Nonparametric Density Estimation using Wavelets

Marina Vannucci [*]
Department of Statistics, Texas A&M University

Discussion Paper 95-26, ISDS, Duke University, USA.

October 1995, first revision April 1997, last revision September 1998

Abstract

Here the problem of density estimation using wavelets is considered. Nonparametric wavelet density estimators have recently been proposed and seem to outperform classical estimators in representing discontinuities and local oscillations. The purpose of this paper is to give a review of different types of wavelet density estimators proposed in the literature. Properties, comparisons with classical estimators and applications are stressed. Multivariate extensions are considered. Performances of wavelet estimators are analyzed using a family of normal mixture densities and the Old Faithful Geyser dataset.

Key words and phrases: Nonparametric Density Estimation, Wavelets.

AMS Subject Classification: 62G07, 42A06.

1 Introduction

In nonparametric theory, density estimation is perhaps one of the most investigated topics. Let $X_1, \cdots, X_n$ be a sample of size n from an unknown probability density function f. The purpose is to estimate f without any assumption on its form. In this paper the attention will be focused on density functions which have compact support and belong to the space $L_2(\mathbb{R})$ of

[*] *Correspondence to:* Marina Vannucci, Dept. of Stats, Texas A&M University, College Station, TX, 77843-3143, USA. *mvannucci@stat.tamu.edu*

square integrable functions. Different types of nonparametric density estimators have been proposed in the literature. A comprehensive class is given by the Delta sequence estimators which includes as special cases the most commonly used density estimators, such as kernel, histogram and orthogonal series estimators. A different class is given by the penalized maximum likelihood estimators, such as spline estimators. Some of the main references for density estimation are Devroye and Györfi (1985), Prakasa-Rao (1983), Eubank (1988), Scott (1992) and Tapia and Thompson (1978). Useful applications of nonparametric density estimators can be found in the context of data analysis, i.e. discriminant analysis, cluster analysis, or the estimation of density functionals, such as the hazard rate in survival analysis.

Recently, nonparametric density estimators have been constructed using wavelets, that is, families of basis functions that have quite interesting properties, such as localization in space as well as in frequency. Wavelet estimators seem to outperform classical estimators in representing discontinuities and local oscillations. The first results about wavelet estimators can be found in Doukhan and Léon (1990), Kerkyacharian and Picard (1992) and Walter (1992).

This article gives a systematic survey of the contributions given in the literature and provides an easy access to the main ideas of wavelet density estimation for nonexperts. The purpose is twofold: first, an overview of different types of wavelet density estimators proposed in the literature is given, stressing relations, advantages and/or disadvantages of wavelet estimators versus more classical density estimators. Wavelet estimators are classified as global estimators that gain their strength from the local properties of the wavelet functions. Second, wavelet estimators are tested using a family of normal mixture densities previously proposed by Marron and Wand (1992) as a very broad class of densities which represent different and challenging problems for density estimation. Our simulation study shows that wavelet estimators have potential in estimating complicated densities and highlights the differences between linear and thresholded estimators. Linear estimators provide good estimates of smooth densities, while non-linear (or thresholded) estimators allow for discontinuities.

The paper presents also some extensions to the multivariate case.

Given the vast literature on classical density estimators, we have chosen not to review those methods. Main references are mentioned above. A comprehensive overview is given in Izenman (1991). We provide, instead, in Section 2, a general description of orthonormal wavelet bases, focusing on the mathematical properties which are essential to the construction of the wavelet density estimators. Section 3 describes the two different types of wavelet density estimators proposed in the literature, i.e. linear and thresholded. Connections to classical density estimators are provided and possible multivariate extensions are explored. Then, the attention is focused on practical problems. Section 4 explains how to avoid estimators which take on negative values. Section 5 is related to parameter choices involved in wavelet estimation. Linear estimators simply depend on the number of coefficients to be included while thresholded estimators require also the choice of thresholding parameters. Proposals are reviewed. Section 6 deals with the problem of choosing the translation parameter of the scaling and wavelet functions involved in the estimators. Section 7 provides insights on wavelet estimators by exploring their performances on simulated normal mixture densities and on a bivariate dataset. Section 8 contains conclusions and remarks.

2 Wavelets

Wavelets, well established in the literature, have been successfully applied in different fields, such as signal and image processing, numerical analysis and geophysics. In statistics, amongst other applications, wavelets have been used to construct suitable nonparametric density estimators. Wavelet bases have several desirable properties that make nonparametric density wavelet estimators competitive. The following sections provide the mathematical framework which is necessary to fully understand how wavelet density estimators are constructed. For a general exposition of the wavelet theory see Daubechies (1992), Chui (1992) and Meyer (1992).

2.1 Multiresolution Analysis

Generally speaking, a wavelet basis in $L_2(\mathbb{R})$ is a collection of functions obtained as translations and dilations of a scaling function ϕ and a *mother wavelet* ψ. The function ϕ is constructed as a solution of the dilation equation $\phi(x) = \sqrt{2}\sum_l h_l \phi(2x - l)$ for a given set of filter coefficients h_l that satisfy suitable conditions. The function ψ is defined from ϕ as $\psi(x) = \sqrt{2}\sum_l g_l \phi(2x - l)$, with filter coefficients g_l often defined as $g_l = (-1)^l h_{1-l}$. The wavelet collection is obtained by translations and dilations as $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$. Wavelet collections are particularly useful to approximate other functions. As it will be shown later, scaling functions give a good approximation of smooth functions while wavelets are particularly useful when dealing with functions that have local fluctuations.

Mallat (1989) introduced orthonormal wavelet bases in the general context of the *multiresolution analysis* (MRA) as a decomposition of $L_2(\mathbb{R})$ into a sequence of linear closed subspaces $\{V_j, j \in \mathbb{Z}\}$ such that

$$(i) \qquad V_j \subset V_{j+1}, \quad j \in \mathbb{Z}$$

$$(ii) \qquad \cap_j V_j = \{0\} \quad \overline{\cup_j V_j} = L_2(\mathbb{R})$$

$$(iii) \;\; f(x) \in V_j \iff f(2x) \in V_{j-1}, \;\; f(x) \in V_j \Rightarrow f(x+k) \in V_j, k \in \mathbb{Z}.$$

Here the dilation function ϕ is such that the family $\{\phi(x - k), k \in \mathbb{Z}\}$ is an orthonormal basis for V_0. The family $\{\phi_{j,k}(x), k \in \mathbb{Z}\}$ is then an orthonormal basis for V_j. If W_j indicates the orthogonal complement of V_j in V_{j+1}, i.e. $V_j \oplus W_j = V_{j+1}$, then $L_2(\mathbb{R})$ can be decomposed as

$$L_2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j \tag{1}$$

or, equivalently, as

$$L_2(\mathbb{R}) = V_{j_0} \oplus \bigoplus_{j \geq j_0} W_j. \tag{2}$$

The family of wavelets $\{\psi_{j,k}(x), j, k \in \mathbb{Z}\}$ forms an orthonormal basis in $L_2(\mathbb{R})$.

