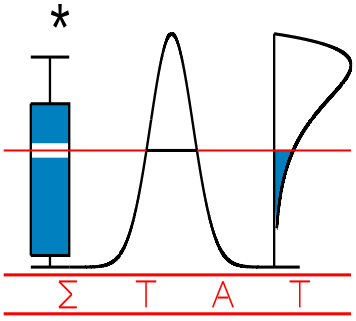
T E C H N I C A L

R E P O R T

0630

LOG-DENSITY DECONVOLUTION BY WAVELET THRESHOLDING

BIGOT J., and S. VAN BELLEGEM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

LOG-DENSITY DECONVOLUTION BY WAVELET THRESHOLDING

Jérémie Bigot & Sébastien Van Bellegem

July 3, 2006

Abstract

This paper proposes a new wavelet-based method for deconvolving a density. The estimator combines the ideas of nonlinear wavelet thresholding with 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 theoretical 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 practical 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, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, F-31062 Toulouse Cedex 9, France, `Jeremie.Bigot@math.ups-tlse.fr`

SÉBASTIEN VAN BELLEGEM, Institut de statistique and CORE, Université catholique de Louvain, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium, `vanbellegem@stat.ucl.ac.be`

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 and suggestions.

1 Introduction

Density deconvolution is a widely studied statistical problem that is encountered in many applied situations. This problem 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

$$Y_i = X_i + \epsilon_i, \quad i = 1, \ldots, n$$

where X_i are iid variables with unknown density f^X, and the added variables ϵ_i model the contamination by some noise. 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.

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

In data analysis, density estimation from noisy sample plays a fundamental role. Applications can be found in communication theory (e.g. Masry, 2003), experimental physics (e.g. Kosarev *et al.*, 2003) or econometrics (e.g. Postel-Vinay and Robin, 2002) to name but a few. 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, standard methods recently studied in the statistical literature include estimation by model selection (e.g. Comte, Rozenholc and Taupin, 2006b), wavelet thresholding (e.g. Fan and Koo, 2002), kernel smoothing (e.g. Carroll and Hall, 1988) or spectral cut-off (e.g. Carrasco and Florens, 2002). However, a problem frequently encountered with most of these techniques is that the proposed estimator is not everywhere positive, therefore is not a valid probability density.

The goal of this paper is to present an estimator that is optimal in terms of asymptotic rates of convergence, and that benefits from good finite sample properties. Furthermore, the proposed estimator 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. In particular, they can estimate irregular functions (in Besov spaces) with optimal rates of convergence. 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. The estimator we propose below, in addition to be a valid density, is fully computable as it only involves finite sums in finite sample. Other recent wavelet estimators for deconvolution problems include the work of Johnstone, Kerkyacharian, Picard and Raimondo (2004) or De Canditiis and Pensky (2006), see also the references theirin.

Our estimator combines wavelet thresholding with information projection that guaranties 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. In the special case of Poisson inverse problems, Antoniadis and Bigot (2006) combined this technique with estimation by wavelet expansions.

It is well-known that the difficulty of the deconvolution problem is quantified by the smoothness of the noise density f^ϵ. If f_ℓ^Y, f_ℓ^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. This phenomenon was systematically studied by Fan (1991), who introduced the following two types of assumption on the smoothness of f^ϵ.

Assumption 1.1 *Ordinary smooth convolution: the Fourier coefficients of f^ϵ decay in a polynomial fashion i.e. there exists a constant C and a real $\nu \geqslant 0$ such that $|f_\ell^\epsilon| \sim C|\ell|^{-\nu}$.*

Assumption 1.2 *Super smooth convolution: the Fourier coefficients of f^ϵ are such that*

$$d_1|\ell|^{\nu_0}\exp(-|t|^\nu/d_0) \leqslant |f_\ell^\epsilon| \leqslant d_2|\ell|^{\nu_1}\exp(-|t|^\nu/d_0) \text{ as } |\ell| \to \infty,$$

where $d_0, d_1, d_2, \nu, \nu_0, \nu_1$ are some positive constants.

In this paper, we also consider these two smoothness assumptions. The optimal rate of convergence we can expect from a linear or a nonlinear wavelet estimator depends on these smoothness assumptions and are well-studied in the literature. To summarize, we know from the work of Pensky and Vidakovic (1999); Fan and Koo (2002) that for ordinary smooth convolution both linear and nonlinear wavelet estimators achieve the optimal rate of convergence. This rate is of polynomial order of the sample size n. However, no adaptive linear estimator are optimal, and only well-calibrated nonlinear wavelet estimators are adaptive. For the case of super smooth convolution, the optimal rate of convergence is only of logarithmic order of the sample size, and there is no difference between the rate of convergence of linear and nonlinear estimators. These results are recalled in Section 3 below. It is worth mentionning that the estimators we define in this paper achieve these optimal rates of convergence.

The next section recalls some general results on wavelet approximation and the definition of the Meyer wavelet which is used for deconvolution. Then Section 3 defines the linear and nonlinear wavelet estimators by information projection. The (optimal) rate of convergence of the proposed estimators are stated in Section 4, and are proved in Section 5. The loss function we consider to calculate this rate is the Kullback-Leibler divergence. Due to the aforementioned difference with the wavelet estimator of Fan and Koo (2002), their technique of proof is very different from the proof presented in this paper. Our proof is actually based on a combinaison of the Gaussian approximation technique developed in Donoho, Johnstone, Kerkyacharian and Picard (1995) and other results on Kullback-Leibler divergence by Csiszár (1975) or Barron and Sheu (1991).

Section 6 addresses the practical issues of the proposed estimation procedure. We compare the performance of the proposed estimator with two of the most recent techniques for density deconvolution. The first is deconvolution via cosine series studied by Hall and Qiu (2005), and the second is the model selection approach of Comte, Rozenholc and Taupin (2006a). While the estimator by model selection showed significant small sample improvements against most of the standard techniques of deconvolution, the proposed wavelet-based estimator by information projection outperforms the results of Comte *et al*. (2006a).

We conclude the paper by a small technical appendix, where we adapt some results

of Barron and Sheu (1991) to the case of estimation by information projection using Meyer wavelets.

2 Meyer wavelets for deconvolution

In this paper, we assume that the support of f^X is compact and included in $[0,1]$. The support of f^ϵ however can be unbounded, so the support of f^Y is in general unbounded[1].

Wavelet systems provide unconditional bases for Besov spaces. Using wavelets, one can express whether or not f^X belongs to a Besov space by a simple requirement on the absolute value of the wavelet coefficients of f^X. More precisely, assume some scaling and wavelet functions (ϕ, ψ) that are both in $C^L(\mathbb{R})$, with $L > s$. If $\sigma = s + (1/2 - 1/p) \geqslant 0$, define the norm $\|\cdot\|_{s,p,q}$ by

$$\|f^X\|_{s,p,q}^q = \sum_{j=0}^{\infty} \left(2^{j\sigma p} \sum_{k=0}^{2^j-1} |\langle g, \psi_{j,k} \rangle|^p \right)^{q/p}.$$

It can be shown (Meyer, 1992) that this norm is equivalent to the norm in traditional Besov space $B_{p,q}^s$, that is, there exist strictly positive constants A and B such that

$$A\|g\|_{s,p,q} \leqslant \|g\|_{B_{p,q}^s} \leqslant B\|g\|_{s,p,q}.$$

Note that the condition $\sigma \geqslant 0$ is imposed to ensure that $B_{p,q}^s[0,1]$ is a subspace of $L^2[0,1]$.

The estimator we shall define in the next section is based on the wavelet decomposition of functions in $L^2([0,1])$ using Meyer wavelets (Meyer, 1992). Let (ϕ, ψ) be the 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 (e.g. $\lambda = (j,k)$ 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 g of $L^2([0,1])$, its wavelet decomposition can be written as:

$$g = \sum_{|\lambda|=j_0} c_\lambda \phi_\lambda + \sum_{j=j_0}^{\infty} \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda,$$

[1]The case where the support of f^X is included in $[0,T]$ is handled by adapting the Fourier tranform (the corresponding exponential orthogonal system is $\exp(-i2\pi x\ell/T)$).

where $c_\lambda = \langle g, \phi_\lambda \rangle = \int_0^1 g(u)\phi_\lambda(u)du$, $\beta_\lambda = \langle g, \psi_\lambda \rangle = \int_0^1 g(u), \psi_\lambda(u)du$ and j_0 denotes the usual coarse level of resolution. One reason of using Meyer wavelets in the context of deconvolution is because they are band limited, i.e. their Fourier transform have compact support. In particular we have that the support of $\mathcal{F}\phi$ is $[-4\pi/3, 4\pi/3]$ and the support of $\mathcal{F}\psi$ is $[-8\pi/3, -2\pi/3] \cup [2\pi/3, 8\pi/3]$, where $\mathcal{F}f$ denotes the Fourier transform of a function f. Let $e_\ell(x) = e^{2\pi i \ell x}, \ell \in \mathbb{Z}$ and denote by $f_\ell = \langle f, e_\ell \rangle = \int_0^1 f(u)e^{-2\pi i \ell u}du$ the Fourier coefficients of a function $f \in L^2([0,1])$. Then, if we denote the Fourier coefficients of ψ_λ by $\psi_\ell^\lambda = \langle \psi_\lambda, e_\ell \rangle$ we obtain with the Plancherel's identity that

$$\beta_\lambda = \langle f^X, \psi_\lambda \rangle = \sum_\ell f_\ell^X \psi_\ell^\lambda.$$

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 $f_\ell^\epsilon = \mathbb{E}(e^{-2\pi i \ell \epsilon_1})$ the characteristic function of the ϵ_j's and by $f_\ell^Y = \mathbb{E}(e^{-2\pi i \ell Y_1})$ the characteristic function of the Y_j's , we have by independence of X_1 and ϵ_1 that

$$f_\ell^Y = \mathbb{E}(e^{-2\pi i \ell Y_1}) = \mathbb{E}(e^{-2\pi i \ell \epsilon_1})\mathbb{E}(e^{-2\pi i \ell X_1}) = f_\ell^\epsilon f_\ell^X.$$

An unbiased estimator of β_λ is thus given by

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

provided that the f_ℓ^ϵ's are non-zero and have a sufficiently smooth decay as ℓ tends to infinity. In (2.1), $n^{-1}\sum_{j=1}^n \exp(-2\pi i \ell Y_j)$ is simply the discrete Fourier transform of the observations $Y_1, \ldots, Y_n$.

