Open access • Journal Article • DOI:10.1111/J.1467-9469.2009.00653.X

Log-density deconvolution by wavelet thresholding — Source link

Jérémie Bigot, Sébastien Van Bellegem

Institutions: Institut de Mathématiques de Toulouse, University of Toulouse

Published on: 01 Dec 2009 - Scandinavian Journal of Statistics (Blackwell)

Topics: Estimator, Wavelet, Density estimation, Information projection and Kullback–Leibler divergence

Working Paper Series n° 635

February, 2009

"Log-Density Deconvolution by Wavelet Thresholding"

Jérôme BIGOT and
Sébastien VAN BELLEGEM

Institut d'Economie Industrielle (IDEI) – Manufacture des Tabacs
Aile Jean-Jacques Laffont – 21, allée de Brienne – 31000 TOULOUSE – FRANCE
Tél. + 33(0)5 61 12 85 89 – Fax + 33(0)5 61 12 86 37 – **www.idei.fr** – contact@cict.fr

Log-density deconvolution by wavelet thresholding

Jérémie Bigot & Sébastien Van Bellegem

This version: February, 11, 2009

Abstract

This paper proposes a new wavelet-based method for deconvolving a density. The estimator combines the ideas of nonlinear wavelet thresholding with periodised Meyer wavelets and estimation by information projection. It is guaranteed to be in the class of density functions, in particular it is positive everywhere by construction. The asymptotic optimality of the estimator is established in terms of rate of convergence of the Kullback-Leibler discrepancy over Besov classes. Finite sample properties is investigated in detail, and show the excellent empirical performance of the estimator, compared with other recently introduced estimators.

Keywords: Deconvolution, Wavelet thresholding, Adaptive estimation, Information projection, Kullback-Leibler divergence, Besov space
AMS classifications: Primary 62G07; secondary 42C40, 41A29

Affiliations

Jérémie Bigot, Institut de Mathématiques de Toulouse, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse Cedex 9, France, `Jeremie.Bigot@math.ups-tlse.fr`

Sébastien Van Bellegem, Toulouse School of Economics (Gremaq), 21, Allée de Brienne, F-31200 Toulouse, France, `svb@tse-fr.eu`

Ackowledgements

This work was supported by the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy). We gratefully ackowledge Yves Rozenholc for providing the Matlab code to compute the model selection estimator, Marc Raimondo for providing the Matlab code for translation invariant deconvolution, and Anestis Antoniadis for helpful comments .

1 Introduction

Density deconvolution arises when the probability density of a random variable X has to be estimated from an independent and identically distributed (iid) sample contaminated by some independent additive noise. Namely, the observations at hand, denoted by Y_i for $i = 1, \ldots, n$, are such that $Y_i = X_i + \epsilon_i, i = 1, \ldots, n$, where X_i are iid variables with unknown density f^X, and ϵ_i is an additive random error. The number n represents the sample size and the contamination variables ϵ_i are supposed iid with a known density function f^ϵ, and independent from the X_i's. In this setting, the density function f^Y of the observed sample Y_i can be written as a convolution between the density of interest f^X, and the density of the additive noise f^ϵ, i.e.

$$f^Y(y) = f^X \star f^\epsilon(y) := \int f^X(u) f^\epsilon(y-u)du, \quad y \in \mathbb{R} . \tag{1.1}$$

The problem of estimating the probability density f^X relates to classical nonparametric methods of estimation, but the indirect observation of the data leads to different optimality properties, for instance in terms of rate of convergence. Among the nonparametric methods of deconvolution, one can find estimation by model selection (e.g. Comte *et al.*, 2006), wavelet thresholding (e.g. Fan and Koo, 2002), kernel smoothing (e.g. Carroll and Hall, 1988), spline deconvolution (e.g. Koo, 1999) or spectral cut-off (e.g. Johannes *et al.*, 2007). However, a problem frequently encountered is that the proposed estimator is not everywhere positive, therefore is not a valid probability density. The main goal of the present paper is to introduce an estimator that is automatically a valid density, in particular because it is guaranteed to be positive. The proposed solution uses wavelet thresholding combined with information projection techniques, and is computationally simple.

The advantage of wavelet methods is their ability in estimating local features of the density, such as peaks or local discontinuities. Wavelet methods for deconvolution have received a special attention in the recent literature. Optimality of the nonlinear wavelet estimator has been established in Fan and Koo (2002), but the given estimator is not computable since it depends on an integral in the frequency domain that cannot be calculated in practice. Other wavelet estimators are presented in Johnstone *et al*. (2004) and De Canditiis and Pensky (2006), see also the references therein. Our estimator combines wavelet thresholding with information projection that guarantees the solution to be positive. This technique was studied by Barron and Sheu (1991) for the approximation of density functions by sequences of exponential families. An extension of this method to linear inverse problems has been studied in Koo and Chung (1998) using expansions in Fourier series.

It is well-known that the difficulty of the deconvolution problem is quantified by the smoothness of the noise density f^{ϵ}. If $f_{\ell}^{Y}, f_{\ell}^{X}$ and f_{ℓ}^{ϵ} denote the Fourier coefficients of the densities f^{Y}, f^{X} and f^{ϵ} respectively, then the convolution equation (1.1) is equivalent to $f_{\ell}^{Y} = f_{\ell}^{X} \cdot f_{\ell}^{\epsilon}$. Depending how fast the Fourier coefficients f_{ℓ}^{ϵ} tend to zero, the reconstruction of f_{ℓ}^{X} will be more or less accurate. In this paper, we consider the case where the f_{ℓ}^{ϵ}'s have a polynomial decay which is usually referred to as ordinary smooth convolution (see e.g. Fan (1991)):

Assumption 1.1 *The Fourier coefficients of f^{ϵ} decay at a polynomial rate i.e. there exist constants $c_1, c_2 > 0$ and a real $\nu \geqslant 0$ such that $c_1|\ell|^{-\nu} \leqslant |f_{\ell}^{\epsilon}| \leqslant c_2|\ell|^{-\nu}$.*

The L^2-rate of convergence that can be expected from a linear or a nonlinear wavelet estimator depends on this assumption and are well-studied in the literature, see e.g. Pensky and Vidakovic (1999); Fan and Koo (2002).

After recalling some general results on Meyer wavelets, we define in Section 3 our linear and nonlinear wavelet estimators by information projection. This paper demonstrates two important features of the non linear estimator. First we prove in Section 4 that its asymptotic rate of convergence, measured in the Kullback-Leibler divergence, is optimal over Besov balls $F_{p,q}^{s}(M)$ (defined below). Moreover, the resulting estimator is positive by construction and shows excellent finite sample properties. As we show in Section 5, it outperforms some of the best nonparametric estimators recently published in the literature.

2 Meyer wavelets for deconvolution

In this paper, we assume that the support of f^{X} is compact and included in $[0,1]$. Of course, this is not an assumption that would hold in many practical applications and it is mainly made for mathematical convenience.

Wavelet systems provide unconditional bases for Besov spaces. Using wavelets, one can characterize whether or not f^X belongs to a Besov space by a condition on the absolute value of the wavelet coefficients of f^X. Assume that (ϕ, ψ) denotes some scaling and wavelet functions that have enough regularity and vanishing moments. Let $s \geq 0$, $p, q \geq 1$, and if $\sigma = s + (1/2 - 1/p) \geqslant 0$, define the norm $\|f^X\|^q_{s,p,q} = \sum_{j=0}^{\infty}(2^{j\sigma p} \sum_{k=0}^{2^j-1} |\langle f^X, \psi_{j,k}\rangle|^p)^{q/p}$. It can be shown (Meyer, 1992) that this norm is equivalent to the norm in the Besov space $B^s_{p,q}$. The parameter s is related to the smoothness of f^X. In particular if f^X is piecewise C^α with a finite number of discontinuities, then f^X belongs to $B^s_{p,q}$ for all $s < \alpha$ and p sufficiently small.

The estimator we shall define in the next section is based on the wavelet decomposition of functions in $L^2([0,1])$ using periodised Meyer wavelets. This wavelet basis is derived through the periodisation of the Meyer wavelet basis of $L^2(\mathbb{R})$. This basis is constructed from a scaling function ϕ with Fourier transform

$$\tilde{\phi}(\omega) = \begin{cases} \tilde{h}(\omega/2)/\sqrt{2} & \text{if} \quad |\omega| \leqslant 4\pi/3, \\ 0 & \text{if} \quad |\omega| > 4\pi/3, \end{cases}$$

where $\tilde{h} : \mathbb{C} \to \mathbb{R}$ is a smooth function (see Meyer (1992), Johnstone *et al.* (2004) for further details). In the simulations below, $\tilde{h}$ is a cubic function known as the *Meyer window* (e.g. Mallat, 1998, p. 248).