2.2 Wavelet Series

A wavelet collection can be used to represent other functions as follows. Let f be any L_2 function. Equation (1) says that f can be represented by a wavelet series as

$$f(x) = \sum_{j,k \in \mathbb{Z}} d_{j,k} \psi_{j,k}(x), \tag{3}$$

with wavelet coefficients defined as $d_{j,k} = \int f(x)\psi_{j,k}(x)dx$. When truncated at a fixed scale $\bar{\jmath}$ the wavelet expansion (3) is an orthogonal projection $P_{\bar{\jmath}}$ of f in the subspace $V_{\bar{\jmath}}$ of $L_2(\mathbb{R})$ and can be expressed in terms of scaling functions as

$$P_{\bar{\jmath}} f(x) = \sum_{k \in \mathbb{Z}} c_{\bar{\jmath},k} \phi_{\bar{\jmath},k}(x) \tag{4}$$

with scaling coefficients defined as $c_{\bar{\jmath},k} = \int f(x)\phi_{\bar{\jmath},k}(x)dx$. Equivalently, equation (2) says that any $f \in L_2$ can be represented also as

$$f(x) = \sum_{k \in \mathbb{Z}} c_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{\infty} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k}(x). \tag{5}$$

Equation (5) can be viewed as an approximation of f at the scale j_0 plus a set of extra information (*details* in the wavelet terminology) about f at finer scales.

Wavelet expansions are often compared with the more classical Fourier representations. There, an orthogonal basis is constructed using sine and cosine functions and a Fourier series is defined. Wavelet bases seem to be more appealing than classical bases, mainly because of their localization properties. Sine and cosine functions are localized in frequency but not in time. Wavelets, on the contrary, can be localized both in time and in frequency/scale. Intuitively, this property enables the wavelet series to describe local characteristics of a function in a parsimonious way.

2.3 Daubechies' Wavelets

There are several families of wavelets, proposed by different authors. We discuss those developed by Daubechies (1988), which are extensively used in statistical applications. Wavelets from these families are orthogonal and compactly supported, they possess different degrees of smoothness and have the maximum number of vanishing moments[1] for a given smoothness. These properties are desirable when representing functions through a wavelet series. As previously pointed out, compact supports are useful to describe local characteristics that change rapidly in time. Moreover, a large number of vanishing moments would lead to high compressibility, since the fine scale wavelet coefficients will be essentially zero where the function is smooth.

Daubechies developed the algorithm for the construction of these wavelets using Mallat's multiresolution analysis and presented them by giving the filter coefficients h_l such that they satisfy orthogonality and the minimum vanishing moments property. The best known of Daubechies' wavelets are the *minimum phase*, to be indicated by the symbol *Daub#N*, and the *Coiflet*, to be indicated by the symbol $Coiflet\#N$. In both cases, N is the number of vanishing moments of the functions. The family *Daub#1* is the well known Haar basis (Haar (1910)).

3 Wavelet Density Estimation

Orthonormal wavelet bases can be used to construct nonparametric wavelet density estimators as follows. Let $X_1, \cdots, X_n$ be an *i.i.d.* sample from $X \sim f$. The wavelet representation of f is given by equation (3) or, equivalently, by (5). Since f is a density function we can write

$$d_{j,k} = \int f(x)\psi_{j,k}(x)dx = E[\psi_{j,k}(X)] \tag{6}$$

and similarly for the scaling coefficients $c_{j,k}$. Thus, a wavelet estimator of f can be written simply by truncating the wavelet expansion of f and using

$$\hat{c}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\phi_{j,k}(X_i), \quad \hat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(X_i) \tag{7}$$

as unbiased estimates of the coefficients.

Given their construction, wavelet estimators can be classified as orthogonal series estimators, a general class of nonparametric density estimators

[1] A function f has N vanishing moments if $\int t^q f(t)dt = 0, \quad q = 0, 1, \ldots, N-1$

introduced in the literature by Čencov (1962). Different orthogonal systems in $L_2(\mathbb{R})$ have been used to construct orthogonal series estimators. Among others, we recall the Fourier system for a function f with compact support, and the Haar system in $[0, 1]$. Results about the consistency of Hermite and Fourier series estimators can be found in Izenman (1991).

What do wavelet estimators have to offer in comparison with more classical estimators of the same type? A major drawback of classical series estimators is that they appear to be poor in estimating local properties of the density. This is due to the fact that orthogonal systems, like the Fourier one, have poor time/frequency localization properties. On the contrary, as previously pointed out, wavelets are localized both in time and in frequency. This makes wavelet estimators well able to capture local features. Examples in Section 7 will provide more insight.

Another common way of comparing nonparametric estimators is through asymptotic analysis. Rates of convergence of wavelet estimators have been studied by several authors. In the next sections we describe the two different types of wavelet estimators, linear and thresholded, proposed in the literature, and give brief summaries of their statistical properties in comparison with classical estimators.

3.1 Linear Estimators

When truncating the wavelet series in the form (3) we obtain a linear estimator of the type

$$\hat{f}(x) = \sum_k \hat{c}_{\bar{\jmath},k} \phi_{\bar{\jmath},k}(x), \tag{8}$$

with sample coefficients $\hat{c}_{\bar{\jmath},k}$ defined as in (7). The estimator is an orthogonal projection of f onto the subspace $V_{\bar{\jmath}}$ of $L_2(\mathbb{R})$. The performance of the wavelet estimator (8) depends on the choice of $\bar{\jmath}$. This parameter adjusts the trade-off between bias and variance of the estimator. See Section 5 for suggested methods.

When using the Haar family the wavelet estimator (8) is the usual histogram with bins centred on $2^{-\bar{\jmath}}k$ and having bin widths of $2^{-\bar{\jmath}}$, k being an integer. The scaling function ϕ is, in fact, the indicator function of the interval $[0,1)$ and, thus, $\hat{c}_{\bar{J},k} = \frac{2^{\bar{J}/2}}{n} \#\{X_i \in [k2^{-\bar{J}},(k+1)2^{-\bar{J}})\}$. The wavelet estimator becomes then

$$\hat{f}(x) = \frac{2^{\bar{J}}}{n} \sum_k N_k I_{T_k}(x), \tag{9}$$

where $T_k = [k2^{-\bar{J}},(k+1)2^{-\bar{J}})$ and N_k is the number of sample values falling into T_k.

Linear wavelet estimators have been studied by several authors, Walter (1992, 1994), Kerkyacharian and Picard (1992, 1993), Antoniadis and Carmona, (1991). Walter (1994) established mean squared error convergence results of wavelet estimators in the case of continuous densities. He also proved that the rate of convergence is faster for smoother functions (for example, functions belonging to Sobolev spaces). These properties are not shared by classical orthogonal series estimators, such as the Fourier series estimators or the Hermite series estimators. Walter (1992) showed that the wavelet estimator can be written in terms of a *reproducing kernel* (see Aronszajn (1950) for a definition) as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} q_{\bar{J}}(X_i, x) \tag{10}$$

with kernel $q_j(y,x) = \sum_k \phi_{j,k}(y)\phi_{j,k}(x)$. He proved that the sequence $\{q_{\bar{J}}(y,x)\}$ is a delta sequence[2]. This result is important in relating wavelet estimators to classical ones. It implies, in fact, that linear wavelet estimators can be classified as delta sequence estimators, a general class introduced by Walter and Blum (1979) that includes as special cases the common density estimators, such as histograms, kernel and orthogonal series estimators.

Kerkyacharian and Picard (1992, 1993) established rates of convergence of the linear wavelet estimators when f is in a Besov space $B_{s,p,q}$. For a definition and properties of Besov spaces see Bergh and Löfström (1976).

