M$ est quant à lui dû à l'estimation de f_ε^*. Pour les vitesses de convergence associées à ce problème, nous renvoyons à Comte et Lacour (2011). Une étude complète, dans le cas bruit connu, peut être trouvée dans Lacour (2006).

4 Procédure d'estimation adaptative

Le but de ce papier est de présenter une procédure adaptative permettant d'obtenir un estimateur $\hat{f}_{\hat{m}}$ réalisant automatiquement le compromis biais-variance de l'Équation (5). Pour cela, nous allons détermininer une pénalité $\widehat{\mathrm{pen}}$ ayant à peu près le même ordre de grandeur que le terme de variance de l'Équation (5). Ensuite, nous pourrons sélectionner $\hat{m}$ comme le minimiseur d'un critère pénalisé en suivant une méthode de sélection de modèle.

Pour cela, nous introduisons un estimateur alternatif de f_ε^* inspiré de Kappus (2014). Suivant l'idée de Neumann (1997), nous modifions le seuil de la manière suivante afin de pouvoir utiliser des inégalités de concentration de type Bernstein dans sa version intégrée

$$k_M(x) = \kappa \left(\log M\right)^{1/2} w(x)^{-1} M^{-1/2} \tag{6}$$

avec $w(x) = (\log(e + |x|))^{-\frac{1}{2}-\delta}$ et $\delta > 0$. Cette fonction w (proposée par Neumann et Reiß (2009)) joue un rôle important dans la méthodologie introduite par Kappus (2014). L'estimateur de f_ε^* retenu est donc défini comme

$$\tilde{f}_\varepsilon^*(x) = \begin{cases} \hat{f}_\varepsilon^*(x) & \text{si} \quad |\hat{f}_\varepsilon^*(x)| \geq k_M(x), \\ k_M(x) & \text{sinon.} \end{cases} \tag{7}$$

On peut alors estimer f_m de la manière suivante

$$\hat{f}_m(x) = \frac{1}{2\pi} \int_{-\pi m}^{\pi m} e^{-ixu} \frac{\hat{f}_Y^*(u)}{\tilde{f}_\varepsilon^*(u)} \, \mathrm{d}u. \tag{8}$$

Maintenant, nous introduisons les termes intervenant dans la pénalité proposée à la suite de ces définitions. Ces termes correspondent à une fonction w près aux termes de variance de l'Équation (5) dans leur version déterministe et aléatoire.

$$\Delta(m) = \frac{1}{2\pi} \int_{-\pi m}^{\pi m} \frac{w(u)^{-2}}{|f_\varepsilon^*(u)|^2} \, du \quad \text{et} \quad \hat{\Delta}(m) = \frac{1}{2\pi} \int_{-\pi m}^{\pi m} \frac{w(u)^{-2}}{|\tilde{f}_\varepsilon^*(u)|^2} \, \mathrm{d}u$$

$$\Delta^f(m) = \frac{1}{2\pi} \int_{-\pi m}^{\pi m} \frac{w(u)^{-2} |f^*(u)|^2}{|f_\varepsilon^*(u)|^2} \, du \quad \text{et} \quad \hat{\Delta}^f(m) = \frac{1}{2\pi} \int_{-\pi m}^{\pi m} \frac{w(u)^{-2} |\hat{f}_Y^*(u)|^2}{|\tilde{f}_\varepsilon^*(u)|^4} \, du$$

La pénalité empirique est alors définie comme

$$\widehat{\mathrm{pen}}(m) = 12\lambda_1^2(m, \hat{\Delta}(m)) \frac{\hat{\Delta}(m)}{n} + 16\kappa^2 \log(Mm) \frac{\hat{\Delta}^f(m)}{M} \tag{9}$$

où κ est défini dans l'Équation (6) et la pénalité déterministe

$$\mathrm{pen}(m) = 12\lambda_1^2(m, \Delta(m)) \frac{\Delta(m)}{n} + 16\kappa^2 \log(Mm) \frac{\Delta^f(m)}{M} \tag{10}$$

avec $\lambda_1(m,D) = \max\left\{\sqrt{8\log\left(1+Dm^2\right)}, \frac{16\sqrt{2}}{3\sqrt{n}}\log\left(1+Dm^2\right)\right\}$.

Remarquons ici l'originalité de notre procédure par rapport à Comte et Lacour (2011) dont la procédure adaptative est associée à la pénalité suivante

$$\widetilde{\text{pen}}(m) = 128\left(\frac{\log\left(\int_{-\pi m}^{\pi m}|\tilde{\varphi}(u)|^{-2}\,\mathrm{d}u\right)}{\log(1+m)}\right)^2 \frac{\int_{-\pi m}^{\pi m}|\tilde{\varphi}(u)|^{-2}\,\mathrm{d}u}{n} \tag{11}$$

et qui n'est valable que lorsque la taille d'échantillon de ε est plus grande que celle des observations. Une hypothèse qui n'est pas nécessaire dans ce papier. De plus, leur procédure ne prend pas en compte explicitement le terme en $\int_{-\pi m}^{\pi m}|f^*(u)|^2/|f_\varepsilon^*(u)|^2\,\mathrm{d}u$ de l'Équation (5).