Meyer wavelets are therefore band-limited which makes them very useful for deconvolution problems. Indeed, let (ϕ, ψ) be the periodised Meyer scaling and wavelet function respectively. Scaling and wavelet functions at scale j (i.e. resolution level 2^j) will be denoted by ϕ_λ and ψ_λ, where the index λ summarizes both the usual scale and space parameters j and k (i.e. $\lambda = (j,k)$ and $\psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$). The notation $|\lambda| = j$ will be used to denote a wavelet at scale j, while $|\lambda| < j$ denotes some wavelet at scale j', with $0 \leqslant j' < j$. For any function f^X of $L^2([0,1])$, its wavelet decomposition can be written as $f^X = \sum_{|\lambda|=j_0} c_\lambda \phi_\lambda + \sum_{j=j_0}^{\infty} \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda$, where $c_\lambda = \langle f^X, \phi_\lambda\rangle = \int_0^1 f^X(u)\phi_\lambda(u)du$, $\beta_\lambda = \langle f^X, \psi_\lambda\rangle = \int_0^1 f^X(u)\psi_\lambda(u)du$ and j_0 denotes the usual coarse level of resolution. Let $e_\ell(x) = \exp(2\pi i \ell x)$, $\ell \in \mathbb{Z}$ and denote by $f^X_\ell = \langle f^X, e_\ell\rangle$ the Fourier coefficients of a function $f^X \in L^2([0,1])$. Then, if we denote the Fourier coefficients of ψ_λ by $\psi^\lambda_\ell = \langle \psi_\lambda, e_\ell\rangle$ we obtain with the Plancherel's identity that $\beta_\lambda = \langle f^X, \psi_\lambda\rangle = \sum_\ell f^X_\ell \psi^\lambda_\ell$.

Given that the Meyer wavelets ψ_λ are band-limited, the above sum only involves a finite number of terms. Now, if we denote by $f^\epsilon_\ell = \mathbb{E}(e^{-2\pi i \ell \epsilon_1})$ the characteristic function of the ϵ_j's and by $f^Y_\ell = \mathbb{E}(e^{-2\pi i \ell Y_1})$ the characteristic function of the Y_j's , we have by independence of X_1 and ϵ_1 that $f^Y_\ell = \mathbb{E}(e^{-2\pi i \ell Y_1}) = \mathbb{E}(e^{-2\pi i \ell \epsilon_1})\mathbb{E}(e^{-2\pi i \ell X_1}) = f^\epsilon_\ell f^X_\ell$. An unbiased estimator of β_λ is thus given by

$$\hat{\beta}_\lambda = \sum_\ell \left(\frac{\psi^\lambda_\ell}{f^\epsilon_\ell} \right) \left(\frac{1}{n} \sum_{j=1}^{n} \exp(-2\pi i \ell Y_j) \right). \tag{2.1}$$

provided that the f^ϵ_ℓ's are non-zero and have a sufficiently smooth decay as ℓ tends to infinity. Analogously the estimators of the scaling coefficients c_λ is defined using the scaling function ϕ instead of ψ.

3 Estimation by information projection

3.1 Linear and nonlinear wavelet estimators

Based on the coefficients $\hat{c}_\lambda$ and $\hat{\beta}_\lambda$, several estimators of the unknown density f^X can be studied. First of all, the linear estimator is such that

$$\hat{f}^X_L = \sum_{|\lambda|=j_0} \hat{c}_\lambda \phi_\lambda + \sum_{j=j_0}^{j_1} \sum_{|\lambda|=j} \hat{\beta}_\lambda \psi_\lambda$$

This estimator was first studied by Pensky and Vidakovic (1999), who showed that for an appropriate scale j_1, it achieves the optimal rate of convergence among the class of linear estimators. In the ordinary smooth situation (Assumption 1.1), the choice of j_1 is such that $2^{j_1} = \mathcal{O}(n^{1/2s+2\nu+1})$ if f^X belongs to the Sobolev space H^s. Note that this choice is not adaptive because j_1 depends on the unknown smoothness class of f^X.

In contrast, adaptive nonlinear estimators by wavelet thresholding have been developed and they can achieve near-optimal rate of convergence (up to logarithmic factors). To simplify the notations, hereafter we write $(\psi_\lambda)_{|\lambda|=j_0-1}$ for the scaling functions $(\phi_\lambda)_{|\lambda|=j_0}$. A non-linear estimator is defined by

$$\hat{f}^X_h = \sum_{j=j_0-1}^{j_1} \sum_{|\lambda|=j} \delta^h_{\tau_{j,n}}(\hat{\beta}_\lambda)\psi_\lambda$$

with $\delta^h_{\tau_{j,n}}(x) = x\mathbb{1}_{\{|x|\geqslant\tau_{j,n}\}}$. This estimator depends on the coarse level of approximation j_0, the high-frequency cut-off j_1 and the threshold $\tau_{j,n}$ that may depend on the level of resolution j. An adaptive estimator is derived with appropriate choices of scales j_0, j_1 and threshold. One possible calibration for an adaptive estimator in ordinary smooth deconvolution is $2^{j_1} = \mathcal{O}(n^{1/2\nu+1})$ and $\delta_{j,n} = \mathcal{O}(2^{\nu j}/\sqrt{n})$ (Pensky and Vidakovic, 1999). The choice $\delta_{j,n} = \mathcal{O}(2^{\nu j}\sqrt{j/n})$ has also been considered (Fan and Koo, 2002).

3.2 Information projection to guarantee positivity

Let $j \geqslant 0$. If θ denotes a vector in $\mathbb{R}^{2^j}$, then θ_λ denotes its λ-th component. The wavelet based exponential family $\mathcal{E}_j$ at scale j is defined as the set of functions:

$$\mathcal{E}_j = \Big\{ f_{j,\theta}(\cdot) = \exp\big(\sum_{|\lambda|<j} \theta_\lambda \psi_\lambda(\cdot) - C_j(\theta)\big),\ \theta = (\theta_\lambda)_{|\lambda|<j} \in \mathbb{R}^{2^j} \Big\},$$

where $C_j(\theta) = \log \int_0^1 \exp(\sum_{|\lambda|<j} \theta_\lambda \psi_\lambda(x))dx$. Following Csiszár (1975), the density function $f_{j,\theta}$ in $\mathcal{E}_j$ that is the closest to the true density f^X in the Kullback-Leibler sense is the unique density function in $\mathcal{E}_j$ for which $\langle f_{j,\theta}, \psi_\lambda \rangle = \langle f^X, \psi_\lambda \rangle$, for all $|\lambda| < j$. It seems therefore natural to estimate the unknown density function f^X, by looking for some $\hat{\theta}_n \in \mathbb{R}^{2^j}$ such that:

$$\langle f_{j,\hat{\theta}_n}, \psi_\lambda \rangle = \sum_{\ell} \left(\frac{\psi_l^\lambda}{f_l^\epsilon} \right) \left(\frac{1}{n} \sum_{j=1}^{n} \exp(-2\pi i \ell Y_j) \right) := \hat{\alpha}_\lambda, \text{ for all } |\lambda| < j. \tag{3.1}$$

Note that the notation $\hat{\alpha}_\lambda$ is used to denote both the estimation of the scaling coefficients $\hat{c}_\lambda$ and the wavelet coefficients $\hat{\beta}_\lambda$.

The positive linear and nonlinear wavelet estimator are then defined as follows:

- The *positive linear wavelet estimator* is $f_{j_1,\hat{\theta}_n}$ such that $\langle f_{j_1,\hat{\theta}_n}, \psi_\lambda \rangle = \hat{\alpha}_\lambda$ for all $|\lambda| < j_1$
- The *positive nonlinear estimator with hard thresholding* is $f^h_{j_1,\hat{\theta}_n}$ such that $\langle f^h_{j_1,\hat{\theta}_n}, \psi_\lambda \rangle = \delta^h_{\tau_{j,n}}(\hat{\alpha}_\lambda)$ for all $|\lambda| < j_1$

The existence of these estimators is questionable. This issue is addressed in the next section and in the technical appendix. Moreover, there is no way to obtain an explicit expression for $\hat{\theta}_n$. In our simulations, we use a numerical approximation of $\hat{\theta}_n$ that is obtained via a Newton-Raphson type of algorithm.