We define the estimators of the scaling coefficients c_λ analogously, with ϕ instead of ψ. From these estimators, we contruct in the next section our estimator of the unknown density function f^X.

3 Estimation by information projection

3.1 Linear and nonlinear wavelet estimators

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

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

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 $2^{j_1} \approx n^{\frac{1}{2s+2\nu+1}}$ 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. For super smooth convolution (Assumption 1.2), the optimal and adaptive choice is $2^{j_1} \approx (\ln n)^{1/\nu}$, see Pensky and Vidakovic (1999) or Fan and Koo (2002).

Contrary to the nonlinear estimator, there exists adaptive nonlinear estimators by wavelet thresholding that can achieve the optimal rate of convergence. The non-linear estimation by hard-thresholding is defined by

$$\hat{f}_h^X = \sum_{|\lambda|=j_0} \hat{c}_\lambda \phi_\lambda + \sum_{j=j_0}^{j_1} \sum_{|\lambda|=j} \delta^h_{\tau_{j,n}}(\hat{\beta}_\lambda) \psi_\lambda$$

with threshold $\delta^h_{\tau_{j,n}}(x) = x \mathbb{1}_{\{|x| \geqslant \delta_{j,n}\}}$ and the non-linear estimation by soft-thresholding is defined by

$$\hat{f}_s^X = \sum_{|\lambda|=j_0} \hat{c}_\lambda \phi_\lambda + \sum_{j=j_0}^{j_1} \sum_{|\lambda|=j} \delta^s_{\tau_{j,n}}(\hat{\beta}_\lambda) \psi_\lambda$$

where $\delta^s_{\tau_{j,n}}(x) = sign(x)(x - \delta_{j,n})_+$. These estimators depend on the coarse level of approximation j_0, the high-frequency cut-off j_1 and the threshold $\tau_{j,n}$ which may depend on the level of resolution j. The estimators $\hat{f}_h^X$ and $\hat{f}_s^X$ have already been proposed as estimators of f^X. An optimal adaptative 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 $2^{j_0} \approx \log n$, $2^{j_1} \approx n^{\frac{1}{2\nu+1}}$ and $\delta_{j,n} \approx 2^{\nu j}/\sqrt{n}$ (Pensky and Vidakovic, 1999). The choice $\delta_{j,n} \approx 2^{\nu j}\sqrt{j/n}$ was also considered (Fan and Koo, 2002).

Since all of these estimators are of a form of orthogonal series estimator, they are not in general in the space of valid densities. In particular, they are not necessarily positive everywhere. In the next section, we modify the linear and nonlinear estimator using a projection step to guaranty positivity.

3.2 Information projection to guaranty positivity

To simplify the notations, we write in the following $(\psi_\lambda)_{|\lambda|=j_0-1}$ for the scaling functions $(\phi_\lambda)_{|\lambda|=j_0}$.

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:

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

where

$$C_j(\theta) = \log \int_0^1 \exp\left(\sum_{|\lambda|<j} \theta_\lambda \psi_\lambda(x) dx \right).$$

The constant $C_j(\theta)$ is used to guarantee that $f_{j,\theta}$ is integrating to one on $[0,1]$, and is thus a probability density function.

Following Csiszár (1975), the density function $f_{j,\theta}$ in the exponential family $\mathcal{E}_j$ that is the closest to the true density f^X in the Kullback-Leibler sense is characterized as the unique density function in the family for which

$$\langle f_{j,\theta}, \psi_\lambda \rangle = \langle f^X, \psi_\lambda \rangle, \text{ for all } |\lambda| < j.$$

It seems therefore natural to estimate the unknown density function f^X, by searching 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 $f_{j,\hat{\theta}_n}$ such that $\langle f_{j,\hat{\theta}_n}, \psi_\lambda \rangle = \hat{\alpha}_\lambda$ for all $|\lambda| < j_1$
- The *positive nonlinear estimator with hard thresholding* is $f_{j,\hat{\theta}_n}$ such that $\langle f_{j,\hat{\theta}_n}, \psi_\lambda \rangle = \delta^h_{\tau_{j,n}}$ for all $|\lambda| < j_1$
- The *positive nonlinear estimator with soft thresholding* is $f_{j,\hat{\theta}_n}$ such that $\langle f_{j,\hat{\theta}_n}, \psi_\lambda \rangle = \delta^s_{\tau_{j,n}}$ for all $|\lambda| < j_1$

The existence of these estimators is questionnable. This issue is addressed in the next section and in the technical appendix. We also derive in the next section the rate of convergence of the estimators.

4 Asymptotic optimality of the estimators

To calculate the rate of convergence of the estimators, we use the loss function given by the Kullback-Leibler discrepancy between two probability density functions p and q:

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

Let M be some fixed constant and let $F^s_{p,q}(M)$ denote the set of density functions such that

$$F^s_{p,q}(M) = \left\{ f \in L^2[0,1] \text{ is a p.d.f. such that for } g = \log f,\ \|g\|_{B^s_{p,q}} \leqslant M \right\}.$$

Note that assuming that $f \in F^s_{p,q}(M)$ implies that f is strictly positive.

4.1 Linear estimation

The following theorem is the general result on 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 $2^{-j(n)} \approx n^{-1/(2s+2\nu+1)}$. Then, the information projection estimator $f_{j(n),\hat{\theta}_n}$ exists and is such that*

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

The above estimator 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 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

Fan and Koo (2002) show that, when the error is supersmooth, optimal rates of convergence are only of logarithmic order of the sample size. In this case, while the linear wavelet estimators cannot be optimal, non linear estimators do not provide much gain for estimating functions in the Besov spaces. For this reason, we only consider the ordinary smooth case in the following.

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 $\tau_{j,n} = \tau_j\sqrt{(\log n)/n}$ with

$$\tau_j = 2^{j\nu}. \tag{4.1}$$

In what follows, we shall assume that the coarse level of approximation j_0 is a fixed parameter whose choice is left to the statistician. 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-possedness 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 indicates that the rate of convergence of the expected Kullback-Leibler discrepancy for the positive nonlinear estimator by soft thresholding achieves the optimal rate of convergence provided that the finest resolution level j_1 is an apppropriate function of the degree of smoothness ν of the convolution.

Theorem 4.2 *Assume that $f \in F^s_{p,p}(M)$ with $1/p = 1/2 + s/(2\nu+1)$, and suppose that the convolution kernel f^ϵ satisfies Assumption 1.1 with $\nu > 1/2$ (ordinary smooth convolution). Moreover, suppose that $s > \nu + 1/2$, that the wavelet ψ is C^L with $L > s$ and that ψ has r vanishing moments with $r > s$. Then, the above described soft-thresholding estimator satisfies*

the minimax rate (up to logarithmic factors)

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

provided that $j_1(n)$ *is such that* $2^{j_1(n)} \geqslant n^{1/2\nu}$.

The proof of Theorem 4.2 is to be found in the next section.

Remark 4.1 *For the nonlinear estimator based on hard thresholding, similar results can be obtained using for instance the maxiset theorems given in Johnstone et al. (2004). The threshold is also level-dependent and the choice of* $j_1(n)$ *depends on* ν. *Similar adaptive results are then obtained with* $2^{j_1(n)} = \mathcal{O}(\frac{n}{log(n)})^{1/(2\nu+1)}$ *and the same choice for the threshold as the one given by equation (4.1).*

5 Proof of the theorems

The proof of the two 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

$$\Delta(f^X; f_{j,\hat{\theta}_n}) = \Delta(f^X; f_{j,\theta^*_j}) + \Delta(f_{j,\theta^*_j}; f_{j,\hat{\theta}_n}) \tag{5.1}$$

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 Pythagorian Theorem derived in Csiszár (1975) and recalled in Lemma A.1 of the appendix. 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 first prove the existence and uniqueness of f_{j,θ^*_j}. Based on some inequalities derived in Barron and Sheu (1991), we show that the approximation error is controlled by the norms $\|g - P_j g\|_{L^2}$ and $\|g - P_j g\|_\infty$, where $g = \log(f^X)$ and $P_j g = \sum_{|\lambda|<j} \langle g, \psi_\lambda \rangle \psi_\lambda$. Bounds for these norms are derived in the technical appendix below.

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 $\mathbb{E}\|\hat{\alpha}_n - \alpha_0\|_2^2$ (using Lemma A.6 from the technical appendix). In the nonlinear situation, we use that our density deconvolution problem is not too far from a usual Gaussian white noise model. This allows to use standard results in the Gaussian setting on soft-threholding estimators with a level-dependent threshold. This reasoning follows the technique initially proposed by Donoho, Johnstone, Kerkyacharian and Picard (1996, Section 6), and adapted to the case of Poisson inverse problems by Cavalier and Koo (2002) and Antoniadis and Bigot (2006). Below we adapt the technique to density deconvolution with Meyer wavelets.

5.1 Proof of Theorem 4.1

This proof concerns the linear, non adaptative estimator.