Enfin, nous sélectionnons $\hat{m}$ comme le minimiseur du critère pénalisé suivant

$$\hat{m} = \underset{m\in\mathcal{M}_n}{\operatorname{argmin}}\left\{-\left\|\hat{f}_m\right\|^2 + \widehat{\text{pen}}(m)\right\}, \tag{12}$$

Nous obtenons alors la majoration suivante du risque dans un cadre non-asymptotique pour une collection de modèle $\mathcal{M}_n = \{1,\ldots,n\}$ et le résulat principal de ce papier.

Théorème 4.1 *Sous l'hypothèse* ***(A1)****, pour $\hat{f}_{\hat{m}}$ défini par (8) et (12), il existe des constantes positives C^{ad} et C telles que*

$$\mathbb{E}\|f-\hat{f}_{\hat{m}}\|^2 \le C^{ad}\inf_{m\in\mathcal{M}_n}\left\{\|f-f_m\|^2 + \text{pen}(m)\right\} + C\left(\frac{1}{n}+\frac{1}{M}\right). \tag{13}$$

Dans les problèmes de déconvolution, les vitesses de convergence sont classiquement complexes et dépendent de la régularité de la fonction à estimer f et de celle du bruit f_ε. Pour les calculer, il faut déterminer la valeur optimale de m en fonction de n à partir des ordres de grandeur du biais et des termes de variance de l'Équation (5). Cette résolution est déjà complexe dans le cas où f_ε est supposée connue. Même une fois cette difficulté surmontée, il est clair que le m optimal dépend des paramètres de régularité des fonctions f et f_ε. Il ne peut donc pas être pas calculé. L'Équation (13) qui est une inégalité oracle est alors de la plus grande importance car elle garantit que le compromis biais-variance est réalisé automatiquement dans un contexte non-asymptotique. Ainsi les vitesses de convergence sont atteintes sans avoir à préciser le contexte. Cependant, il est légitime de se demander quel est effet de l'adaptation sur les vitesses. Pour y répondre, il faut comparer les termes des Équations (5) et (13). Plus précisément, puisque le biais $\|f-f_m\|^2$ ne change pas et que les termes $n^{-1}+M^{-1}$ sont négligeables, il suffit de comparer pen(m) à

$$\frac{\Delta(m)}{n} + \frac{\log M}{M}\Delta^f(m).$$

Il est clair que la différence repose sur les termes logarithmiques. La perte occasionnée est ainsi négligeable.

Remarques conclusives. La méthode présentée repose sur un modèle avec de faibles hypothèses et reste donc générale. Comme le montre Kappus et Mabon (2014), cette méthode peut encore être utilisée dans le cadre du modèle de déconvolution en présence de données répétées. De plus, il semble que cette méthodologie recèle de multiples prolongements dès lors que le modèle peut être vu comme un modèle de déconvolution.

Bibliographie

[1] Butucea, C. (2004). Deconvolution of supersmooth densities with smooth noise. *The Canadian Journal of Statistics*, 32(2):181–192.

[2] Butucea, C. and Tsybakov, A. (2008a). Sharp optimality in density deconvolution with dominating bias I. *Theory Proba. Appl.*, 52(1):24–39.

[3] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186.

[4] Comte, F. and Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *Journal of the Royal Statistical Society: Series B*, 73:601–627.

[5] Comte, F., Rozenholc, Y., and Taupin, M.-L. (2006). Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, (34):431–452.

[6] Efromovich, S. (1997). Density estimation for the case of supersmooth measurement errors. *Journal of the American Statistical Association*, 92:526–535.

[7] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272.

[8] Johannes, J. (2009). Deconvolution with unknown error distribution. *The Annals of Statistics*, 37(5a):2301–2323.

[9] Johannes, J. and Schwarz, M. (2012). Adaptive circular deconvolution by model selection under unknown error distribution. arXiv:0912.1207v2. To appear in Bernoulli.

[10] Kappus, J. (2014). Adaptive nonparametric estimation for Lévy processes observed at low frequency. *Stochastic Processes and their Applications*, 124:730–758.

[11] Kappus, J. and Mabon, G. (2013). Adaptive density estimation in deconvolution problems with unknown error distribution. Preprint, hal-00915982.

[12] Lacour, C. (2006). Rates of convergence for nonparametric deconvolution. *C. R. Acad. Sci. Paris*, Ser. I 342.

[13] Neumann, M. and Reiß, M. (2009). Nonparametric estimation for Lévy processes from low frequency observations. *Bernoulli*, 15(1):223–248.

[14] Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7(4):307–330.

[15] Stefanski, S. and Carroll, R. (1990). Deconvoluting kernel density estimators. *Statistics*, 21:169–184.