Wu (1994, 1996) suggested wavelet estimators obtained as a convex combination of M linear wavelet estimators. The sample is decomposed into two

[2] A delta sequence $\delta_\lambda(x,y)$ is a sequence of bounded measurable functions on $J \times J$, with J an open interval in $\mathbb{R}$, such that $\int_J \delta_\lambda(x,y)\phi(y)dy \to \phi(x)$ as λ goes to infinity for every $x \in J$ and every infinitely differentiable function ϕ on J.

disjoint subsamples $Y^\star$ and $Z^\star$ (of size N_1 and N_2 respectively) and a *multiscale wavelet estimator* is defined as

$$\hat{f}(x,t) = \sum_{m=1}^{M} \lambda_m(Y^\star,t)\hat{f}_{N_2}^{(m)}(Z^\star,t) \tag{11}$$

where $\hat{f}_{N_2}^{(m)}$, $m = 1, \cdots, M$, are M linear wavelet estimators calculated using $Z^\star$ and where $\lambda_m(Y^\star,t),\ m = 1,\cdots,M$ are non-negative random variables that constitute a random partition of the unity. The author proved some statistical properties of these estimators, such as asymptotic unbiasedness and normality, and presented applications on mixtures of normal and double exponential densities.

3.2 The Donoho *et al.* Thresholded Estimators

When dealing with densities with spatially varying (or heterogeneous) smoothness properties, thresholding of wavelet coefficients allows wavelet estimators to automatically adapt to the local smoothness of the density. Donoho *et al.*, (1996), suggested a thresholded estimator obtained by truncating the wavelet series in the form (5) and then thresholding the smallest empirical coefficients $\hat{d}_{j,k}$ as

$$\hat{f}(x) = \sum_k \hat{c}_{j_0,k}\phi_{j_0,k}(x) + \sum_{j_0}^{j_1}\sum_k \tilde{d}_{j,k}\psi_{j,k}(x) \tag{12}$$

with $\tilde{d}_{j,k}$ thresholded version of $\hat{d}_{j,k}$. This estimator can be seen as a coarse approximation of f at scale j_0, plus some extra information on f that is added in order to improve the estimate. The linear part of the estimator describes "average", or low-frequency, features of the density while the non-linear one adapts to local fluctuations like discontinuities or high-frequency oscillations. This construction can be seen as a special type of tapering (see Section 5). The purpose is to exclude from the series those terms that cannot be estimated in a feasible way given the large number of zero terms in $\hat{d}_{j,k}$, while the thresholding selects the extra terms that allow for discontinuities.

Two different types of thresholding can be performed: the choice $\tilde{d}_{j,k} = \hat{d}_{j,k}I(|\hat{d}_{j,k}| > \delta)$ leads to a "hard" thresholding that sets to zero coefficients with absolute value smaller than δ, while the choice $\tilde{d}_{j,k} = sign(\hat{d}_{j,k})(|\hat{d}_{j,k}| - \delta)_+$ leads to a "soft" thresholding that, additionally, shrinks the larger coefficients towards zero. The trade-off between bias and variance is now adjusted by the parameters j_1 and δ.

From a mathematical point of view, heterogeneous smoothness is modeled by assuming that the density belongs to specific class of Besov spaces. Donoho *et al.* gave a lower bound result providing the optimal minimax rates of convergence that a single estimator can achieve over a large variety of Besov spaces. They proved that, when f is in a Besov space $B_{s,p,q}$ and the global error measure adopted is $L_\pi = E\|\hat{f} - f\|^2_{s,\pi}$ with $\| \cdot \|_{s,\pi}$ the norm in the Sobolev space W^s_π, the thresholded estimator achieves the asymptotic optimal minimax rate, outperforming the linear estimator (8) in the case $\pi > p$. They suggested robust choices of the truncation parameters to achieve the optimal rates.

Donoho and Johnstone proposed also an alternative method for density estimation, briefly described in Donoho (1993). This method uses a variance stabilizing transformation to reduce the density estimation problem into a regression model. Standard wavelet shrinkage techniques (see, for example, Donoho *et al.* (1995b)) can then be applied.

3.3 The Hall and Patil Thresholded Estimators

Hall and Patil (1995a, 1995b, 1996) proposed a different class of thresholded wavelet estimators by employing a second smoothing parameter. These authors defined the wavelet collection as

$$\phi_k(x) = p^{1/2}\phi(px - k) \tag{13}$$

$$\psi_{j,k}(x) = p_j^{1/2}\psi(p_j x - k) \tag{14}$$

where ϕ and ψ are the usual scaling function and mother wavelet, as defined in Section 2.1. Given $p > 0$, $k \in \mathbb{Z}$, $j > 0$ and $p_j = p2^j$, the functions ϕ_k and $\psi_{j,k}$ are orthogonal and a generalized Fourier expansion of f can be considered. The nonlinear (or thresholded) wavelet estimator proposed by the authors is then given by

$$\hat{f}(x) = \sum_k \hat{c}_k \phi_k(x) + \sum_0^{q-1} \sum_k \hat{d}_{j,k} w(\hat{d}_{j,k}/\delta) \psi_{j,k}(x) \tag{15}$$

with coefficients $\hat{c}_k = n^{-1} \sum_{i=1}^n \phi_k(X_i)$ and $\hat{d}_{j,k} = n^{-1} \sum_{i=1}^n \psi_{j,k}(X_i)$. The parameters q, the smoothing or truncation parameter, and δ, the threshold, are adjustable constants that need to be chosen. See Section 5 for suggested methods. The function w is the threshold or "weight" function. It can be chosen as $w(u) = I(|u| > 1)$, obtaining a "hard" thresholding, or as $w(u) = 0$ for $0 < u < c_1$, $w(u) \in [0, 1]$ for $c_1 < u < c_2$ and $w(u) = 1$ for $u > c_2$, with $0 < c_1 < c_2 < +\infty$ constants, obtaining a "soft" thresholding.

The parameter p is an additional smoothing parameter. Notice that the nonlinear wavelet estimators proposed by Donoho *et al.* can be obtained by choosing $p = 2^{j_0}$. Hall and Patil showed that a theoretical advantage of allowing more flexibility in the choice of p is that, for specific parameter spaces, no logarithmic factors affect the convergence rates.

Hall and Patil (1995b, 1996) pointed out the analogy of the linear part of the proposed estimator with the classical generalized kernel estimators when considering p^{-1} as a bandwidth, and developed formulae for the variance, the bias and the mean squared error. They provided an asymptotic formula for the mean integrated squared error of the nonlinear wavelet estimator (15) showing that, unlike the kernel methods, wavelet estimators have good convergence rates even when the unknown density is only piecewise continuous. They proved also that nonlinear wavelet estimators are highly robust against oversmoothing, that is, against choosing p too small. Hall, McKay and Turlach (1996) studied the rates of convergence of wavelet estimators when the discontinuities of the density increase in number and decrease in size as the sample size grows. Essentially, wavelet estimators are consistent if the number of jumps multiplied by the size of the jumps is of a smaller order than the sample size (multiplied by a factor).

3.4 Multivariate Extensions

Only a small amount of work has been done in the multivariate case. Contributions have been given by Walter (1995), Vannucci (1996) and Neumann

(1996).