4 Rates of convergence of the estimators

Below we study the convergence of the estimators for the Kullback-Leibler discrepancy loss between two probability density functions p and q, that is given by:

$$\Delta(p;q) = \int_0^1 p(x)\log(\frac{p(x)}{q(x)})dx,$$

where dx denotes the Lebesgue measure on $[0,1]$. Let M be some fixed constant and let $F^s_{p,q}(M)$ denote the set of density functions such that $F^s_{p,q}(M) = \{f \in L^2[0,1]$ is a p.d.f. such that for $g = \log f,\ \|g\|^q_{s,p,q} \leqslant M\}$.

4.1 Linear estimation

The following theorem is about the nonadaptive information projection estimator of the unknown density function.

Theorem 4.1 *Assume $f^X \in F^s_{2,2}(M)$ with $s > 1$, and suppose that the convolution kernel f^ϵ satisfies Assumption 1.1 (ordinary smooth convolution). Let $j(n)$ be such that $2^{-j(n)} = \mathcal{O}(n^{-1/(2s+2\nu+1)})$. Then, the information projection estimator $f_{j(n),\hat{\theta}_n}$ exists with probability tending to one as $n \to +\infty$, and is such that*

$$\mathbb{E}\Delta\left(f^X; f_{j(n),\hat{\theta}_n}\right) = \mathcal{O}\left(n^{-\frac{2s}{2s+2\nu+1}}\right).$$

In the case of ordinary smooth deconvolution, Koo and Chung (1998) have shown that $n^{-\frac{2s}{2s+2\nu+1}}$ is the fastest rate of convergence for the problem of estimating a density f such that $\log(f)$ belongs to Sobolev ball of order s which corresponds to the space $F^s_{2,2}(M)$. The above estimator $f_{j(n),\hat{\theta}_n}$ therefore converges with the optimal rate for densities in $F^s_{2,2}(M)$. However, this estimator is not adaptive since the choice of $j(n)$ depends on the unknown smoothness class of the function f^X. Moreover, the result is only suited for smooth functions (as $F^s_{2,2}(M)$ corresponds to a Sobolev space of order s) and does not attain the optimal rates when for example $g = \log(f^X)$ has singularities. In the next section, we therefore propose another estimator based on an appropriate nonlinear thresholding procedure.

4.2 Non-linear estimation

In non-linear estimation, we need to define an appropriate thresholding of the estimated coefficients $\hat{\alpha}_\lambda$. This threshold is level-dependent and takes the form $\tau_{j,n} = \eta \tau_j \sqrt{(\log n)/n}$ with $\tau_j = 2^{j\nu}$, and for some constant $\eta > 0$. The size of the exponential family used for the estimation depends on the high-frequency cut-off j_1 which is typically related to the ill-posedness ν of the inverse problem e.g. $2^{j_1} \geqslant n^{1/2\nu}$ as in Antoniadis and Bigot (2006) or $2^{j_1} = \mathcal{O}\left((\frac{n}{log(n)})^{1/(2\nu+1)} \right)$ as in Johnstone *et al*. (2004).

The following theorem gives the rate of convergence of the expected Kullback-Leibler discrepancy for the positive nonlinear estimator by hard thresholding.

Theorem 4.2 *Assume that $f^X \in F^s_{p,q}(M)$, and suppose that the convolution kernel f^ϵ satisfies Assumption 1.1 with $\nu > 0$ (ordinary smooth convolution). Suppose*

$$0 \leqslant q \leqslant \min((4\nu+2)/(2s+2\nu+1), 4\nu/(2s+2\nu-2/p+1))$$

$$1 \leqslant p \leqslant 2, \quad s \geqslant 1/p + 1/2, \quad \nu \geqslant 1/2, \tag{4.1}$$

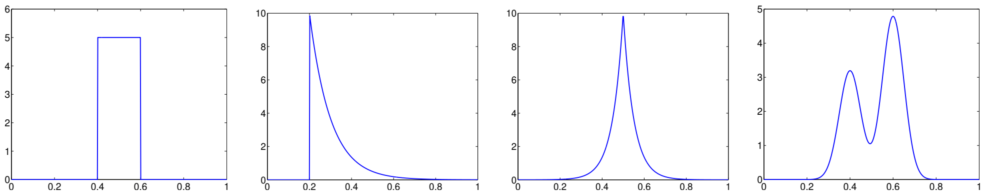
$$s \geqslant (2\nu+1)(1/p+1/2), \quad s \geqslant 1/2 + 1/(4\nu) \tag{4.2}$$

Then, the above described hard thresholding estimator exists with probability tending to one as $n \to +\infty$, and satisfies

$$\mathbb{E}\Delta(f^X; f^h_{j_1(n),\hat{\theta}_n}) = \mathcal{O}\left(\left(\frac{\log n}{n} \right)^{2s/(2s+2\nu+1)} \right),$$

provided that $2^{j_1(n)} = \mathcal{O}((n/log(n))^{1/(2\nu+1)})$.

The space $F^s_{p,q}(M)$ with $1 \leqslant p < 2$ contains piecewise smooth functions with local irregularities such as peaks or discontinuities. In the classical density estimation problem (without an additive noise), Koo and Kim (1996) have studied the optimal rate of convergence in the minimax sense for the Kullback-Leibler discrepancy over the density class $F^s_{p,q}(M)$. It is shown in Koo and Kim (1996) that $n^{-2s/(2s+1)}$ is the lowest rate of convergence if $s > 1/2$ and $p, q \geqslant 1$. However, to the best of our knowledge, studying optimal rates of convergence for ordinary smooth deconvolution has not been investigated for the Kullback-Leibler discrepancy for the class $F^s_{p,q}(M)$. We conjecture that $n^{-2s/(2s+2\nu+1)}$ is a lower bound for the problem of estimating $f^X \in F^s_{p,q}(M)$ in the case of ordinary smooth deconvolution. Hence, the above theorem



(a) (b) (c) (d)

Figure 5.1: Test densities: (a) Uniform, (b) Exponential, (c) Laplace, (d) MixtGauss (mixture of two Gaussian)

shows that our information projection estimate based on hard thresholding is adaptive and converges with a near-optimal rate. Note that in Johnstone *et al*. (2004), the case $1/p - 1/2 - \nu \leqslant s < (2\nu + 1)(1/p - 1/2)$ is also considered for which a different rate of convergence is derived. This is known as the 'Elbow' phenomenon which has been commonly observed in direct models and recently noticed by Johnstone *et al*. (2004) for deconvolution problems, but for simplicity we have not considered this case.

The conditions (4.1) and (4.2) in particular guarantee the existence with probability tending to one of the information projection estimates, see the proof of Theorem 4.2 where Lemma 5 of Barron and Sheu (1991) is also used. Moreover, note that our condition on the high-frequency cut-off yields a choice for j_1 which is similar to the one obtained by the conditions in Johnstone *et al*. (2004). The problem of determining the choice of j_1 in practice is further discussed in the next section.

5 Simulations

Given a density f^X with variance σ_X^2 and a noise density f^ϵ with variance σ_ϵ^2 we generate observations $Y_i, i = 1, \ldots, n$ from the additive model $Y_i = X_i + \epsilon_i$, where X_i (resp. ϵ_i) are independent realizations from f^X (resp. f^ϵ). Important quantities in the simulations are the sample size n and the root signal-to-noise ratio defined by $s2n := \sigma_X/\sigma_\epsilon$. For the sake of conciseness, we only present results with a Laplace measurement error, that is $f^\epsilon(x) = (\sqrt{2}\sigma_\epsilon)^{-1}\exp(-\sqrt{2}|x|/\sigma_\epsilon)$, $x \in \mathbb{R}$. The Fourier coefficients of this density are given by $f_\ell^\epsilon = (1 + 2\sigma_\epsilon^2\pi^2\ell^2)^{-1}$, $\ell = 0, \pm1, \pm2, \ldots$. This noise density corresponds to the case of ordinary smooth deconvolution with $\nu = 2$.