We first control the approximation error term $\Delta(f^X; f_{j(n),\theta^*_{j(n)}})$. Let $g = \log(f^X) = \sum_{j=-1}^{\infty}\sum_{|\lambda|=j}\beta_\lambda\psi_\lambda$, and for all $|\lambda| < j$, define the wavelet coefficients $\alpha_{j,\lambda} = \langle \exp(P_j g), \psi_\lambda\rangle$ and $\alpha_\lambda = \langle f^X, \psi_\lambda\rangle$. The Bessel's inequality gives $\|\alpha_j - \alpha\|_2^2 \leqslant \|f^X - \exp(P_j g)\|_{L^2}^2$. Therefore, Lemma A.4 implies

$$\|\alpha_j - \alpha\|_2^2 \leqslant M_1 \int \frac{(f^X - \exp(P_j g))^2}{f^X} d\mu.$$

Now, using Lemma 2 of (Barron and Sheu, 1991), we can write

$$\|\alpha_j - \alpha\|_2^2 \leqslant M_1 e^{2\|g - P_j g\|_\infty} \int f^X (g - P_j g)^2 d\mu \leqslant M_1^2 e^{2\gamma_j} D_j^2.$$

where $D_j = \|g - P_j g\|_{L^2}$ and $\gamma_j = \|g - P_j g\|_\infty$.

Define $\epsilon_j = 2M_1^2 e^{2\gamma_j + 1} D_j A_j$. Lemma A.2 with $\theta_{0,\lambda} = \beta_\lambda$, $\alpha_\lambda = \langle f^X, \psi_\lambda\rangle$ for all $|\lambda| < j$ and $b = \exp\{\|\log(\exp(P_j g))\|_\infty\}$ implies that $\theta_j^* = \theta(\alpha)$ exists provided that $M_1 e^{\gamma_j} D_j \leqslant (2ebA_j)^{-1}$. This last condition is fulfilled if $\epsilon_j \leqslant 1$ because $\|\log(\exp(P_j g))\|_\infty \leqslant \log M_1 + \gamma_j$.

From Lemma A.1 we can write $\Delta(f^X; f_{j,\theta_j^*}) \leqslant \Delta(f^X; \exp(P_j g))$. Thence, by Lemma 1 of Barron and Sheu (1991),

$$\begin{aligned}\Delta(f^X; f_{j,\theta_j^*}) &\leqslant \frac{1}{2}\exp(\|g - P_j g\|_\infty) \int f^X (g - P_j g)^2 d\mu \\ &\leqslant \frac{1}{2} M_1 e^{\gamma_j} D_j^2 \end{aligned} \tag{5.2}$$

Now let $j(n)$ be such that $2^{j_1(n)} \geqslant n^{1/2\nu}$. As $f^X \in F^s_{2,2}(M)$ with $s > 1$ by assumption, it follows from the bounds on A_j, D_j and γ_j given in Lemma A.5 that $\gamma_{j(n)} \to 0$ as $n \to \infty$ and so $\epsilon_j = \epsilon_{j(n)} = \mathcal{O}(A_{j(n)}\Delta_{j(n)}) = \mathcal{O}(2^{-j(n)(s-1)})$. Since $\epsilon_{j(n)} \to 0$ as $n \to \infty$, equation (5.2) implies 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

$$\Delta(f^X; f_{j(n),\theta^*_{j(n)}}) = \mathcal{O}\left(2^{-2j(n)s}\right). \tag{5.3}$$

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 $\hat{\alpha}_{n,\lambda} = \sum_l \left(\frac{\psi^\lambda_l}{f^\epsilon_l}\right)\left(\frac{1}{n}\sum_{j=1}^n e^{-2\pi i l Y_j}\right)$. To prove the existence of a vector $\hat{\theta}_n \in \mathbb{R}^{2^{j(n)}}$ such that

$$\langle f_{j,\hat{\theta}_n}, \psi_\lambda \rangle = \hat{\alpha}_{n,\lambda}, \text{ for all } |\lambda| < j(n),$$

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 Lemma A.2 with $\theta_0 = \theta^*_{j(n)}$, $\alpha = \hat{\alpha}_n$ and $b = \exp\{\|\log(f_{j(n),\theta^*_{j(n)}})\|_\infty\}$. Given our assumption on f^X and f^ϵ we have that $|f^Y_l| \leqslant C|l|^{-(s+\nu)}$ with $s + \nu > 1$, and we can therefore apply Lemma A.6 to obtain that

$$\mathbb{E}\|\hat{\alpha}_n - \alpha_0\|_2^2 \leqslant \frac{C}{n} 2^{j(n)(2\nu+1)}$$

Note we have that

$$\|\log(f_{j(n),\theta^*_{j(n)}} / \exp(P_{j(n)}g))\|_\infty \leqslant \epsilon_{j(n)}$$

and so $b \leqslant M_1 e^{\epsilon_{j(n)}+\gamma_{j(n)}}$. Hence, if we set $\delta_{j(n)} = 2M_1 e^{\epsilon_{j(n)}+\gamma_{j(n)}+1} A_{j(n)} 2^{j(n)(\nu+1/2)}/\sqrt{n}$, we can write $\delta_{j(n)} = \mathcal{O}(2^{j(n)(\nu+3/2)}/\sqrt{n}) = \mathcal{O}(2^{-j(n)(s-1)}) \to 0$ as $n \to \infty$. Hence, by Lemma A.6 we have that for n sufficiently large, $\hat{\theta}_n$ exists and is such that

$$\mathbb{E}\left(\Delta(f_{j(n),\theta^*_{j(n)}}; f_{j(n),\hat{\theta}_n})\right) = \mathcal{O}\left(2^{-2j(n)s}\right), \tag{5.4}$$

The result of the theorem now follows from the control of the approximation and estimation error terms, using the identity (5.1). □

5.2 Proof of Theorem 4.2

We consider in this proof the nonlinear, adaptative estimator, and first control the approximation error term in a very similar way than in the preceeding proof. By

proceeding as in the proof of Theorem 4.1, we easily show that for n sufficiently large

$$\Delta(f^X; f_{j_1(n),\theta^*_{j_1(n)}}) = \mathcal{O}\left(2^{-4j_1(n)s\frac{\nu}{2\nu+1}}\right),$$

using the notations from the proof of Theorem 4.1 for $f_{j_1(n),\theta^*_{j_1(n)}}$. Now, one obtains the optimal order $n^{-2s/(2s+2\nu+1)}$ provided $j_1(n)$ is large enough so that $2^{-j_1(n)} \leqslant n^{-\frac{\nu+1/2}{\nu 2s+2\nu+1}}$. Note that for $n > 1$, $n^{-1/(2\nu)} \leqslant n^{-\frac{\nu+1/2}{\nu(2s+2\nu+1)}}$. Therefore, if $2^{j_1(n)} \geqslant n^{1/(2\nu)}$,

$$\Delta(f^X; f_{j_1(n),\theta^*_{j_1(n)}}) = \mathcal{O}\left(n^{-2s/(2s+2\nu+1)}\right)$$

We can now consider the estimation error term. Define $\hat{\alpha}_{n,\lambda}$ and α_λ as in the proof of Theorem 4.1. Note that

$$\mathbb{E}\|\delta^s_{\tau_{j,n}}(\hat{\alpha}_n) - \alpha_0\|^2 = \sum_{|\lambda|=j_0-1} \mathbb{E}(\hat{\alpha}_{n,\lambda} - \alpha_\lambda)^2 + \sum_{j_0\leqslant|\lambda|<j_1(n)} \mathbb{E}(\delta^s_{\tau_{j,n}}(\hat{\alpha}_{n,\lambda}) - \alpha_\lambda)^2$$

Given our assumptions on f^X and the fact that the space $B^s_{p,p}$ is continuously embedded in H^t whenever $t \leqslant s + 1/2 - 1/p = 2\nu s/(2\nu+1)$ we have that $|f^Y_l| \leqslant C|l|^{-2\nu s/(2\nu+1)-\nu}$. Given the fact that $s > \nu + 1/2$ and $\nu > 1/2$, we have that $2\nu s/(2\nu+1) + \nu > 1$ and we can then apply Lemma A.6 to show that

$$\sum_{|\lambda|=j_0-1} \mathbb{E}(\hat{\alpha}_{n,\lambda} - \alpha_\lambda)^2 \leqslant C2^{j_0(2\nu+1)}/n \tag{5.5}$$

We will now derive an upper bound for the risk $\mathbb{E}\|\delta^s_{\tau_{j,n}}(\hat{\alpha}_{n,\lambda}) - \alpha_\lambda\|^2$. As $|f^\epsilon_l| \sim |l|^{-\nu}$ we have that

$$\sum_{j_0\leqslant|\lambda|<j_1(n)} \mathbb{E}(\delta^s_{\tau_{j,n}}(\hat{\alpha}_{n,\lambda}) - \alpha_\lambda)^2 \leqslant C \sum_{j_0\leqslant|\lambda|<j_1(n)} 2^{2|\lambda|\nu}\mathbb{E}(\delta^s_{\tau_{j,n}}(\hat{\alpha^*}_{n,\lambda}) - \alpha^*_\lambda)^2 \tag{5.6}$$

where $\alpha^*_\lambda = \sum_l \psi^\lambda_l f^Y_l$ and $\hat{\alpha^*}_{n,\lambda} = n^{-1}\sum_{j=1}^n \sum_l \psi^\lambda_l e^{-2\pi i l Y_j}$