Constructing wavelet estimators is a trivial extension of the univariate case if we consider that in $L_2(\mathbb{R}^d)$ a multiresolution analysis is defined as tensor products of d one-dimensional MRA. A wavelet collection can be constructed as translations and dilations of a scaling function $\phi(\mathbf{x}) = \phi(x_1)\cdots\phi(x_d)$, where $\phi(x)$ is a one-dimensional scaling function, and of $2^d - 1$ wavelets $\psi^{(l)}(\mathbf{x})$ defined as

$$\psi^{(l)}(\mathbf{x}) = \prod_{i=1}^{d} \xi(x_i), \quad \text{with } \xi = \phi \text{ or } \psi, \text{ but not all } \xi = \phi \tag{16}$$

where ψ is the one-dimensional mother wavelet associated with the function ϕ.

Thus, given $\mathbf{X}_i, i = 1, \cdots, n$ a sample from $\mathbf{X}$ with probability density function f, linear and thresholded wavelet estimators can be constructed by truncating the wavelet expansion and defining sample coefficients as for the univariate case. See Vannucci (1996) for more details. Walter (1995) presented convergence results of linear estimators and showed how they can be expressed in terms of reproducing kernels. Properties of thresholded estimators need to be investigated. Neumann (1996) proposed an alternative construction of a multidimensional wavelet basis, which involves tensor products of one-dimensional MRA's based on different scale parameters. He mainly explored the estimation of noisy signals but mentioned also applications to density estimation.

4 Bona Fide Estimates

Except for the Haar family, wavelet estimators may take on some negative values in the tails of the distribution or in the part of support without data. This is a recurrent problem in nonparametric estimation and can be addressed in many different ways. We discuss two. A simple way is to truncate the estimate to its positive part and then re-normalize the truncation. Alternatively, one may estimate a transformed version of f, usually $\log f$ or $f^{1/2}$, and then transform back to have a non-negative estimate of f. The transformation $g = f^{1/2}$ was introduced by Good and Gaskins (1971) in the context of penalized maximum likelihood estimation to satisfy the nonnegativity constraint. This approach was criticized later by De Montricher, Tapia and Thompson (1975) who pointed out theoretical problems related to the uniqueness of the solution. The transformation $g = \log(f)$ was used by Leonard (1973) in a Bayesian approach to density estimation and by Silverman (1982) and Gu and Qiu (1993) in the context of penalized maximum likelihood estimation.

Pinheiro and Vidakovic (1997) applied wavelet estimators to the square root of the density. The wavelet representation of $\sqrt{f}$ involves wavelet coefficients defined as

$$d_{j,k} = \langle \sqrt{f}, \psi_{j,k} \rangle = \int \sqrt{f(x)}\psi_{j,k}(x)dx = \int \frac{\psi_{j,k}(x)}{\sqrt{f(x)}} f(x)dx \tag{17}$$

and similarly for the scaling coefficients. Thus, thresholded and linear estimators are constructed by using sample coefficients

$$\hat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \frac{\psi_{j,k}(X_i)}{\sqrt{\hat{f}_n(X_i)}} \quad \text{and} \quad \hat{c}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \frac{\phi_{j,k}(X_i)}{\sqrt{\hat{f}_n(X_i)}}. \tag{18}$$

In the equation (18) $\hat{f}_n$ is a pilot estimator of f that should be computationally simple and give sensible “weights” to ϕ_{jk}’s and ψ_{jk}’s. The simple choice

$$\hat{f}_n(X_i) = \#\{X_j \in (X_i - r, X_i + r)\}, \tag{19}$$

with radius $r \in \mathbb{R}^+$, produces good results. A nice feature of estimating the square root of a density is that, since $\int \hat{f}(x)dx = \|\sqrt{\hat{f}}\|^2$, re-normalization of the sample coefficients leads to a density estimate that integrates to one, that is to a *bona fide* estimate. The result follows by applying the Parseval identity to the normalized sample coefficients.

This approach can be easily generalized to the multivariate case using pilot estimators of the type

$$\hat{f}_n(\mathbf{X}_i) = \#\{\mathbf{X}_j \in B(\mathbf{X}_i, r), \quad j = 1, \cdots, n\}, \tag{20}$$

with $B(\mathbf{X}_i, r)$ a ball centered at $\mathbf{X}_i$ with radius r (see Vannucci (1996)).

Estimation of the square root of a density has been studied also by Penev and Dechevsky (1995). The proposed pilot estimator is expressed here through order statistics $X_{n,1} \leq X_{n,2} \leq \cdots \leq X_{n,n}$ as

$$\hat{f}_n(X_{n,i}) = \frac{C}{n(X_{n,i+1} - X_{n,i})}.$$

The authors prove that the estimator achieves asymptotically optimal minimax rates, as in Donoho *et al.* but on more restricted spaces.

5 Choice of Parameters

When using orthogonal series estimators a crucial issue is the choice of the truncation parameter, that is, of the number of terms to be included in the estimator. Truncation is needed not only for practical but also for theoretical reasons. The estimator may not be well defined for an infinite set of indices, because it has an infinite variance and it is not consistent in the integrated squared error sense, that is, $\|\hat{f} - f\|_2$ does not converge to zero. See Izenman (1991), page 214, for a nice discussion and related references. Standard practices are to *taper* the estimator by multiplying the empirical coefficients with symmetric weights that shrink them towards the origin or, more simply, to select a finite number of coefficients. Performances of the estimator strongly depend on this. A smaller bias but a greater variance can be achieved by including more terms. Generally, a small number of terms can lead to smooth estimators, losing important features of the density, while a large number to rough estimators, containing spurious artifacts. As an extreme case, the estimator can be close to a sum of delta functions centered on the observations.

When using wavelet estimators a number of parameter choices are required. Specifically, the linear wavelet estimator (8) depends only on the truncation parameter $\bar{\jmath}$. The Donoho *et al.* thresholded estimators require the choice of a low scale j_0 and a finer scale j_1. The Hall and Patil estimators (15) depend on the smoothing parameter p, the equivalent of a bandwidth in kernel estimation, and on the truncation parameter q. Clearly, when using thresholded estimators, the choice of the threshold δ needs to be addressed too.

Notice that, roughly speaking, the problem of choosing a suitable scale parameter is more important in the linear case; when using estimators (12) and (15), thresholding of wavelet coefficients may permit us to eliminate possible "redundancy" due to the choice of a large scale.

We now give a brief review of different proposals, starting with linear estimators. Examples in Section 7 will provide more insights and practical comparisons.

5.1 Linear Case

Walter (1994) suggested choosing the truncation parameter of (8) by minimizing the estimated Mean Integrated Squared Error (MISE). Let e_j indicate the MISE with $\hat{f}(x)$ depending on the scale j. Proceeding from one scale to the next coarser, the increment $e_j - e_{j-1}$ can be estimated. Specifically,

$$e_j - e_{j-1} = \frac{1}{n}\int 2^j q(2^j x, 2^j x) f(x) dx - \frac{n+1}{n}\sum_k d^2_{j-1,k}, \tag{21}$$

where the integral can be estimated by $\frac{1}{n} 2^j \sum_1^n q(2^j X_i, 2^j X_i)$. Moreover, when using the Haar family,

$$e_j - e_{j-1} = \frac{2^{j-1}}{n} - \frac{n+1}{n}\sum_k d^2_{j-1,k}. \tag{22}$$

Walter suggested starting from a given scale, which should be chosen not to be finer than that required to put every X_i into a different interval, and using the empirical wavelet coefficients to estimate the increments for the next coarser scale until the error increases by a large amount. The achieved scale will be the scale of interest.