As for the density of interest f^X, we consider the four following functions: (1)

The Uniform distribution $f(x) = 5\mathbb{1}_{[0.4,0.6]}(x)$; the Exponential distribution $f(x) = 10e^{-10(x-0.2)}\mathbb{1}_{[0.2,+\infty[}(x)$; (3) the Laplace distribution $f(x) = 10e^{-20|x-0.5|}$ and (4) the MixtGauss distribution which is a mixture of two Gaussian variables i.e. $X \sim \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$ with $\pi_1 = 0.4, \pi_1 = 0.6$, $\mu_1 = 0.4, \mu_2 = 0.6$ and $\sigma_1 = \sigma_2 = 0.05$. The four densities f^X are displayed in Figure 5.1, where we can observe that they show various types of smoothness. The Uniform distribution is a piecewise constant function with two jumps, the Exponential distribution is a piecewise smooth function with a single jump, the Laplace density is a continuous function with a cusp at $x = 0.5$ and is thus non-differentiable at this point, whereas the MixtGauss density is infinitely differentiable.

5.1 Computation of the estimators

The computation of the wavelet deconvolution by information projection is described below. It is compared with two among the most recent estimators found in the literature : the estimator by model selection of Comte, Rozenholc and Taupin (2007) and cosine series deconvolution of Hall and Qiu (2005). Simulations use the wavelet toolbox *Wavelab* of Matlab (Buckheit *et al.*, 1995).

5.1.1 Wavelet deconvolution

The empirical Fourier coefficients $\sum_{j=1}^{n} \exp(-2\pi i \ell Y_j)/(n f_\ell^\epsilon)$ are computed for $\ell = -n/2+1, \ldots, n/2$. They are used as an input of the efficient algorithm of Kolaczyk (1994) in order to compute the Meyer wavelet coefficients of a discrete signal.

According to Theorem 4.2, the optimal cut-off is $j_1^* = (2\nu+1)^{-1}\log_2(n)$. As we will show below, this choice is too small in practice. However, this choice is crucial because a too high level of resolution might unacceptably introduce instability in the estimator (for instance when a large wavelet coefficient due to the noise at a fine scale is erroneously kept by the thresholding procedure). One objective of the simulation study is to identify a reasonable empirical range of scales j_1. We will investigate every possible values of j_1 between 3 and $\log_2(n) - 1$.

For a non linear wavelet estimator, Theorem 4.2 suggests to set the threshold $\tau_{j,n} = \eta \tau_j \sqrt{(\log n)/n}$, where η is a tuning constant and $\tau_j = 2^{j\nu}$. Based on extensive simulations, we have found that the best results were obtained with the choice $\eta = \sqrt{2}$

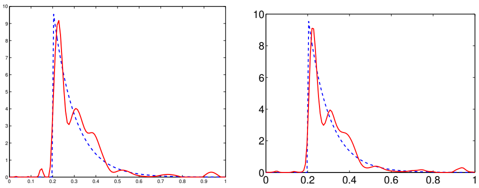
(universal thresholding). In the context of Meyer wavelet-based deconvolution in a regression setting, Johnstone *et al.* (2004) use the same type of level-dependent thresholding but the scale parameter τ_j depends on the noise distribution f^{ϵ} and on the support of the Meyer wavelet in the Fourier domain. It is given by $\tilde{\tau}_j = |C_j|^{-1} \sum_{\ell \in C_j} |f_{\ell}^{\epsilon}|^{-2}$, where C_j denotes the set of non-zero Fourier coefficients ψ_{ℓ}^{λ} at scale $|\lambda| = j$ (recall that the Meyer wavelets are band-limited) and $|C_j| = 4\pi 2^j$ is the cardinal of C_j. As it can be seen from the proof of Lemma 7.1, the above choice $\tau_j = 2^{j\nu}$ comes from the bound $\tilde{\tau}_j^2 = \mathcal{O}(2^{2j\nu})$, under the assumption of ordinary smooth deconvolution. It is not clear whether the scale parameters τ_j and $\tilde{\tau}_j$ yield similar estimators. In our simulations, we have therefore chosen to compare the results obtained from the "theoretical" scale parameter τ_j and from the "distribution dependent" scale parameter $\tilde{\tau}_j$.

Once we have computed the coefficients $\delta_{\tau_{j,n}}^{h}(\hat{\alpha}_{\lambda})$ with hard thresholding for all $|\lambda| < j_1$, it remains to compute the empirical version of the information projection estimate $f_{j_1,\hat{\theta}_n}^{h}$. In this step we use a Newton-Raphson type algorithm as described in Antoniadis and Bigot (2006).

As it was suggested by a referee, one may wonder what are the advantages of the information projection step over a simple truncation to its positive part of the unconstrained estimator $\hat{f}_h^X$ obtained by simple thresholding of the $\hat{\alpha}_{\lambda}$'s. In Figure 5.2, we display an example of the estimation of the Exponential density by $\hat{f}_h^X$ and $\hat{f}_{j_1,\hat{\theta}_n}^{h}$. The projection step yields significant improvements as it removes some of the oscillating parts of the unconstrained estimator $\hat{f}_h^X$, and it gives a smoother estimation in the regions where the true density is close to zero. As the mass of $\hat{f}_{j_1,\hat{\theta}_n}^{h}$ is equal to one, it also gives a better estimation of the peak of the Exponential density.

5.1.2 Density deconvolution via model selection

The adaptive density deconvolution estimator of Comte *et al.* (2007) is based on penalized contrast minimization over a collection of model S_m, $m \in \mathcal{M}_n = \{1, \ldots, m_n\}$ where S_m is the space of square integrable functions with Fourier transform supported included in $[-l_m, l_m]$ with $l_m = m\Delta$, $\Delta > 0$. It is therefore a band-limited function $\hat{f} \in S_{\hat{m}}$ where $\hat{m}$ is the model selected by minimization of an appropriate penalized criteria based on the Y_i's and the probability distribution of ϵ, see Comte *et al.* (2007). Hence, this estimator can be viewed as a kind of adaptive linear wavelet



(a) (b)

Figure 5.2: The usefulness of information projection is illustrated here on the estimation of the Exponential density ($n = 128$, $s2n = 10$): (a) unconstrained wavelet estimator $\hat{f}_h^X$ truncated to its positive part, (b) corresponding positive estimator $\hat{f}^h_{j_1,\hat{\theta}_n}$ via wavelet thresholding and information projection.

estimator with a Shannon wavelet basis which is also a band-limited function like the Meyer wavelet but less localized in the time domain.

5.1.3 Cosine series deconvolution

We also compare the results with the recently introduced estimator of Hall and Qiu (2005). The estimator is based on the cosine-series expansion $\hat{f}(x) = 1 + \sum_{j=1}^{m} 2\hat{a}_j \cos(j\pi x)$ where $\hat{a}_j$ is an estimator of the coefficient $a_j = \int_0^1 f(x)\cos(j\pi x)dx$ and $m \geqslant 1$ is an integer defining a high frequency cut-off. In our simulations, the error follows a Laplace distribution, which is symmetric about its mean 0. A simple estimator of the cosine a_j is therefore given by $\hat{a}_j = \hat{b}_j/\alpha_j \delta_{\tau_n}(|\hat{b}_j|)$, where $\alpha_j = \mathbb{E}(\cos(j\pi\epsilon_1))$, and $\delta_{\tau_n}(|\hat{b}_j|) = \mathbb{1}_{|\hat{b}_j|>\tau_n}$ is a simple hard-thresholding rule with $\tau_n = C\sqrt{\log(n)/(2n)}$ and C is a tuning constant. Based on Hall and Qiu (2005), we set $m = n$ and $C = 2$.

5.2 Results of the simulations

Figure 5.3 shows typical estimates of f^X for $n = 512$ and $s2n = 10$ with all methods. Note that for the sake a better visual quality, we only plot the positive part of the estimators. Our wavelet estimator is by construction a probability density function and, with that respect, is more satisfactory than the two competitors that may take negative values. When f^X is not smooth (i.e. for Uniform, Exponential and Laplace distribution) the reconstruction of the singularities (discontinuities and cusp) of the signals is much better with our wavelet estimator. For the smooth density MixtGauss, the model selection estimator performs slightly better than the two other methods.