To bound equation (5.6) we define a suitable Gaussian approximation of $\hat{\alpha^*}_{n,\lambda}$, following the construction in Section 6 of Donoho *et al*. (1996). We first note that $\mathbb{E}\hat{\alpha^*}_{n,\lambda} = \alpha^*_\lambda$ and $\mathrm{Var}\hat{\alpha^*}_{n,\lambda} = n^{-1}\sigma^2_\lambda$ with $\sigma^2_\lambda \leqslant C$ (by Lemma A.6 using again the fact that $|f^Y_l| \leqslant C|l|^{-2\nu s/(2\nu+1)-\nu}$). The normalized statistic

$$U_j = \frac{\sum_l \psi^\lambda_l \exp(-2\pi i l Y_j) - \alpha^*_\lambda}{\sigma_\lambda}$$

is such that $|U_j| \leqslant C2^{-|\lambda|/2}/\sigma_\lambda$. Now define the Gaussian process $\hat{\gamma}_\lambda = \alpha^*_\lambda + Z_\lambda\sigma_\lambda/\sqrt{n}$ where Z_λ are independant standard normal variables. If $\sigma^2_\lambda \geqslant C2^{|\lambda|}(\log^3 n)/n$, then

$U_i^2 \leqslant Cn/(\log^3 n)$ and from Lemma 2 of (Donoho *et al.*, 1996), we get $\mathbb{E}(\hat{\alpha^*}_{n,\lambda} - \hat{\gamma}_\lambda)^2 \leqslant C2^{|\lambda|}/n^2$. If $\sigma_\lambda^2 < C2^{|\lambda|}(\log^3 n)/n$, then $\mathbb{E}(\hat{\alpha^*}_{n,\lambda} - \hat{\gamma}_\lambda)^2 \leq 2\mathrm{Var}(\hat{\alpha^*}_{n,\lambda}) + 2\sigma_\lambda^2/n \leqslant C2^{|\lambda|}(\log^3 n)/n^2$. We finally get for all λ, T, and for all σ_λ^2,

$$\mathbb{E}\left(\hat{\alpha}^*_{n,\lambda} - \hat{\gamma}_\lambda\right)^2 \leqslant C2^{|\lambda|}\frac{\log^3 n}{n^2} .$$

If $r(\delta^s_{\tau_{j,n}}, \sigma_\lambda/\sqrt{n}; \alpha^*_\lambda)$ denotes the mean square error for the Gaussian model $\mathbb{E}(\delta^s_{\tau_{j,n}}(\hat{\gamma}_\lambda) - \alpha^*_\lambda)^2$, then we can write

$$\mathbb{E}(\delta^s_{\tau_{j,n}}(\alpha^*_\lambda) - \alpha^*_\lambda)^2 \leq 2\mathbb{E}(\delta^s_{\tau_{j,n}}(\hat{\alpha}^*_{n,\lambda}) - \delta^s_{\tau_{j,n}}(\hat{\gamma}_\lambda))^2 + 2r(\delta^s_{\tau_{j,n}}, \sigma_\lambda/\sqrt{n}; \alpha^*_\lambda) .$$

Using that the mapping $y \to \delta^s_{\tau_{j,n}}(y)$ is a contraction, the first term is bounded by $2\mathbb{E}(\hat{\alpha^*}_{n,\lambda} - \hat{\gamma}_\lambda)$ and thus bounded by $C2^{|\lambda|}n^{-2}\log^3 n$. Moreover, as $\sigma_\lambda \leqslant C$, we use Lemma 3 of Donoho *et al.* (1996) to bound the second term by $4r(\delta^s_{\tau_{j,n}}, C/\sqrt{n}; \alpha^*_\lambda)$. Finally, we obtain that

$$\sum_{j_0 \leqslant |\lambda| < j_1(n)} \mathbb{E}(\delta^s_{\tau_{j,n}}(\hat{\alpha}_{n,\lambda}) - \alpha_\lambda)^2$$
$$\leqslant C \sum_{j_0 \leqslant |\lambda| < j_1(n)} 2^{|\lambda|(2\nu+1)} n^{-2}\log^3 n + 4r(\delta^s_{\tau_{j,n}}, C2^{\nu|\lambda|}/\sqrt{n}; 2^{\nu|\lambda|}\alpha^*_\lambda)$$

Lemma A.7 below shows that the sequence α^*_λ are the wavelet coefficients of a function in $B^{s+\nu}_{p,p(\tilde{M})}$ for some some finite constant $\tilde{M}$. Then, using the level dependent threshold $\tau_{j,n} = 2^{\nu j} n^{-1/2}\sqrt{|\log n|}$ and the fact that $2^{j_1(n)} \geqslant n^{1/(2\nu)}$ we obtain the final rate from standard results in the Gaussian setting on soft wavelet thresholding (e.g. Donoho *et al.*, 1995) and level-dependent thresholding estimators (e.g. Cohen, DeVore and Hochmuth, 2000). □

6 Simulations

In this section we report the result of simulations, and compare our procedure with other deconvolution methods recently introduced in the literature.

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 realisations 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 consider for f_ϵ the Laplace density function, given by

$$f^\epsilon(x) = \frac{1}{\sqrt{2}\sigma_\epsilon} \exp\left(-\sqrt{2}\frac{|x|}{\sigma_\epsilon}\right), \qquad x \in \mathbb{R}.$$

The Fourier coefficients of this density are given by

$$f^\epsilon_\ell = \frac{1}{1 + 2\sigma_\epsilon^2\pi^2\ell^2}, \qquad \ell = 0, \pm 1, \pm 2, \ldots$$

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 five following situations:

1. *Uniform distribution*: $f(x) = 5\mathbb{1}_{[0.4,0.6]}(x)$.
2. *Exponential distribution*: $f(x) = 10e^{-10(x-0.2)}\mathbb{1}_{[0.2,+\infty[}(x)$
3. *Laplace distribution*: $f(x) = 10e^{-20|x-0.5|}$
4. *Gaussian distribution*: $X \sim N(\mu, \sigma^2)$ with $\mu = 0.5$ and $\sigma = 0.1$.
5. *MixtGauss distribution (mixture of two Gaussian variables)*: $X \sim \pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2)$ 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 five densities f^X are displayed in Figure 6.1, where we can observe that they show various types of smoothness. The Uniform distribution is a piecewise constant function with two jumps, the Exponential disitbution 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 Gaussian and the MixtGauss densities are very smooth signals (analytical functions). Due to the excellent localization properties of the wavelets for the reconstruction of irregular signals, it is expected that our wavelet-based estimator is well-adapted to these types of irregularity. Although all these distributions are not necessarily compactly supported on $[0, 1]$, they have been chosen so that their mass is essentially concentrated on $[0, 1]$ and it is therefore very unlikely to have observations X_i outside the interval $[0, 1]$.

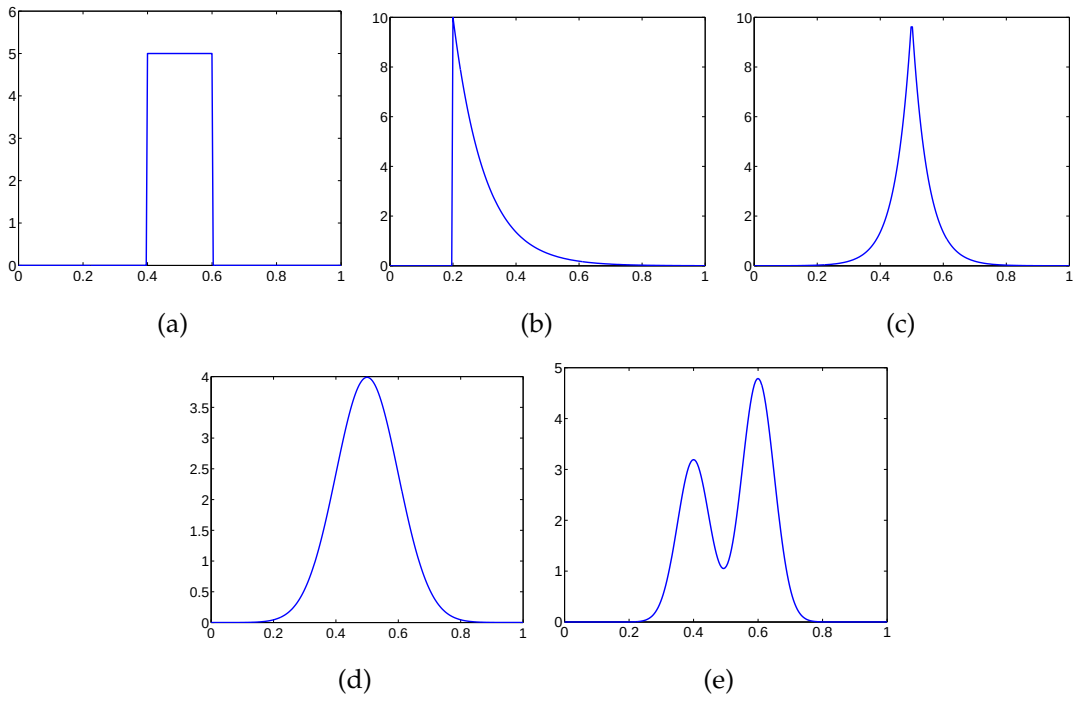


Figure 6.1: Test densities: (a) Uniform, (b) Exponential, (c) Laplace, (d) Gaussian, (e) MixtGauss (mixture of two Gaussian)

6.1 Computation of the estimators

In the following, we describe in detail the computation algorithm of the wavelet deconvolution by information projection described in the previous sections. We also introduce two competitors, the estimator by model selection studied by Comte *et al.* (2006a) and cosine series deconvolution of Hall and Qiu (2005). These two procedures have been recently introduced in the literature and their properties on finite samples have been well studied.