Another possible procedure was proposed, again by Walter (1995), for the special case of wavelet estimators constructed using the Haar family. These estimators are, in fact, histograms. Scott (1992), among others, showed that histogram estimators are MISE consistent and that the MISE can be asymptotically minimized by choosing asymptotically optimal bin widths. The optimal bin width depends on the norm of f, so approximated values have to be chosen for practical purposes. Scott found that the approximate optimal bin width $\hat{h} = 3.5 s n^{-1/3}$, with s the sample standard deviation, works well for Gaussian samples and leads to oversmoothing otherwise. Thus, when using the Haar linear estimator, a starting point to achieve the finest scale of interest could be $j = 3^{-1}\log_2 n - 2 - \log_2 \sigma$, where σ can be estimated with the sample standard deviation. Since the Haar wavelets generate the worst approximating family, the "optimal" j for that case could be an upper bound for the Daubechies wavelet families $Daub\#N$, when $N \neq 1$.

Finally, we give a brief description of a method proposed by Vannucci and Vidakovic (1995) that involves the Fisher information functional as a roughness measure. Given f absolutely continuous, the Fisher information functional is defined as

$$F(f) = \int_{\mathbb{R}} \frac{1}{f(x)} \Big[\frac{d}{dx} f(x)\Big]^2 dx \tag{23}$$

and corresponds to the usual parametric definition in the case of density functions $f(\cdot - \theta)$ with θ a location parameter (see, for example, Cox and Hinkley (1974)). The roughness of a function f can be measured by the ease of discriminating the function from itself when it is subjected to a small translation. Fisher information is invariant over translations; thus, it can be used to measure the roughness of f. Fisher information was used by Good and Gaskins (1971) in the context of maximum likelihood estimation.

Using a wavelet representation of the derivative of the linear estimator, Vannucci and Vidakovic derived an estimate of the Fisher information of

$$\hat{f}(x) = \left(\sum_k \hat{c}_{\bar{\jmath},k}\phi_{\bar{\jmath},k}(x)\right)^2 . \tag{24}$$

This estimate depends on the truncation parameter $\bar{\jmath}$ and provides a roughness measure of wavelet estimators. The authors provided also an "optimal" roughness value for densities assuming zero values at the extremes of their (compact) support, and suggested choosing the value of $\bar{\jmath}$ that gives a wavelet estimator with estimated Fisher information close to this optimum.

5.2 Thresholded Case

We start with the thresholded wavelet estimators (12). The two parameters j_0 and j_1 and the threshold δ need to be chosen. Donoho *et al.* suggested selecting j_0 according to the regularity of the function f and j_1 as

$\lfloor \log_2 n - \log_2(\log n) \rfloor$. Moreover they proposed a scale-dependent threshold $\delta_j = C\sqrt{j/n}$ with C a suitably chosen constant.

Thresholded wavelet estimators (12) have been studied by other authors. Delyon and Juditsky (1993) suggested $j_1 = \lfloor \log_2 n - \log_2(\ln n) \rfloor$ and $\delta_j = Cn^{-1/2}(max(j - j_0, 0))^{1/2}$. A rather different approach was taken by Pinheiro and Vidakovic (1997). In order to choose the scale parameters they considered the wavelet *scalogram* of f,

$$\Pi_{f(j)} = \sum_k |\langle f, \psi_{jk} \rangle|^2 \tag{25}$$

that describes the distribution of the total energy of f (that is the L_2 norm) at each scale j. They showed that the scalogram "behaves well" for small values of j and increases exponentially for j large. Thus, given $[a, b]$ the sample interval, they suggested starting from a coarse scale j_0, that can be selected such that $supp(\phi_{j_0+1,k}) \subset [a, b] \subset supp(\phi_{j_0,k})$ for some k, and using the empirical scalogram $\Pi(j) = \sum_k |\hat{d}_{jk}|^2$ to select the maximum scale j_1 as the scale after which the scalogram begins its exponential growth. In choosing the thresholding parameter, the authors used the Lorentz curve of the energies of the empirical wavelet coefficients. Most of the total energy of the density is concentrated in a few large coefficients. Thus, only those with energy larger than κ times the average energy are retained.

We now focus on the Hall and Patil estimators (15). Hall and Patil (1996) gave necessary and sufficient conditions on the form of the threshold to achieve optimal rates in the case of continuous and piecewise-smooth functions. In the sequel we concentrate on r-times differentiable densities and on wavelet estimators (15) constructed using r-th order wavelets and the hard thresholding function $w(u) = I(|u| > \delta)$. Similar results apply in the soft thresholding case and can be easily generalized to the class of piecewise-smooth densities.

Given the simple choice $p_j = 2^j$ and assuming that $q_0 - C_1 \leq q \leq C_2(\log n)(\log 2)^{-1}$ with q_0 the integer part of $(2r+1)^{-1}(\log 2)^{-1}(\log n)$ and arbitrary $C_1, C_2 \in ((2r+1)^{-1}, 1)$, three different thresholding rules can be used:

- $\delta_j = 0$ for $0 \leq j \leq q_1$ and $\delta_j = C_3 n^{-1/2}(j - q_1)^{1/2}$ for $j > q_1$

- $\delta_j = 0$ for $0 \le j \le q_1$ and $\delta_j = C_3 n^{-1/2} (\log n)^{1/2}$ for $j > q_1$

where $|q_1 - q_0| \le C_4,\ C_4 > 0$ arbitrary and C_3 sufficiently large.

- $\delta_j = C_5 (\log n)^{1/2}$ for all j.

The first two rules produce optimal mean squared convergence rates uniformly over the class of r-times differentiable densities. The third one, however, produces an oversmoothed estimator with squared bias contribution to the MISE greater than the variance. In the more general case $p_j = p2^j$ with $p \in [1, 2)$, the first two thresholding rules can be written as

- $\delta_j = 0$ for $0 \le j \le q_1$ and $\delta_j = C n^{-1/2} (j - q_1)^{1/2}$ for $j > q_1$
- $\delta_j = 0$ for $0 \le j \le q_1$ and $\delta_j = C n^{-1/2} (\log n)^{1/2}$ for $j > q_1$.

For both the rules $C > (2(\log 2) \sup f)^{1/2}$ is adequate. The parameters that allow for smoothing (that is, that adjust the trade-off between squared bias and variance) are C and p_{q_1} using the first rule, and only p_{q_1} using the second. Hall and Patil suggested also spatially adaptive rules (that is, with the parameter δ depending not only on j but also on k) that permit local smoothing.

5.3 Multivariate Extensions

In addressing the multivariate linear case, Tribouley (1995) suggested the use of a cross validation procedure by choosing the scale $\bar{\jmath}$ that minimizes

$$CV(j) = \sum_{\mathbf{k}} \Big[\frac{2}{n(n-1)} \sum_{i=1}^{n} (\phi_{j,\mathbf{k}}(\mathbf{X}_i))^2 - \frac{n+1}{n^2(n-1)} \Big(\sum_{i=1}^{n} \phi_{j,\mathbf{k}}(\mathbf{X}_i) \Big)^2 \Big]. \quad (26)$$

This criterion is equivalent to minimizing either the integrated squared error or the integrated mean squared error.

Multivariate generalizations of some of the methods described above, for the linear and thresholded case, may be possible. We notice, however, that the (multivariate) normal histogram rule suggested by Scott (1992) selects different bin widths for each dimension and therefore it cannot be applied to the linear multivariate wavelet estimator. Such an estimator is still a histogram, in the Haar wavelet case, but with the same bin width, given by $2^{-\bar{\jmath}/d}$, for each dimension.