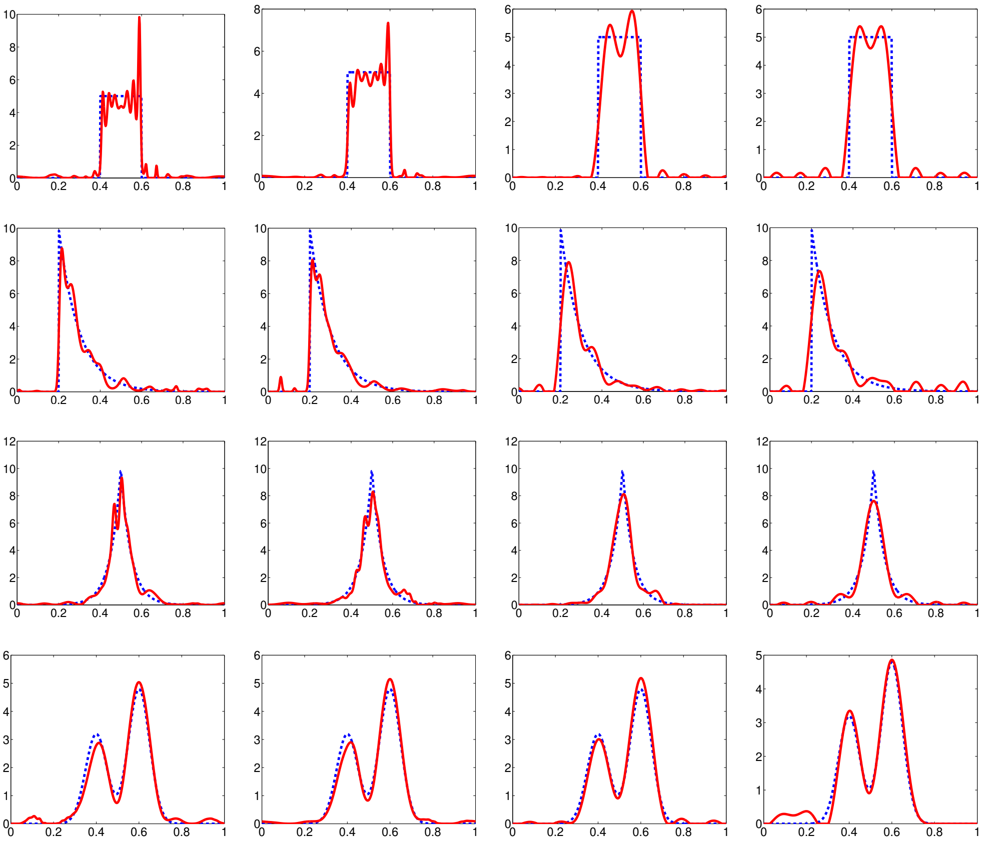


Figure 5.3: Typical reconstructions from a single simulation of four contaminated densities: Uniform, Exponential, Laplace, and MixtGauss. Estimators are: non-TI wavelet thresholding (1st column), TI wavelet thresholding (2nd column), model selection (3rd column) and cosine series (4th column). The scale j_1 considered in the wavelet estimators depend on the true density: $j_1 = 3$ for the Laplace density, and $j_1 = 4$ for the three other densities. In all figures, the dotted lines is the true density and the solid lines is the estimator ($n = 512$ and $s2n = 10$).

By inspecting the first column in Figure 5.3 we see that the wavelet estimator is affected by pseudo-Gibbs phenomena. A possible remedy to this defect is to use a translation invariant (TI) procedure such as the one suggested by Donoho and Raimondo (2004) for Meyer wavelet-based deconvolution in a regression setting. In the second column of Figure 5.3 we display the TI version of the wavelet estimators plotted in the first column. Observe that TI estimators remarkably exhibit very small oscillations while preserving a good reconstruction of the singularities of the non-smooth densities. Note that smoother estimates can also be obtained by using a soft thresholding rule.

We also give the result of some Monte Carlo exercises. Here, we consider the four test densities for various sample sizes ($n = 128, 512$) and various levels of noise ($s2n = 100, 10, 3$). Note that $s2n = 100$ corresponds to a noise with a very small variance, and therefore that model is very close to the direct density estimation problem with uncontaminated data. For each combination of these factors, we simulate 100 independent samples of size n and compute the integrated square error (ISE), $\sum_{i=1}^{m}(\hat{f}_n(t_i) - f(t_i))^2/m$, where $t_i = i/m$, $i = 0, \ldots, m-1$. The ISE is computed in Table 1 for $m = n$ (this choice is not critical and the conclusions below remain for $m \neq n$). Note also that our choice of computing the ISE instead of the Kullback Leibler divergence was guided by the fact that the competing methods do not always provide a stricly positive estimator of the unknown density.

Table 1 shows the mean and the variance of the integrated square error (ISE) for various estimators. For wavelet deconvolution, we report the results of the two thresholdings: *wavtheo* is based on τ_j whereas *wavemp* is constructed with $\tilde{\tau}_j$. We also indicate the level j_1 leading to the smallest empirical mean of ISE's over the 100 simulations. As it can be observed from Table 1, the new estimator outperforms the competitors for all type of non-smooth densities f^X. It confirms the superiority of wavelet-based positive estimators over those based on Fourier decompositions for the reconstruction of signals with local singularities. The wavelet thresholding with the scale parameter $\tau_j = 2^{2j\nu}$ gives generally better results. When the true density is a mixture of Gaussian random variables, the wavelet approach is better for $n = 128$ while the model selection procedure is slightly better than wavelet thresholding for $n = 512$. Note that the fine level j_1 that gives the best results is generally quite low (although it is larger than the theoretically optimal level j_1^*). For some combinations of the parameters of the Monte Carlo simulation, the choices $j_1 = 3, 4$ yield to the best results. This observation is consistent with the condition of Theorem 4.2 that suggests a smaller j_1 for ill-posed inverse problems than in the direct case. It also confirms that introducing higher level of resolution does not necessarily improve the quality of the estimator.

	Wavemp		Wavtheo		Modsel	Cos
Uniform distribution:						
$n = 128$, $s2n = 100$	**0.4** (0.14)	$j_1 = 3$	0.42 (0.14)	$j_1 = 3$	0.87 (0.11)	0.56 (0.04)
$n = 128$, $s2n = 10$	**0.41** (0.14)	$j_1 = 3$	0.412 (0.14)	$j_1 = 3$	0.88 (0.11)	0.56 (0.05)
$n = 128$, $s2n = 3$	**0.48** (0.16)	$j_1 = 3$	0.522 (0.23)	$j_1 = 3$	0.91 (0.12)	0.58 (0.03)
$n = 512$, $s2n = 100$	0.26 (0.04)	$j_1 = 4$	**0.22** (0.06)	$j_1 = 3$	0.28 (0.02)	0.31 (0.05)
$n = 512$, $s2n = 10$	0.26 (0.04)	$j_1 = 4$	**0.212** (0.05)	$j_1 = 3$	0.28 (0.02)	0.31 (0.04)
$n = 512$, $s2n = 3$	0.3 (0.08)	$j_1 = 3$	**0.282** (0.19)	$j_1 = 3$	0.3 (0.03)	0.43 (0.07)
Exponential distribution:						
$n = 128$, $s2n = 100$	**0.44** (0.17)	$j_1 = 4$	0.452 (0.1)	$j_1 = 3$	0.86 (0.46)	1.2 (0.17)
$n = 128$, $s2n = 10$	**0.49** (0.13)	$j_1 = 3$	0.492 (0.13)	$j_1 = 3$	1.03 (0.57)	1.28 (0.21)
$n = 128$, $s2n = 3$	1.2 (0.37)	$j_1 = 3$	**0.952** (0.61)	$j_1 = 3$	1.53 (0.51)	1.55 (0.21)
$n = 512$, $s2n = 100$	**0.27** (0.07)	$j_1 = 4$	0.282 (0.07)	$j_1 = 4$	0.51 (0.03)	0.64 (0.07)
$n = 512$, $s2n = 10$	0.38 (0.1)	$j_1 = 4$	**0.312** (0.08)	$j_1 = 4$	0.53 (0.03)	0.7 (0.07)
$n = 512$, $s2n = 3$	**0.59** (0.25)	$j_1 = 3$	0.792 (0.75)	$j_1 = 3$	0.7 (0.04)	0.95 (0.07)
Laplace distribution:						
$n = 128$, $s2n = 100$	0.24 (0.13)	$j_1 = 3$	**0.222** (0.09)	$j_1 = 5$	1.16 (0.27)	0.38 (0.12)
$n = 128$, $s2n = 10$	**0.23** (0.13)	$j_1 = 3$	0.242 (0.13)	$j_1 = 5$	1.15 (0.24)	0.38 (0.13)
$n = 128$, $s2n = 3$	0.43 (0.2)	$j_1 = 3$	**0.362** (0.23)	$j_1 = 3$	1.23 (0.23)	0.49 (0.15)
$n = 512$, $s2n = 100$	0.07 (0.04)	$j_1 = 3$	**0.062** (0.03)	$j_1 = 3$	0.08 (0.04)	0.15 (0.04)
$n = 512$, $s2n = 10$	0.08 (0.04)	$j_1 = 3$	**0.062** (0.03)	$j_1 = 3$	0.08 (0.03)	0.16 (0.04)
$n = 512$, $s2n = 3$	0.2 (0.08)	$j_1 = 3$	**0.082** (0.05)	$j_1 = 3$	0.15 (0.03)	0.25 (0.05)
MixtGauss distribution:						
$n = 128$, $s2n = 100$	0.19 (0.1)	$j_1 = 3$	**0.132** (0.07)	$j_1 = 5$	0.95 (0.05)	0.4 (0.13)
$n = 128$, $s2n = 10$	0.24 (0.1)	$j_1 = 3$	**0.162** (0.08)	$j_1 = 5$	0.95 (0.07)	0.42 (0.14)
$n = 128$, $s2n = 3$	**0.41** (0.21)	$j_1 = 3$	0.542 (0.31)	$j_1 = 3$	0.98 (0.07)	1.07 (3.65)
$n = 512$, $s2n = 100$	0.05 (0.02)	$j_1 = 3$	0.042 (0.02)	$j_1 = 3$	**0.03** (0.01)	0.12 (0.05)
$n = 512$, $s2n = 10$	0.07 (0.04)	$j_1 = 4$	0.042 (0.02)	$j_1 = 3$	**0.03** (0.01)	0.14 (0.05)
$n = 512$, $s2n = 3$	0.24 (0.2)	$j_1 = 4$	0.152 (0.15)	$j_1 = 3$	**0.07** (0.02)	0.22 (0.1)