In all simulations, we used the Matlab program and the wavelet toolbox Wavelab (see Buckheit, Chen, Donoho and Johnstone, 1995).

6.1.1 Wavelet deconvolution

For $\ell = -n/2+1, \ldots, n/2$ we compute the coefficients

$$\hat{f}_\ell^X = n^{-1} \sum_{j=1}^{n} \exp(-2\pi i \ell Y_j) / f_\ell^\epsilon.$$

This gives an estimation of the Fourier coefficients of the unknown function f^X, and we then use the efficient algorithm of Kolaczyk (1994) to compute the Meyer wavelet coefficients of a discrete signal. This algorithm only requires $\mathcal{O}(n(\log(n))^2)$ operations to compute the empirical wavelet coefficients from a sample of size n.

According to Theorem 4.2, the high-frequency cut-off $j_1(n)$ must be chosen such that $j_1(n) \geqslant (2\nu)^{-1}\log_2(n)$. In practice the optimal level $(2\nu)^{-1}\log_2(n)$ is too small, and in our simulations we have therefore investigated the choices $j_1 = 3$ to $j_1 = \log_2(n) - 1$. For any of these choices, the optimal theoretical level is always smaller than j_1 and introducing a higher level of resolution may only introduce some instability in our estimator (for instance when a large wavelet coefficient due to the noise at a fine scale is erroneously kept by the thresholding procedure). This behavior has also been noticed by Johnstone *et al*. (2004). As we shall see in the simulation results, the best empirical level j_1 depends on the amount of noise and is proportional to the signal to noise ratio. For all the simulations and all test densities, the coarse level j_0 is equal to 3.

For a non-linear wavelet estimator, the results of Theorem 4.2 suggest to take a threshold of the form

$$\tau_{j,n} = C\tau_j\sqrt{(\log n)/n},$$

where $C \geqslant 1$ 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 $C = \sqrt{2}$ rather than $C = 1$. 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

$$\tilde{\tau}_j = \frac{1}{|C_j|}\sum_{\ell\in C_j}|f^\epsilon_\ell|^{-2},$$

where C_j denotes the set of of non-zero Fourier coefficients ψ^λ_ℓ 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 A.6, the choise $\tau_j = 2^{j\nu}$ comes from the bound

$$\tilde{\tau}_j^2 = \frac{1}{|C_j|}\sum_{\ell\in C_j}|f^\epsilon_\ell|^{-2} = \mathcal{O}(2^{2j\nu})$$

under the assumption of ordinary smooth deconvolution. It is not clear whether the scale parameters τ_j and $\tilde{\tau}_j$ yield to 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 $\tilde{\tau}_j$.

Finally, once we have computed the thresholded coefficients $\delta^s_{\tau_{j,n}}(\hat{\alpha}_\lambda)$ for all $|\lambda| < j_1$, it remains to compute the empirical version of the information projection estimate $f^s_{j_1,\hat{\theta}_n}$. To do so, we use a Newton-Raphson type algorithm as described in Antoniadis and Bigot (2006).

6.1.2 Density deconvolution via model selection

The adaptive density deconvolution estimator of Comte *et al*. (2006a) 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 $l_m = m\Delta$, $\Delta > 0$. The adaptive estimator by model selection 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 Fourier transform of the error distribution f^ϵ, see Comte *et al*. (2006a). Based on extensive simulations with various sample sizes and signal to noise ratios, Comte *et al*. (2006a) show that the model selection procedure performs very well for finite samples, compared with the standard estimators. This estimator outperforms the kernel estimator, even when the bandwidth parameter is selected in a data-driven way. In consequence, we see this procedure as the most challenging competitor in our simulations.

6.1.3 Cosine series deconvolution

As an alternative competitor, we also consider the deconvolution estimator recently introduced by Hall and Qiu (2005). The estimator if 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 cosine 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. Since the Laplace density is symmetric

about its mean 0 (recall that this is our choice for the error distribution for all the simulations), a simple estimator of the cosine coefficient a_j is given by

$$\hat{a}_j = \frac{\hat{b}_j}{\alpha_j} \delta_{\tau_n}(|\hat{b}_j|)$$

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. Slight modifications of this threholding rule are also considered in Hall and Qiu (2005), but these modifications have the same theoretical and empirical properties as those based on δ_{τ_n}. Moreover, the simulations carried out by Hall and Qiu (2005) have shown that the simple choice $m = n$ and $C = 2$ leads to satisfactory results and that there is not much to be gained by employing cross-validation to choose m and C. So, in all the simulations presented in this paper we take $m = n$ and $C = 2$.

6.2 Results of the simulations

We present results for sample sizes $n = 128, 256$ and 512 (respectively a small, moderate and large sample size) and $s2n = 100, 10, 3$ (respectively a very low, a moderate and a high level of noise). Note that for $s2n = 100$, as the variance of the noise is extremely low, we are therefore 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 for each sample the quality of an estimator $\hat{f}_n$ of the test density f is measured by the empirical mean square error (MSE) defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_n(t_i) - f(t_i) \right)^2$$

where $t_i = i/n$, $i = 0, \ldots, n-1$. In Figure 6.2, we illustrate the performance of each method and show typical reconstructions of the test densities f^X for $n = 256$ and $s2n = 10$. 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. It is therefore visually much more satisfactory than the model selection estimator and the cosine series estimator which may take negative values. For the three non-smooth

densities (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 two smooth densities (Gauss and MixtGauss), the model selection estimator performs slightly better than the two other methods.

By inspecting the first column in Figure 6.2 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. Their algorithm yields thresholded coefficients $\delta^{s}_{\tau_{j,n}}(\hat{\alpha}_\lambda)$ (for all $|\lambda| < j_1$) invariant to translation that can be used to calculate a TI information projection estimate. In the second column of Figure 6.2 is displayed the TI version of the wavelet estimators plotted in the first column. One can see that the TI estimators are visually much better since they exhibit very small oscillations while preserving a good reconstruction of the singularities of the non-smooth densities. However, from the overall simulations, we have found that the TI version of our wavelet estimator does not yield significant improvements in terms of MSE. Therefore we only present results for the comparison between our wavelet estimator (non-TI version) and the two alternative methods by Comte *et al*. (2006a) and Hall and Qiu (2005).

In Figures 6.3 to 6.7, we depict for each test f^X density the boxplot of the MSE over the 100 replications. All combinations of sample sizes and signal-to-noise ratios are considered. For wavelet deconvolution, we give boxplots for each type of thresholding, either with the scale parameter τ_j (abbreviated as *wavtheo*) or $\tilde{\tau}_j$ (abbreviated as *wavemp*). We also indicate the level j_1 which gives the best result in term of averaged MSE over the 100 simulations. As it can be observed from these boxplots, our wavelet approach outperforms the other methods 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. For the Gaussian distribution, the wavelet approach with scale parameter $\tilde{\tau}_j$ yields generally better results, in particular for a small sample sizes ($n = 128$). With sample sizes $n = 256$ or $n = 512$, the results obtained with the three methods are very similar. Finally, for the MixtGauss distribution, the wavelet approach is clearly better for $n = 128$ while the model selection procedure is slightly