6 Practical Computations

Here we address more practical issues that arise in applications. One is finding the range of the translation parameter k for $\phi_{j,k}$, ϕ_k and $\psi_{j,k}$ in (8), (12) or (15). Using wavelet functions with compact support ensures a finite range of values. Considering, for example, the minimum phase Daubechies wavelets: the support of ϕ is $[0, 2N-1]$ and the support of $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k)$ is $[k/2^j, (2N-1+k)/2^j]$. Thus, given $[a, b]$, the sample range, one has to calculate only the values of k for which the support of the corresponding functions $\phi_{j,k}$ intersects $[a, b]$. Straightforward calculations give

$$\lceil a2^j \rceil - 2N + 1 \leq k \leq \lfloor b2^j \rfloor. \tag{27}$$

For example, for $[a, b] = [0, 1]$ and $j = 0$ the range of k is $-2N + 1 \leq k \leq 1$. Similarly, the support of $\phi_k(x) = p^{1/2}\phi(px - k)$ is $[k/p, (2N-1+k)/p]$ and the values of k for which the support of ϕ_k intersects $[a, b]$ are

$$\lceil ap \rceil - 2N + 1 \leq k \leq \lfloor bp \rfloor. \tag{28}$$

Accordingly, the support of ψ is $[1-N, N]$ and wavelets defined as $\psi_{j,k} = 2^{j/2}\psi(2^j x - k)$ can be chosen such that

$$\lceil a2^j \rceil - N \leq k \leq \lfloor b2^j \rfloor + N - 1 \tag{29}$$

while those defined as $\psi_{j,k} = p_j^{1/2}\psi(p_j x - k)$ such that

$$\lceil ap_j \rceil - N \leq k \leq \lfloor bp_j \rfloor + N - 1. \tag{30}$$

The same procedure can be repeated with $Coiflet$ wavelets.

Clearly, similar calculations can be done in the multivariate case. Let $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$ be the sample range. When using minimum phase Daubechies wavelets, the wavelet estimators will involve scaling functions $\phi_{j,\mathbf{k}}$ with $\mathbf{k} = (k_1, \cdots, k_d)$ such that

$$\lceil a_i 2^j \rceil - 2N + 1 \leq k_i \leq \lfloor b_i 2^j \rfloor, \quad i = 1, \cdots, d \tag{31}$$

and wavelet functions $\psi_{j,\mathbf{k}}$ with $\mathbf{k} = (k_1, \cdots, k_d)$ such that

$$\lceil a_i 2^j \rceil - N \leq k_i \leq \lfloor b_i 2^j \rfloor + N - 1, \quad i = 1, \cdots, d. \tag{32}$$

Similar computations can be performed in the *Coiflet* wavelets case. We suggest rescaling the data to the same interval in order to simplify the calculations.

A second problem that can arise in practical applications is choosing an algorithm to calculate the values of the scaling function and the mother wavelet at arbitrary points. Daubechies wavelets do not have, in fact, analytic form, except for the Haar case. Several algorithms have been proposed. In the examples described in the next section we use the well known *cascade algorithm.* See Daubechies (1992) for a description.

7 Comparisons

Here we compare the wavelet estimators using both simulated and real data. The main purpose of the following examples is to provide a better understanding of the different types of wavelet estimators. Practical comparisons will demonstrate the usefulness of thresholding methods in estimating densities with discontinuities. Comparisons will also be given between wavelet estimators and more classical estimators of the same type, such as Fourier series estimators.

7.1 Simulated Data

We first analyze the performance of wavelet estimators using a family of normal mixture densities. This family of functions is a very broad class which represents different and challenging problems for density estimation. See Marron and Wand (1992) for formulae and descriptions of these densities.

Notice that in applying wavelet estimators an important choice is the wavelet family. Each family is characterized by a different degree of regularity and an ideal choice is to use functions with regularity similar or comparable to the regularity of the unknown density. See Tribouley (1995) for a discussion and an empirical method to choose the wavelet family.

Example 1: The Gaussian density. We start with a simple case by generating 500 observations from a Gaussian density. Given the smoothness of the density, we expect linear wavelet estimators to give good results.

While testing different wavelet families, we noticed that the Daubechies wavelets lead to oversmoothed estimates while better results can be obtained using the Coiflet families. Moreover, as we expected, wavelets with a higher degree of smoothness give better estimates. In order to choose the scaling parameter $\bar{\jmath}$ all the methods described in Section 5.1 seem to suggest $\bar{\jmath} = -1$ or $\bar{\jmath} = 0$ as the "optimal" value. Figure 1*b*. shows the linear wavelet estimator obtained using Coiflet wavelets with 5 vanishing moments and $\bar{\jmath} = 0$. In order to avoid negative values, the estimator is applied to the square root of the density. The pilot estimator (19) with radius $r = 0.3$ was chosen. A comparison with Figure 1*a*. shows that the linear wavelet estimator performs slightly better than a Fourier orthogonal series estimator. Results did not improve when using thresholded wavelet estimators.

Example 2: A separated bimodal density. We then generated 500 observations from a separated bimodal density. Again we expected linear wavelet estimators to give good results. We used the Daubechies wavelets. No improvements were obtained using Coiflet. The family *Daub*#10 gave the best results. The selected scale was $\bar{\jmath} = 0$. Figure 1*d*. shows the linear wavelet estimator with pilot estimator (19) and radius $r = 0.2$. A comparison with Figure 1*c*. shows that for this case wavelet estimators definitely perform better than the Fourier estimators. Localization in space clearly helps in estimating the modes of the density. Results did not improve when using thresholded wavelet estimators.

Example 3: The kurtotic unimodal density. More challenging problems may arise if we consider densities that present some kind of discontinuities. We first analyze the kurtotic unimodal density shown in Figure 2*a*. Less smooth wavelets need to be used and we got the best results using the *Daub*#3 family. We first applied the linear wavelet estimators. The selected scale was $\bar{\jmath} = 1$. Figure 2*c*. shows the linear wavelet estimator with pilot estimator (19) and radius $r = 0.3$. A comparison with Figure 2*b*. shows that, again, wavelet estimators perform better than more classical Fourier series estimators.

We then tried to improve the estimate by using thresholding techniques.

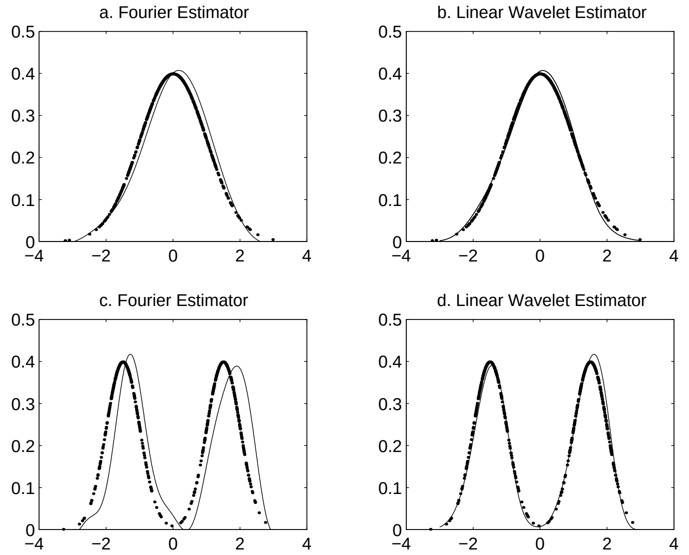

Figure 1: Fourier and wavelet estimators for Gaussian (*a.* and *b.*) and a separated bimodal (*c.* and *d.*) densities. Dotted lines: true densities. Continuous lines: estimated densities.