Table 1: Empirical mean and standard deviation (in brackets) of the ISE over $M = 100$ repetitions for each method and some combination of the factors n and $s2n$. In the wavelet-based methods, only the level j_1 leading to the smallest empirical mean is reported. The smallest ISE over lines is bolded.

6 Conclusion and perspectives

Compared to the some recent deconvolution methods, the above results demonstrate the significant improvement given by the nonlinear wavelet thresholding estimator by information projection: The estimator is showed to have an optimal rate of convergence over a reasonable class of functions and a thorough empirical study proved the satisfactory behaviour of the estimator on finite sample.

The empirical study also showed that the theoretical optimal level $j_1^* = (2\nu + 1)^{-1}\log_2(n)$ is usually too small for the practice. This phenomenon is not surprising and has also been noticed e.g. in Johnstone *et al.* (2004). Future research could be devoted to a tighter, non asymptotic control for the risk of estimation (e.g. via oracle-type inequalities). This would be most useful in order to develop an automatic, data-driven selection of j_1. Similarly, a specific work on $\tau_{j,k}$ is also needed. A possible way to address the problem is to extend to our setting the corresponding work provided by (Juditsky and Lambert-Lacroix, 2004) in the standard regression model.

7 Appendix

We start by a technical lemma used in the proof of the main results. In what follows, C denotes a generic constant whose value may change from line to line.

Lemma 7.1 *Assume that the Fourier coefficients of f^Y are such that $|f_l^Y| \leqslant C|l|^{-u}$ with $u > 1$. Then,*

$$\mathbb{E}(\hat{\alpha}_{n,\lambda} - \alpha_\lambda)^2 \leqslant \frac{C}{n} 2^{2|\lambda|\nu}$$

where $\hat{\alpha}_{n,\lambda} = \sum_l \left(\frac{\psi_l^\lambda}{f_l^\epsilon}\right)\left(\frac{1}{n}\sum_{j=1}^n e^{-2\pi i l Y_j}\right)$ and $\alpha_\lambda = \sum_l \frac{\psi_l^\lambda}{f_l^\epsilon} f_l^Y$.

PROOF: For $|\lambda| = j$, let $C_j = \{\ell \ : \ \psi_\ell^\lambda \neq 0\}$. Since the Meyer wavelets are band-limited, $C_j = \{\ell \ : \ 2^j \leqslant |l| \leqslant 2^{j+r}\}$ for some fixed $r > 0$. To simplify the notation, we shall assume that $C_j = \{\ell \ : \ 2^j \leqslant l \leqslant 2^{j+r}\}$ noticing that all the bounds below also hold for negative values of ℓ. Then, using Assumption 1.1, we use $|\psi_\ell^\lambda| \leqslant C2^{-|\lambda|/2}$ and the independence of the Y_i's in order to write

$$\mathbb{E}(\hat{\alpha}_{n,\lambda} - \alpha_\lambda)^2 \leqslant \frac{C}{n} 2^{2|\lambda|\nu} 2^{-|\lambda|} \sum_{\ell,\ell'=2^{|\lambda|}}^{2^{|\lambda|+r}} \mathbb{E} e^{-2\pi i(\ell-\ell')Y_1} \leqslant \frac{C}{n} 2^{2|\lambda|\nu} + \frac{C}{n} 2^{2|\lambda|\nu} 2^{-|\lambda|} \sum_{\ell \neq \ell'} f_{\ell-\ell'}^Y$$

As $|f_\ell^Y| \leqslant C|\ell|^{-u}$ with $u > 1$, the double sum $\sum_{\ell \neq \ell'} f_{\ell-\ell'}^Y$ in the equation above is bounded which yields the result. □

Proof of the main theorems. The proof of the two main theorems is based on a decomposition of the relative entropy between the true and the estimated density function into the sum of two terms which correspond to approximation error and estimation error (bias and variance in a familiar mean squared error analysis). This decomposition is given by

$$\Delta(f^X; f_{j,\hat{\theta}_n}) = \Delta(f^X; f_{j,\theta_j^*}) + \Delta(f_{j,\theta_j^*}; f_{j,\hat{\theta}_n}) \tag{7.1}$$

where f_{j,θ_j^*} denotes the closest function of $\mathcal{E}_j$ to the true density f^X for the Kullback-Leibler divergence. This identity comes from the Pythagorean Theorem derived in Csiszár (1975). It allows in particular to write the risk $\mathbb{E}\Delta(f^X; f_{j(n),\hat{\theta}_n})$ as the sum of an approximation error term $\Delta(f^X; f_{j(n),\theta_{j(n)}^*})$ and an estimation error term $\mathbb{E}\Delta(f_{j(n),\theta_{j(n)}^*}; f_{j(n),\hat{\theta}_n})$.

The control of the approximation error term is similar for the linear and the nonlinear estimators. Below, we only sketch the proof of the existence and uniqueness of f_{j,θ_j^*} as this follows from the arguments in Antoniadis and Bigot (2006) and by applying Barron and Sheu (1991, Lemma 5). To do so, note that the technical lemmas in Appendix A of Antoniadis and Bigot (2006) need to be adapted to the case of Meyer wavelets.

The control of the estimation error term differs for the linear or the nonlinear estimators. In the linear case, it simply relates to the control of the risk $\mathbb{E}\|\hat{\alpha}_n - \alpha_0\|_2^2$ which is given by Lemma 7.1. In the nonlinear situation, we use some classical moment bounds (Rosenthal (1972)) and Bernstein's inequality to control the difference between the estimated wavelet coefficients and their true values, together with the maxiset theorem of Johnstone *et al.* (2004).

For the periodised Meyer wavelet basis and under the conditions of Theorem 4.2 for $s, p, q, \nu, \tau_{j,n}, j_1$, this maxiset theorem says that an estimator of the form $\hat{f}_h = \sum_{j=j_0-1}^{j_1} \sum_{|\lambda|=j} \delta^h_{\tau_{j,n}}(\hat{\beta}_\lambda)\psi_\lambda$ satisfies the asymptotic rate of converge $\mathbb{E}\|\hat{f}_h - f\|^2_{L^2([0,1])} \leqslant C(\frac{\log n}{n})^{2s/(2s+2\nu+1)}$ provided that $2^{j_1(n)} = \mathcal{O}\left((\frac{n}{log(n)})^{1/(2\nu+1)}\right)$, and if for η large enough, there exists two constant C_1 and C_2 such that for all $n \in \mathbb{N}^*$ and $|\lambda| = j$

$$\mathbb{E}|\hat{\beta}_\lambda - \beta_\lambda|^4 \leqslant C_1 \frac{\tau_j^4}{n^2}, \tag{7.2}$$

$$\mathbb{P}\left(|\hat{\beta}_\lambda - \beta_\lambda| \geqslant \eta\tau_j\sqrt{(\log n)/n}\right) \leqslant C_2(\frac{\log n}{n})^2, \tag{7.3}$$

where the β_λ's are the wavelet coefficients of f (see Johnstone *et al.* (2004) for further details).