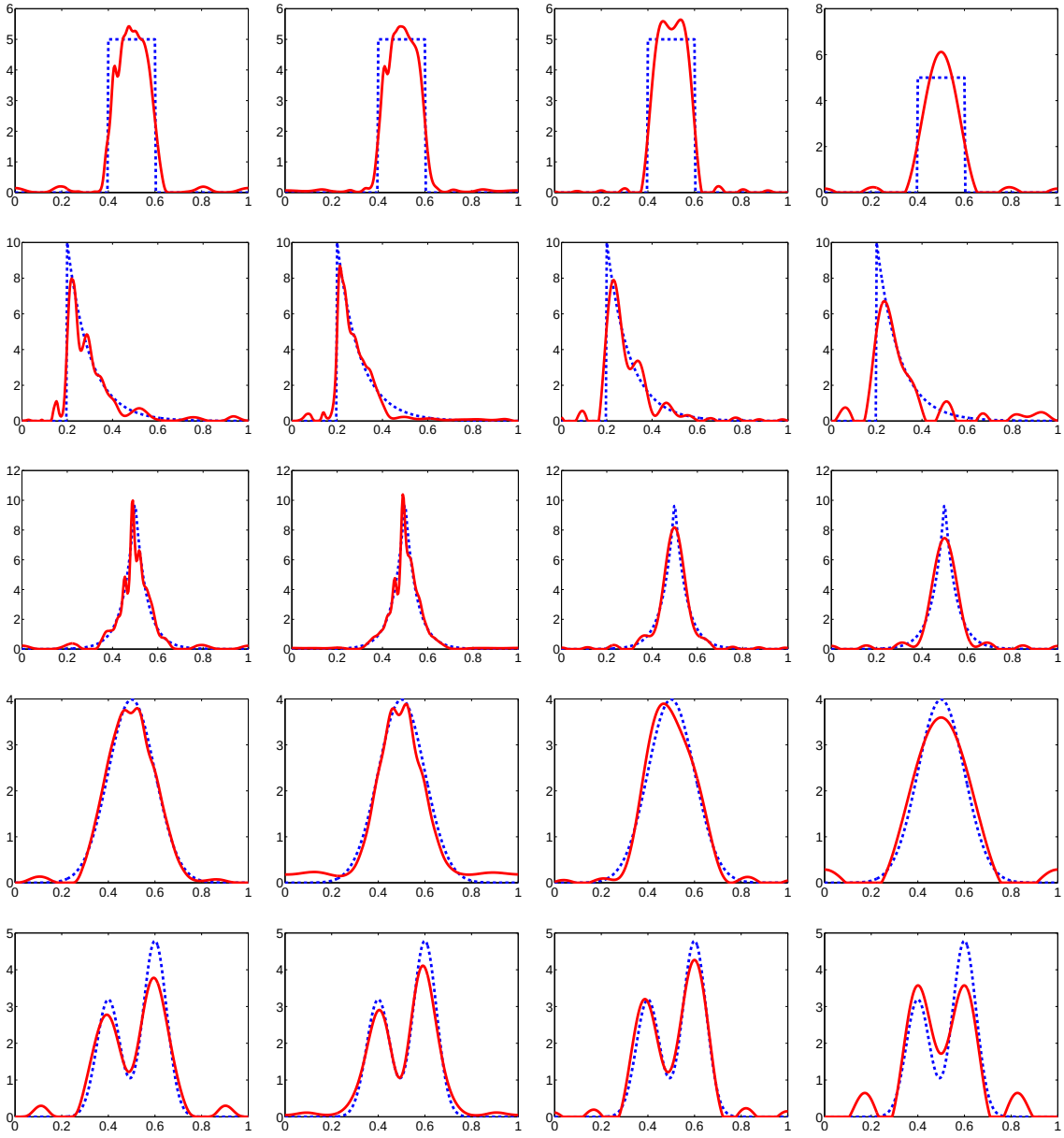


Figure 6.2: Typical reconstructions for one realization of simulations by: wavelet thresholding with $j_1 = 5$ and the "distribution dependent" scale parameter $\tilde{\tau}_j$ (first column the non-TI version, second column the TI version), model selection (third column) and cosine series (fourth column) for the five test densities: Uniform, Exponential, Laplace, Gaussian and MixtGauss. The dotted lines show the true densities and the solide lines correspond to the various estimators ($n = 256$ and $s2n = 10$)

better than wavelet thresholding for $n = 256$ and $n = 512$. Note that the fine level j_1 which gives the best results is generally quite low and depends on the signal to noise ratio. For almost all combinations of the factors, 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.

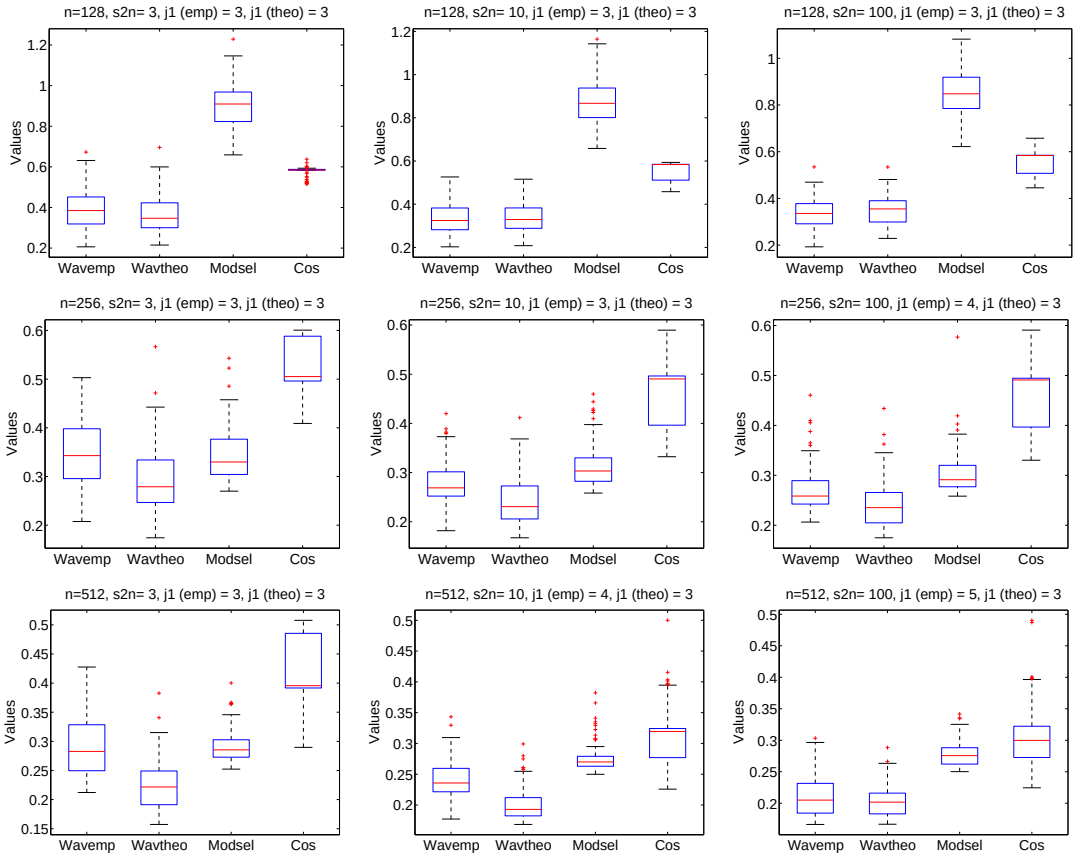


Figure 6.3: **Uniform distribution**: graphical display (boxplots) of the MSE with 100 repetitions for each method and all combination of the factors n and $s2n$.

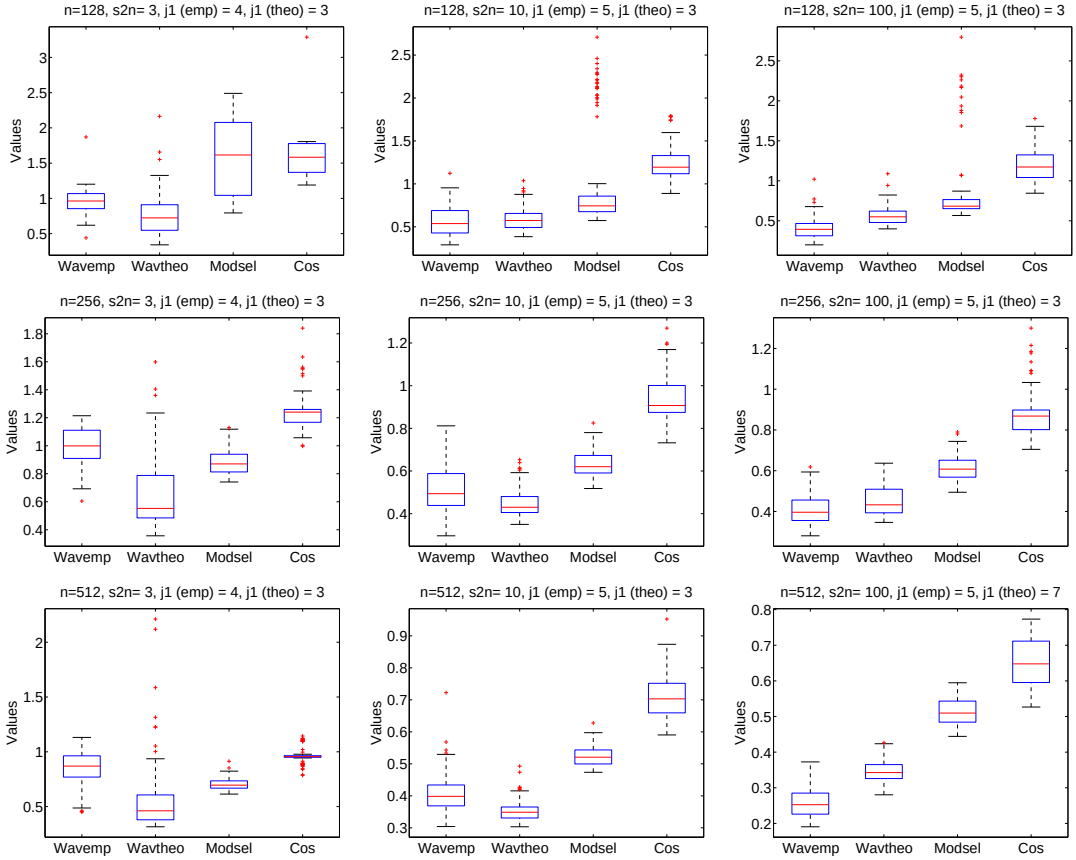


Figure 6.4: **Exponential distribution**: graphical display (boxplots) of the MSE with 100 repetitions for each method and all combination of the factors n and $s2n$.