Notice that the nonlinear estimator (15) proposed by Hall and Patil corresponds in the case $p = 1$ to the estimator (12) of Donoho *et al.* with $j_0 = 0$ and $j_1 = q - 1$. In such a case the main difference between the two types of estimators consists of the proposed thresholding policies. Specifically, the choice suggested by Hall and Patil allows thresholding to be done at a selected number of finer scales ($j > q_1$ in the notation of Section 5.2) rather than at each scale $j > j_0$ as in the case of the Donoho *et al.* estimators. This allows more flexibility. On the other hand, negative values for the scale parameter j are not possible when using the estimators (15).

When applying these estimators to the kurtotic unimodal density we found that adding any of the coefficients at scales finer than 1 (that is either $q > 1$ or $j_1 > 0$) does not improve the performance of the estimators. However, focusing on the nonlinear estimator (15) of Hall and Patil we found that varying the smoothing parameter p can improve the estimate in a significant way. Figure 2*d.* shows the estimates obtained when selecting $q = 1$, thresholding wavelet coefficients at scale 1 and choosing $p = 1.8$.

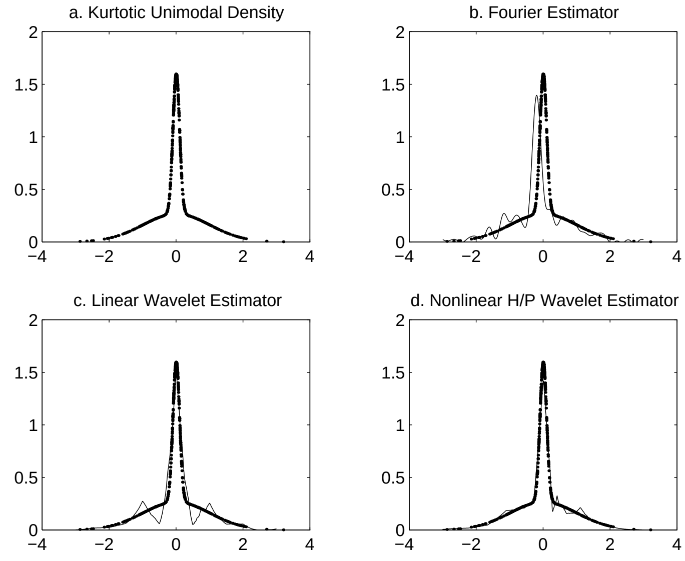

Figure 2: Fourier and wavelet estimators for a kurtotic unimodal density. Dotted lines: true densities. Continuous lines: estimated densities.

Example 4: The "claw" density. We finally analyze the "claw" density shown in Figure 3*a*. As in the previous example, we first applied Fourier and linear wavelet estimators. Figure 3*b*. shows the Fourier estimate. Linear wavelet estimators constructed with *Daub*#4 wavelets gave the best results. Figure 3*c*. shows the linear estimator with scale $\bar{\jmath} = 2$ and radius $r = 0.2$. Notice again how the modes are better estimated when using a wavelet estimator. We then tried to improve the obtained estimates by using thresholding techniques. In this case we got better results by employing the estimators (12) proposed by Donoho *et al.*. The estimator plotted in Figure 3*d*. was obtained choosing $j_0 = -1$, $j_1 = 1$ and applying the thresholding rule suggested by Delyon and Juditsky. The results did not improve when using the Hall and Patil estimators and varying the parameter p. We found that adding any of the coefficients at finer scales (that is $j_1 > 1$) does not improve the performance of the estimators.

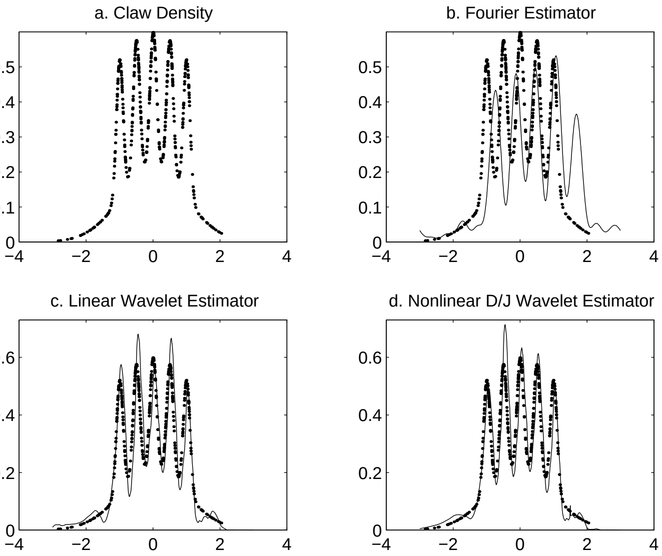

Figure 3: Fourier and wavelet estimators for a claw density. Dotted lines: true densities. Continuous lines: estimated densities.

7.2 A Real Data Example

A complete evaluation of the performance of any statistical method requires applications to real data. Here we apply wavelet density estimators to the Old Faithful Geyser dataset. This dataset has often been used as a benchmark for density estimators. Two-dimensional density plots of this dataset, obtained with the Average Shifted Histogram proposed by Scott (1992) and with a two-dimensional quartic kernel estimator, can be found in Hardle (1991) and Venables and Ripley (1992). The dataset is bivariate with components the *waiting time* between successive eruptions of the Old Faithful geyser in the Yellowstone National Park, Wyoming, and the *duration time* of eruptions. The waiting time is the time interval between starts of successive eruptions and the duration time is the duration of the subsequent eruption. Times are measured in minutes. The data were collected continuously from August 1st to August 15th, 1985, for a total of 299 pairs of observations. This data set was studied by Azzalini and Bowman (1990). A similar set of data was studied by Denby and Pregibon (1987) who pointed out a correlation among the data. For a description of the conjectured geyser mechanism see Rinehart (1969). The scatter plot of waiting time versus previous duration, Figure 4, shows that short durations can be associated with short waiting times and long durations with long waiting times.

We use wavelet estimators to estimate the univariate density function of the *duration time* component as well as the bivariate density. In order to avoid negative values, wavelet estimators are applied to the square root of the density using the pilot estimator (19) in the univariate case and the pilot estimator (20) in the bivariate case. Moreover, wavelet coefficients were re-normalized in order to get an estimate that integrates to one. We used linear estimators with Daubechies' wavelet families $Daub\#N$. Figure 5 shows the univariate linear wavelet estimator with $Daub\#7$ wavelet family, scale $j = 0$ and radius $r = 0.2$.

Figure 6 shows the bivariate linear wavelet estimator with $Daub\#7$ and pilot estimator (20) with radius $r = 0.25$. The value of the scale parameter, $j = 0$, was chosen by looking at the plots of the wavelet estimators obtained for different values of j.