Proof of Theorem 4.1. We first consider the control of the approximation term. By arguing as in Barron and Sheu (1991) and Antoniadis and Bigot (2006), and under the assumptions of Theorem 4.1 , one can prove that for n sufficiently large, there exists some $\theta^*_{j(n)}$ such that $\langle f^X, \psi_\lambda \rangle = \langle f_{j(n),\theta^*_{j(n)}}, \psi_\lambda \rangle$ for all $|\lambda| < j(n)$, which satisfies for $2^{-j(n)} = \mathcal{O}(n^{-1/(2s+2\nu+1)})$

$$\Delta(f^X; f_{j(n),\theta^*_{j(n)}}) = \mathcal{O}\left(2^{-2j(n)s}\right) = \mathcal{O}\left(n^{-2s/(2s+2\nu+1)}\right). \tag{7.4}$$

We now turn to the estimation error term. For all $|\lambda| < j(n)$, define $\alpha_{0,\lambda} = \langle f^X, \psi_\lambda \rangle = \langle f_{j,\theta^*_j}, \psi_\lambda \rangle$ and let $\hat{\alpha}_{n,\lambda} = \sum_l (\psi_l^\lambda / f_l^\epsilon) \sum_{j=1}^n \exp(-2\pi i l Y_j)/n$. To prove the existence of a vector $\hat{\theta}_n \in \mathbb{R}^{2^{j(n)}}$ such that $\langle f_{j,\hat{\theta}_n}, \psi_\lambda \rangle = \hat{\alpha}_{n,\lambda}$, for all $|\lambda| < j(n)$, we need to control the term $\|\hat{\alpha}_n - \alpha_0\|_2^2 = \sum_{|\lambda|<j(n)} (\hat{\alpha}_{n,\lambda} - \alpha_{0,\lambda})^2$ and then to apply Barron and Sheu (1991, Lemma 5). Given our assumption on f^X and f^ϵ we have that $|f_l^Y| \leqslant C|l|^{-(s+\nu)}$ with $s + \nu > 1$, and we can therefore apply Lemma 7.1 to obtain that $\mathbb{E}\|\hat{\alpha}_n - \alpha_0\|_2^2 \leqslant C 2^{j(n)(2\nu+1)}/n$ Then, under the assumptions of Theorem 4.1, and arguing as in Antoniadis and Bigot (2006) and by applying Barron and Sheu (1991, Lemma 5), we have that for n sufficiently large, $\hat{\theta}_n$ exists and is such that

$$\mathbb{E}\left(\Delta(f_{j(n),\theta^*_{j(n)}}; f_{j(n),\hat{\theta}_n})\right) = \mathcal{O}\left(n 2^{j(n)(2\nu+1)}\right) = \mathcal{O}\left(n^{-2s/(2s+2\nu+1)}\right), \tag{7.5}$$

for $2^{-j(n)} = \mathcal{O} n^{-1/(2s+2\nu+1)})$. The result of the theorem now follows from the control of the approximation and estimation error terms, using the identity (7.1). □

Proof of Theorem 4.2. By proceeding as in Antoniadis and Bigot (2006), one can show that for n sufficiently large, there exists some $\theta^*_{j_1(n)}$ such that for $1 \leqslant p \leqslant 2$ and $s > 1/2 + 1/p$, it holds $\Delta(f^X; f_{j_1(n),\theta^*_{j_1(n)}}) = \mathcal{O}(2^{-2j_1(n)(s-1/2-1/p)})$, where we have used the notations from the proof of Theorem 4.1. Then, since $2^{j_1(n)} = \mathcal{O}(\{\log(n)/n\}^{1/(2\nu+1)})$, we can write $\Delta(f^X; f_{j_1(n),\theta^*_{j_1(n)}}) = \mathcal{O}(\{\log(n)/n\}^{2(s-1/2-1/p)/(2\nu+1)})$. Since $s \geqslant (2\nu + 1)(1/p + 1/2)$ by assumption, we therefore obtain $s - 1/2 - 1/p \geqslant 2s\nu/(2\nu+1)$ and the condition $s \geqslant 1/2 + 1/4\nu$ finally implies that $2s\nu/(2\nu+1)^2 \geqslant 2s/(2s+2\nu+1)$ which yields the near-optimal order of convergence for the approximation term $\Delta(f^X; f_{j_1(n),\theta^*_{j_1(n)}}) = \mathcal{O}(\{\log(n)/n\}^{2s/(2s+2\nu+1)})$.

We can now consider the estimation error term. Define $\hat{\alpha}_{n,\lambda}$ and α_λ as in the proof of Theorem 4.1. Define $\mathbb{E}\|\delta^h_{\tau_{j,n}}(\hat{\alpha}_n) - \alpha_0\|_2^2 = \sum_{j_0-1\leqslant|\lambda|<j_1(n)} \mathbb{E}(\delta^h_{\tau_{j,n}}(\hat{\alpha}_{n,\lambda}) - \alpha_\lambda)^2$ with $\tau_{j,n} = \eta \tau_j \sqrt{(\log n)/n}$ and $\tau_j = 2^{j\nu}$. In order to control the above sum, we use the

maxiset theorem of Johnstone *et al.* (2004). Given our conditions imposed on p, q, s, ν, j_1 and $\tau_{j,n}$ it remains to check (7.2) and (7.3) with $\hat{\beta}_\lambda = \hat{\alpha}_{n,\lambda}$ and $\beta_\lambda = \alpha_\lambda$.

Before we recall a useful result for moment bounds of iid variables (Rosenthal, 1972): If $Z_1, \ldots, Z_n$ are iid random variables such that $\mathbb{E}Z_j = 0$, $\mathbb{E}Z_j^2 \leqslant \sigma^2$, then if $m \geqslant 2$, there exists a positive c_m such that $\mathbb{E}|\sum_{j=1}^n Z_j/n|^m \leqslant c_m(\sigma^m/n^{m/2} + \mathbb{E}|Z_1|^m/n^{m-1})$.

Recall that $\hat{\alpha}_{n,\lambda} - \alpha_\lambda = n^{-1}\sum_{j=1}^n(\sum_l(e^{-2\pi i l Y_j} - f_l^Y)\psi_l^\lambda/f_l^\epsilon)$. For $|\lambda| = j$, let $C_j = \{l : \psi_l^\lambda \neq 0\}$. Since the Meyer wavelets are band-limited, $C_j = \{l : 2^j \leqslant |l| \leqslant 2^{j+r}\}$ for some fixed $r > 0$. To simplify the notation, we shall assume that $C_j = \{l : 2^j \leqslant l \leqslant 2^{j+r}\}$ noticing that all the bounds below also hold for negative values of ℓ. Hence, we have that $\hat{\alpha}_{n,\lambda} - \alpha_\lambda = \frac{1}{n}\sum_{j=1}^n Z_j$, where the Z_j's are iid variables such that $Z_j = \sum_{l=2^{|\lambda|}}^{2^{|\lambda|+r}} \frac{\psi_l^\lambda}{f_l^\epsilon}\left(e^{-2\pi i l Y_j} - f_l^Y\right)$.