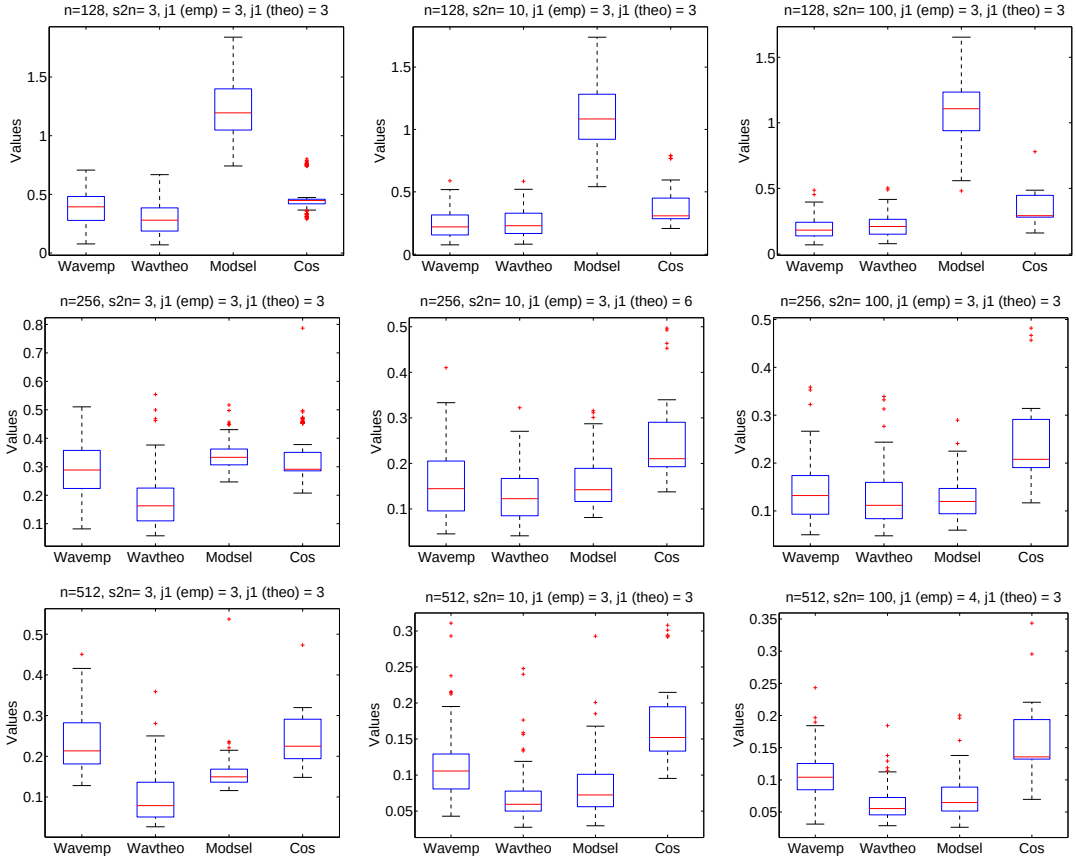


Figure 6.5: **Laplace distribution**: graphical display (boxplots) of the MSE with 100 repetitions for each method and all combination of the factors n and $s2n$.

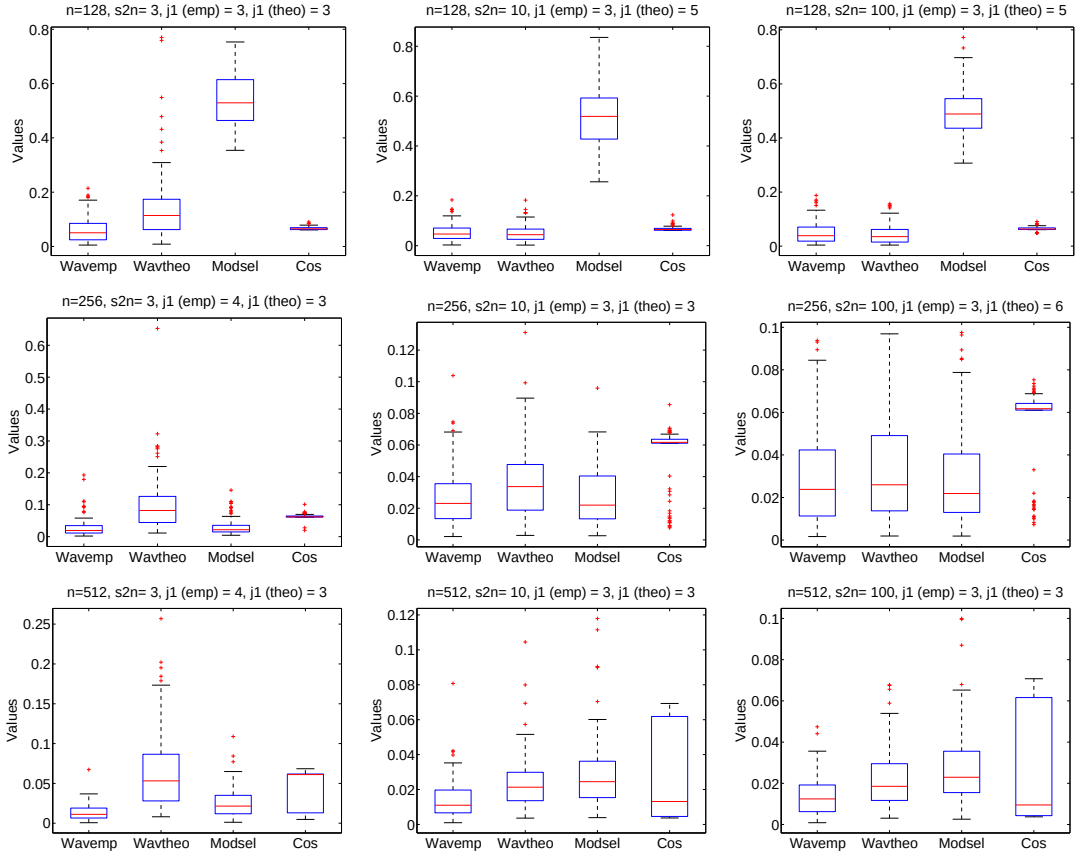


Figure 6.6: **Gaussian distribution**: graphical display (boxplots) of the MSE with 100 repetitions for each method and all combination of the factors n and $s2n$.

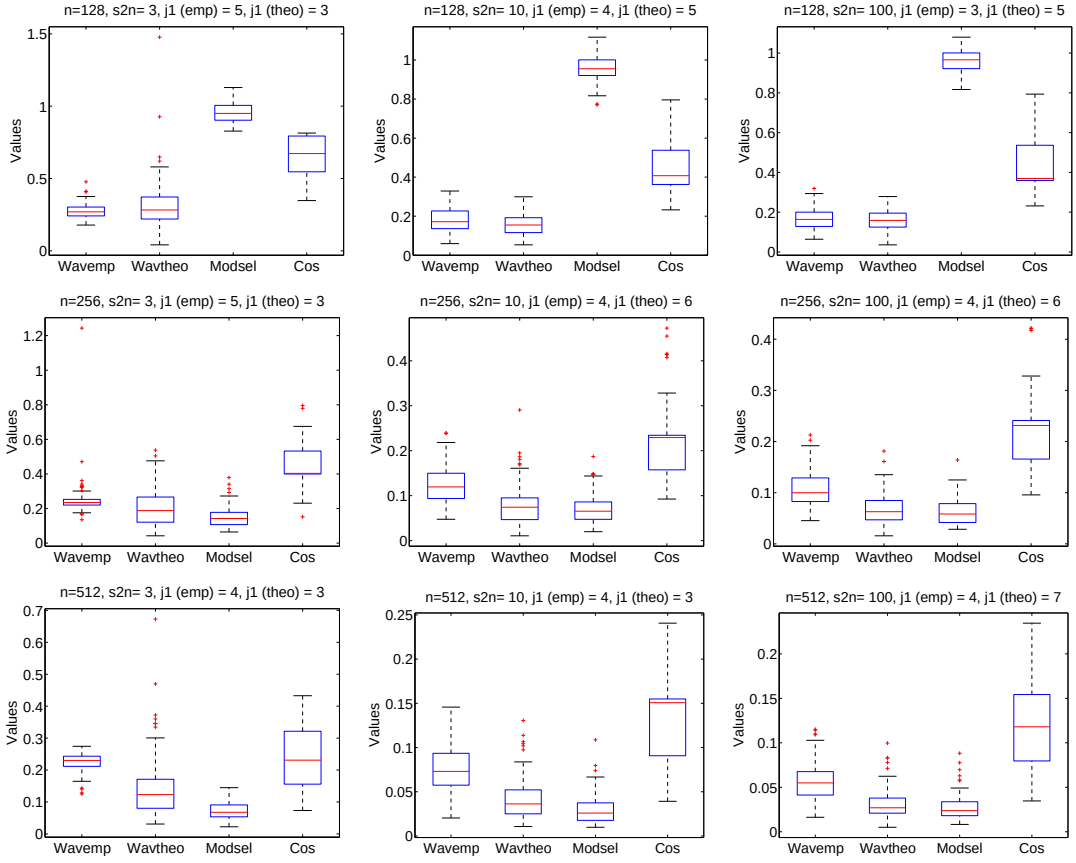


Figure 6.7: **MixtGaussian distribution**: graphical display (boxplots) of the MSE with 100 repetitions for each method and all combination of the factors n and $s2n$.

A Appendix

The estimation of density function based on information projection has been studied by Barron and Sheu (1991). To apply this method in our context of density deconvolution using Meyer wavelets, we need to recall and to adapt a set of results that are useful to prove the optimality of our estimator.

The first lemma derives a Pythagorian-like identity for the Kullback-Leibler divergence onto $\mathcal{E}_j$. This result is proved in Csiszár (1975).

Lemma A.1 *Let $\alpha \in \mathbb{R}^{2^j}$. Assume that there exists some $\theta(\alpha) \in \mathbb{R}^{2^j}$ such that for all $|\lambda| < j$:*

$$\langle f_{j,\theta(\alpha)}, \psi_\lambda \rangle = \alpha_\lambda.$$

Then, for any density function $f \in L^2([0,1])$ such that $\langle f, \psi_\lambda \rangle = \alpha_\lambda$ and for all $\theta \in \mathbb{R}^{2^j}$, the identity

$$\Delta(f; f_{j,\theta}) = \Delta(f; f_{j,\theta(\alpha)}) + \Delta(f_{j,\theta(\alpha)}, f_{j,\theta}).$$

holds true.

Note that the divergence $\Delta(f;g)$ is stricly positive, unless $f = g$ almost everywhere. Therefore, the Lemma of Csiszár (1975) implies that $\theta(\alpha)$ (if it exists) uniquely minimizes $\Delta(f; f_{j,\theta})$ for $\theta \in \mathbb{R}^{2^j}$.

The next lemma is a key result which gives sufficient conditions for the existence of the vector $\theta(\alpha)$ as defined in Lemma A.1. This lemma also relates distances between the densities in the parametric family to distance between the corresponding wavelet coefficients. Its proof relies upon a series of lemmas on bounds within exponential families for the Kullback-Leibler distance and can be found in Barron and Sheu (1991, Lemma 5).

Lemma A.2 *Let $\theta_0 \in \mathbb{R}^{2^j}$, $\alpha_{0,\lambda} = \langle f_{j,\theta_0}, \psi_\lambda \rangle$ and $\alpha \in \mathbb{R}^{2^j}$ a given vector. Let $b = \exp(\| \log(f_{j,\theta_0}) \|_\infty)$ and $e = \exp(1)$. If $\|\alpha - \alpha_0\|_2 \leqslant \frac{1}{2ebA_j}$, then the solution $\theta(\alpha)$ to*

$$\langle f_{j,\theta(\alpha)}, \psi_\lambda \rangle = \alpha_\lambda \text{ for all } |\lambda| < j$$

exists and satisfies:

$$\|\theta(\alpha) - \theta_0\|_2 \leqslant 2eb\|\alpha - \alpha_0\|_2 \tag{A.1}$$

$$\| \log(\frac{f_{j,\theta(\alpha_0)}}{f_{j,\theta(\alpha)}}) \|_\infty \leqslant 2ebA_j\|\alpha - \alpha_0\|_2 \tag{A.2}$$

$$\Delta(f_{j,\theta(\alpha_0)}; f_{j,\theta(\alpha)}) \leqslant 2eb\|\alpha - \alpha_0\|_2^2. \tag{A.3}$$

We conclude the appendix by a set of technical lemmas that are needed for the proof of our main results. These lemmas are an adaptation of similar results in Barron and Sheu (1991) or Antoniadis and Bigot (2006) to the case of periodoc, Meyer wavelets on $L^2[0,1]$. We start with some definitions. For $f \in F^s_{p,q}(M)$, let $g = \log_e(f)$ and define

$$D_j = \|g - P_j g\|_{L^2} \qquad \text{and} \qquad \gamma_j = \|g - P_j g\|_\infty.$$

The scaling Meyer functions $(\phi_\lambda)_{|\lambda|=j}$ span a finite dimensional space V_j within a multiresolution hierarchy $V_0 \subset V_1 \subset \ldots \subset L^2([0,1])$, such that $dim(V_j) = 2^j$ (see e.g. Meyer, 1992). In the following results, we use the inequalities $\|\phi_\lambda\|_\infty = \|\phi\|_\infty 2^{|\lambda|/2}$ and $\|\psi_\lambda\|_\infty = \|\psi\|_\infty 2^{|\lambda|/2}$, and assume that there exists some constant $A_j < \infty$ such that for all $v \in V_j$:

$$\|v\|_\infty \leqslant A_j \|v\|_{L^2}.$$

In the following, C denotes a generic constant whose value may change from line to line.

Lemma A.3 *Let $v \in V_j$, then $\|v\|_\infty \leqslant C 2^j \|v\|_{L^2}$.*

PROOF: Let $v = \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda$. By the Cauchy-Schwartz inequality and by the fact that $\|\psi_\lambda\|_\infty \leqslant C 2^{j/2}$, we obtain that uniformly in $x \in [0,1]$

$$|v(x)|^2 \leqslant \sum_{|\lambda|=j} |\psi_\lambda(x)|^2 \sum_{|\lambda|=j} |\beta_\lambda|^2 \leqslant C 2^{2j} \|\beta_j\|_2^2$$

which establishes the result. □

Lemma A.4 *Assume that $f \in F^s_{p,q}(M)$ with $p \leqslant 2$. If $s > 1/p + 1/2$, then there exists a constant M_1 such that*

$$0 < \frac{1}{M_1} \leqslant f \leqslant M_1 < \infty.$$

PROOF: Let $g = \log(f) = \sum_{j=-1}^{\infty} \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda$. Since $\|g\|_{B^s_{p,q}} \leqslant M$, we can write

$$\|\beta_j\|_p^p = \sum_{|\lambda|=j} |\beta_\lambda|^p \leqslant M 2^{-jps'},$$

where $s' = s + (1/2 - 1/p)$. As $p \leqslant 2$, we also get

$$\|\beta_j\|_2 \leqslant \|\beta_j\|_p \leqslant C 2^{-js'}. \tag{A.4}$$

Therefore, Lemma A.3 implies

$$\|g\|_\infty \leqslant \sum_{j=-1}^{\infty} \| \sum_{|\lambda|=j} \beta_\lambda \psi_\lambda \|_\infty \leqslant \sum_{j=0}^{\infty} C 2^j \|\beta_j\|_2 \leqslant C \sum_{j=0}^{\infty} 2^{j(1-s')} \leqslant C \sum_{j=0}^{\infty} 2^{-j(s-1/p-1/2)}.$$

Since $s > 1/p + 1/2$, $\sum_{j=0}^{\infty} 2^{-j(s-1/p-1/2)} < \infty$ and therefore there exists some constant $M_1 > 1$ such that $\|g\|_\infty = \|\log f\|_\infty \leqslant \log M_1$. □

The next lemma derive bounds for A_j, D_j and γ_j.

Lemma A.5 *The inequality*

$$A_j \leqslant C2^j$$

holds true. Moreover, assume that $f \in F^s_{p,q}(M)$ with $p \leqslant 2$. If $s > 1/p + 1/2$, then

$$\begin{aligned} D_j &\leqslant C2^{-j(s+1/2-1/p)} \\ \gamma_j &\leqslant C2^{-j(s-1/p-1/2)} \end{aligned}$$

PROOF: The result for A_j immediately follows from Lemma A.3. Note that from equation (A.4),

$$D_j^2 = \sum_{j' \geqslant j} \|\beta_{j'}\|_2^2 \leqslant C \sum_{j' \geqslant j} 2^{-2j'(s+1/2-1/p)} = \mathcal{O}(2^{-2j(s+1/2-1/p)}).$$

By definition, $\gamma_j = \|g - P_j g\|_\infty \leqslant A_j D_j \leqslant C2^{-j(s-1/p-1/2)}$, which completes the proof. □

The following lemma controls the mean square error for $\hat{\alpha}_{n,\lambda} - \alpha_\lambda$ 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 $\alpha_\lambda = \sum_l \frac{\psi_l^\lambda}{f_l^\epsilon} f_l^Y$.

Lemma A.6 *Assume that the Fourier coefficients of f^Y are such that $|f_l^Y| \leqslant C|l|^{-u}$ with $u > 1$. Then,*

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

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 the fact that under Assumption 1.1, $|f_\ell^\epsilon| \sim |\ell|^{-\nu}$, that $\psi_\ell^\lambda \leqslant C2^{-|\lambda|/2}$ and the independence of the Y_i's, we get the bound

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

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. □

Lemma A.7 *Let $f^X \in B^s_{p,q}[0,1]$, f^ϵ such that $|f_\ell^\epsilon| \sim |\ell|^{-\nu}$ for $\nu > 1/2$ (ordinary smooth error) and $f^Y = f^X \star f^\epsilon$. Consider the sequence $d_\lambda = \sum_\ell f_\ell^Y \psi_\ell^\lambda$ defined with Meyer wavelets ψ_λ. Then there exists a function of $B^{s+\nu}_{p,q}[0,1]$ with wavelet coefficients d_λ.*

PROOF: Consider the function $f^{\epsilon\star}$ in $L^2[0,1]$ such that $f_\ell^{\epsilon\star} = f_\ell^\epsilon$ for all ℓ. This function exists because, using that $f_\ell^\epsilon \sim |\ell|^{-\nu}$, the norm $\|f^{\epsilon\star}\|_{L^2[0,1]} = \sum_\ell (f_\ell^{\epsilon\star})^2$ is finite provided that $\nu > 1/2$. Now, consider the function $f^{Y\star} = f^X \star f^{\epsilon\star}$ in $L^2[0,1]$. By construction, we have that $\sum_\ell \psi_\ell^\lambda f_\ell^Y = \langle f^{Y\star}, \psi^\lambda \rangle_{L^2[0,1]}$ and $f^{Y\star} \in B_{p,q}^{s+\nu}[0,1]$ because $f^X \in B_{p,q}^s[0,1]$ and $f_\ell^{\epsilon\star} \sim |\ell|^{-\nu}$. □

References

Antoniadis, A. and Bigot, J. (2006). Poisson inverse problems. *Ann. Statist.* (to appear)

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. (`http://www-stat.stanford.edu/software/wavelab`)

Carrasco, M. and Florens, J.-P. (2002). *Sepctral method for deconvolving a density* (Working Paper No. 138). Université de Toulouse I: IDEI.

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

Cavalier, L. and Koo, J.-Y. (2002). Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. *IEEE Trans. Information Theory*, *48*, 2794–2802.

Cohen, A., DeVore, R. and Hochmuth, R. (2000). Restricted nonlinear approximation. *Constr. Approx.*, *16*, 85–113.

Comte, F., Rozenholc, Y. and Taupin, M.-L. (2006a). Finite sample penalization in adaptive density deconvolution. *J. Stat. Comput. Simul.*, to appear.

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

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., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B*, *57*, 301–369.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, *24*, 508–539.

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.

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.

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

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

Kosarev, E., Shul'man, A., Tarasov, M. and Lindstroem, T. (2003). Deconvolution problems and superresolution in Hilbert-transform spectroscopy based on a.c. Josephson effect. *Comput. Phys. Comm.*, *151*, 171–186.

Masry, E. (2003). Deconvolving multivariate kernel density estimated from contaminated associated observations. *IEEE Trans. Inform. Theory*, *49*, 2941–2952.

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.

Postel-Vinay, F. and Robin, J.-M. (2002). Equilibrium wage dispersion with worker and employer heterogeneity. *Econometrica*, *70*, 2295–2350.