First notice that $\mathbb{E}Z_j = 0$. In order to apply Rosenthal's inequality, it remains to derive a bound for $\mathbb{E}|Z_j|^2$ and $\mathbb{E}|Z_j|^4$. Denote by g_λ the function $g_\lambda(x) = \sum_{l=2^{|\lambda|}}^{2^{|\lambda|+r}}(\psi_l^\lambda/f_l^\epsilon)\exp(-2\pi i l x)$, and observe that the inequality

$$\mathbb{E}|Z_j|^2 = \mathbb{E}|g_\lambda(Y_j) - \alpha_\lambda|^2 \leqslant C\left(\int |g_\lambda(y)|^2 dy + |\alpha_\lambda|^2\right) \tag{7.6}$$

holds. Then by Parseval equality, one has that

$$\int |g_\lambda(y)|^2 dy = \sum_{l=2^{|\lambda|}}^{2^{|\lambda|+r}} |\psi_l^\lambda/f_l^\epsilon|^2 \leqslant C2^{|\lambda|2\nu}, \tag{7.7}$$

where we have used the fact that $\alpha_\lambda^2 = (\int_0^1 f(x)\psi_\lambda(x)dx)^2 \leqslant \int_0^1 f(x)^2 dx \int_0^1 \psi_\lambda(x)^2 dx$ is thus bounded by some constant C for any λ, that $\sum_{l=2^{|\lambda|}}^{2^{|\lambda|+r}}\left|\psi_l^\lambda\right|^2 = 1$, and Assumption 1.1. Inserting (7.6) into (7.7) yields $\mathbb{E}|Z_j|^2 = \mathbb{E}|g_\lambda(Y_j) - \alpha_\lambda|^2 \leqslant C2^{|\lambda|2\nu}$. Using the inequality $(a+b)^4 \leqslant 8(a^4 + b^4)$ that is valid for any real a, b, we get

$$\mathbb{E}|Z_j|^4 = \mathbb{E}|g_\lambda(Y_j) - \alpha_\lambda|^4 \leqslant C\left(\int |g_\lambda(y)|^4 dy + |\alpha_\lambda|^4\right). \tag{7.8}$$

Then, observe that $\int |g_\lambda(y)|^4 dy \leqslant |g_\lambda|_\infty^2 \int |g_\lambda(y)|^2 dy \leqslant C2^{|\lambda|(2\nu+1)}2^{|\lambda|2\nu}$, where we have used $|g_\lambda|_\infty \leqslant \sum_{l=2^{|\lambda|}}^{2^{|\lambda|+r}}|\psi_l^\lambda/f_l^\epsilon| \leqslant C2^{|\lambda|\nu}2^{|\lambda|/2}$ which comes from the fact $|\psi_l^\lambda| \leqslant C2^{-|\lambda|/2}$ and from Assumption 1.1. With (7.8) and using the fact that $|\alpha_\lambda|^4 \leqslant C$ finally leads to $\mathbb{E}|Z_j|^4 \leqslant C2^{|\lambda|(4\nu+1)}$.

Now, if we apply Rosenthal's inequality with $m = 4$ and for $|\lambda| = j$ we obtain $\mathbb{E}\left|\hat{\alpha}_{n,\lambda} - \alpha_\lambda\right|^4 \leqslant C(2^{4j\nu}/n^2 + 2^{j(4\nu+1)}/n^3)$. As the thresholding parameter is such that

$\tau_j^2 = 2^{2j\nu}$ and given that for $j \leqslant j_1(n)$, one has that $\frac{2^j}{n} \leqslant C$, we finally obtain that $2^{4j\nu}/n^2 \leqslant C\tau_j^4/n^2$ and $2^{j(4\nu+1)}/n^3 \leqslant C\tau_j^4/n$. It leads to $\mathbb{E}|\hat{\alpha}_{n,\lambda} - \alpha_\lambda|^4 \leqslant C\tau_j^4/n^2$, holds true. This development proves that $\hat{\alpha}_{n,\lambda} - \alpha_\lambda$ satisfies the condition (7.2).

Now, recall the standard Bernstein's inequality: let $Z_1, \ldots, Z_n$ be i.i.d. random variables with $\mathbb{E}Z_j = 0$, $\mathbb{E}Z_j^2 \leqslant \sigma^2$, $|Z_j| \leqslant \|Z\|_\infty < +\infty$, then for any $\lambda > 0$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n Z_j\right| > \lambda\right) \leqslant 2\exp\left(-\frac{n\lambda^2}{2(\sigma^2 + \|Z\|_\infty \lambda/3)}\right).$$

Now, let us apply Bernstein's inequality with the Z_j's as defined previously. From (7), we have that $\mathbb{E}Z_j^2 \leqslant C2^{|\lambda|2\nu}$ and arguing as previously, one has that $|Z_j| \leqslant C2^{|\lambda|(\nu+1/2)}$. Therefore, the following bound holds for $|\lambda| = j$ (for some constant C_1 and C_2)

$$\begin{aligned}\mathbb{P}\left(|\hat{\alpha}_{n,\lambda} - \alpha_\lambda| > \eta\tau_j\sqrt{(\log n)/n}\right) &\leqslant 2\exp\left(-\frac{\eta^2\log(n)}{2(C_1 + C_2 2^{j/2}\eta(\log n/n)^{1/2})}\right), \\ &\leqslant 2\exp\left(-C\eta^2\log(n)\right).\end{aligned}$$

Hence, for η large enough one has that for all $n \geqslant 1$ $\mathbb{P}(|\hat{\alpha}_{n,\lambda} - \alpha_\lambda| > \eta\tau_j\sqrt{\log(n)/n}) \leqslant C\{\log(n)/n\}^2$, which proves that $\hat{\alpha}_{n,\lambda} - \alpha_\lambda$ satisfies the condition (7.3).

Hence, from the maxiset theorem in Johnstone *et al.* (2004), we finally derive the following upper bound : $\mathbb{E}\|\delta^h_{\tau_{j,n}}(\hat{\alpha}_n) - \alpha_0\|_2^2 = \mathcal{O}(\{\log(n)/n\}^{2s/(2s+2\nu+1)})$.

In order to prove the existence of the projection estimate $f^h_{j_1(n),\hat{\theta}_n}$ we proceed as in Antoniadis and Bigot (2006). Under the assumptions of Theorem 4.1 and by applying Barron and Sheu (1991, Lemma 5), one can show that for n sufficiently large, $f^h_{j(n),\hat{\theta}_n}$ exists and is such that $\mathbb{E}(\Delta(f_{j(n),\theta^*_{j(n)}}; f^h_{j(n),\hat{\theta}_n})) = \mathcal{O}(\{\log(n)/n\}^{2s/(2s+2\nu+1)})$. The result of the theorem now follows from the control of the approximation and estimation error terms, using the identity (7.1). □

References

Antoniadis, A. and Bigot, J. (2006). Poisson inverse problems. *Ann. Statist.*, *34*, 2132-2158.

Barron, A. R. and Sheu, C. H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.*, *19*, 1347–1369.

Buckheit, J., Chen, S., Donoho, D. and Johnstone, I. (1995). *Wavelab reference manual* (Tech. Rep.). Department of Statistics, Stanford University.

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

Comte, F., Rozenholc, Y. and Taupin, M.-L. (2006). Penalized contrast estimator for density deconvolution. *Canad. J. Statist.*, *34*, XXX.

Comte, F., Rozenholc, Y. and Taupin, M.-L. (2007). Finite sample penalization in adaptive density deconvolution. *J. Stat. Comput. Simul.*, *7*, 977–1000.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, *3*, 146–158.

De Canditiis, D. and Pensky, M. (2006). Simultaneous wavelet deconvolution in periodic setting. *Scand. J. Statist.*, *33*, 293–306.

Donoho, D. L. and Raimondo, M. (2004). Translation invariant deconvolution in a periodic setting. *Int. J. Wavelets Multiresolut. Inf. Process.*, *4*, 415–431.

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

Fan, J. and Koo, J.-Y. (2002). Wavelet deconvolution. *IEEE Trans. Inform. Theory*, *48*, 734–747.

Hall, P. and Qiu, P. (2005). Discrete-transform approach to deconvolution problems. *Biometrika*, *92*, 135–148.

Johannes, J., Van Bellegem, S. and Vanhems, A. (2007). A unified approach to solve ill-posed inverse problems in econometrics. (`www.stat.ucl.ac.be/ISpub`)

Johnstone, I., Kerkyacharian, G., Picard, D. and Raimondo, M. (2004). Wavelet deconvolution in a periodic setting. *J. Roy. Statist. Soc. Ser. B*, *66*, 547–573.

Juditsky, A. and Lambert-Lacroix, S. (2004). On minimax density estimation on R. *Bernoulli*, *10*, 187–220.

Kolaczyk, E. (1994). *Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data*. Ph.d. thesis, Stanford University.

Koo, J.-Y. (1999). Logspline deconvolution in Besov space. *Scand. J. Statist.*, *26*, 73–86.

Koo, J.-Y. and Chung, H.-Y. (1998). Log-density estimation in linear inverse problems. *Ann. Statist.*, *26*, 335–362.

Koo, J.-Y. and Kim, W.-C. (1996). Wavelet density estimation by approximation of log-densities. *Statistics and Probability Letters*, *26*(3), 271-278.

Mallat, S. (1998). *A wavelet tour of signal processing*. New York: Academic Press.

Meyer, Y. (1992). *Wavelets and operators*. Cambridge: Cambridge University Press.

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

Rosenthal, H. P. (1972). On the span in L^p of sequences of independent random variables. II. 149–167.