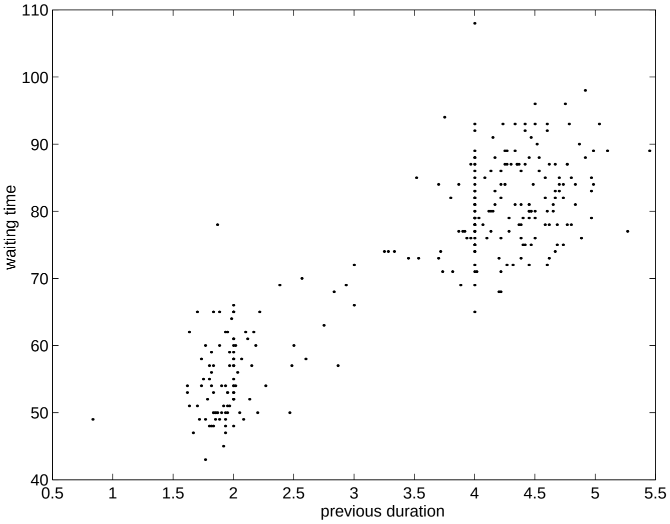

Figure 4: Dataset *Geyser*. Scatter plot of *waiting time* versus *previous duration*.

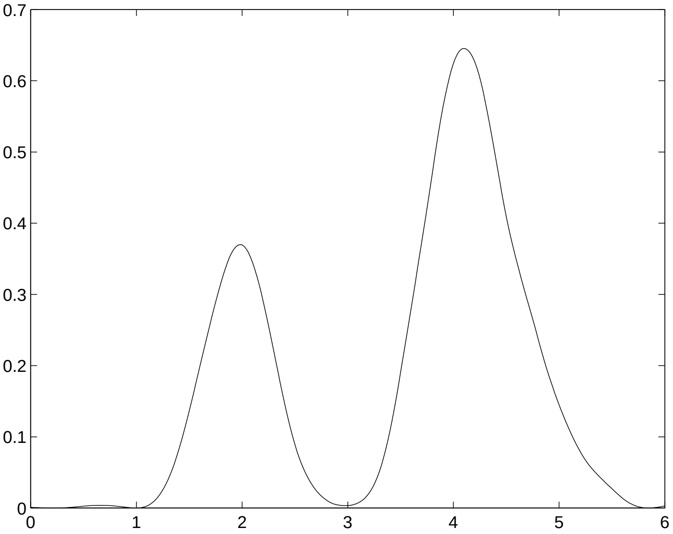

Figure 5: Dataset *Geyser*, *duration time* component. Wavelet density estimator with *Daub*#7 wavelet family, radius $r = 0.2$ and scale $j = 0$.

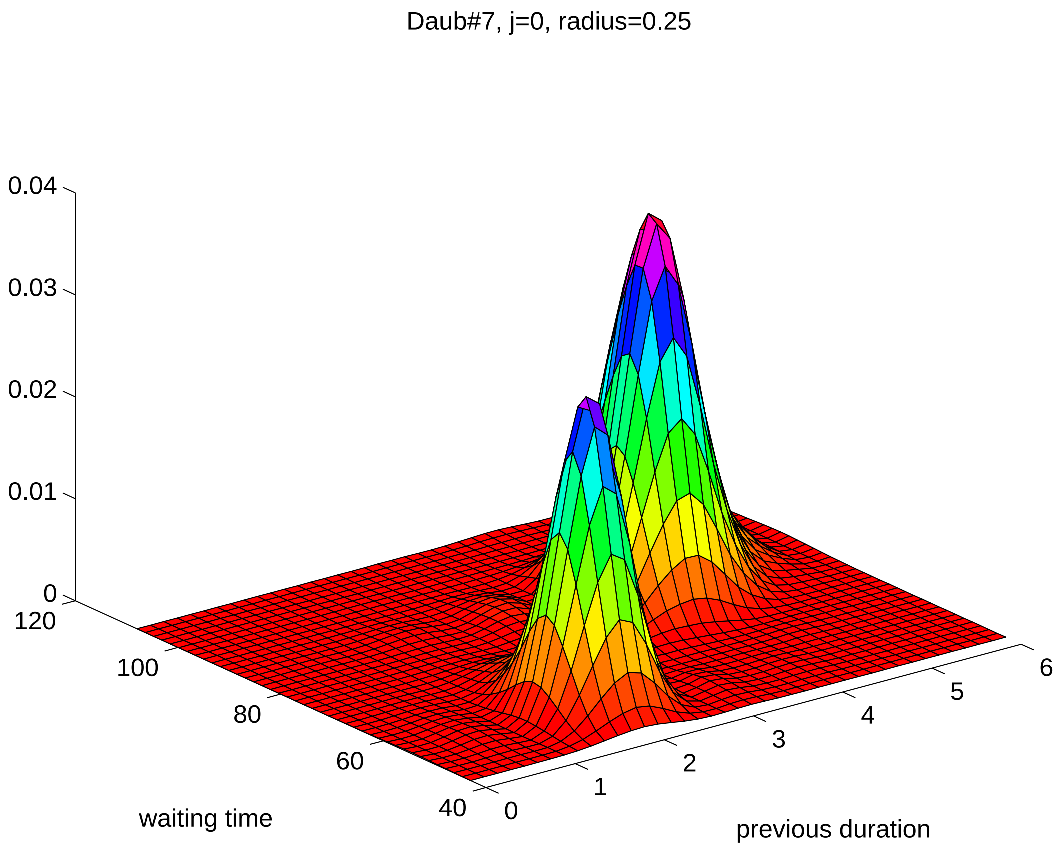

Figure 6: Wavelet density estimator for the Geyser dataset at the finest scale achieved ($j = 0$).

8 Concluding Remarks

Here we have explored the problem of density estimation using wavelets. Different types of wavelet density estimators, linear and thresholded, have been reviewed, stressing comparisons with classical estimators. Wavelet estimators have been classified as global estimators that gain their strength from the local properties of the wavelet functions. Practical problems, arising in applications, have been addressed. Examples have shown that linear wavelet estimators provide good estimates of smooth densities, while nonlinear (or thresholded) estimators allow for discontinuities.

Recently the wavelet density estimation problem has been approached from a Bayesian point of view. Among others, Müller and Vidakovic (1995) employ mixture priors and use Markov chain Monte Carlo methods to simulate from the posterior. Possible extensions to prior distributions that take into account the correlation among coefficients, as in Vannucci and Corradi (1997), may be possible. A review paper on these methods is in preparation (see Vannucci (1999)).

Acknowledgments

Discussions with Brani Vidakovic, Duke University, Antonio Moro, University of Florence, and Pol D. Spanos, Rice University, are gratefully acknowledged. Programs have been written in Matlab[3] with the aid of the toolbox *Wavbox 4.3* (Copyright Carl Taswell (1995)).

References

Antoniadis, A. and Carmona, R. (1991). Multiresolution analyses and wavelets for density estimation. Technical Report. University of California at Irvine.

Aronszajn, N. (1950). Theory of reproducing kernels. *Math. Soc.*, **68,** 337–404.

[3] The MathWorks, Inc, Natick, Mass.,USA

Azzalini, A. and Bowman, A.W. (1990). A look at some data on the Old Faithful Geyser. *Applied Statistics*, **39,** 357–365.

Bergh, J. and Löfström, J. (1976). *Interpolation Spaces, an Introduction.* Springer Verlag.

Čencov, N. N. (1962). Evaluation of an unknown distribution density from observations. *Doklady*, **3,** 1559–1562.

Chui, K. (1992). *An Introduction to Wavelets.* Academic Press.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* London, Chapman and Hall.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics* X, **LI,** 909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, volume 61. SIAM, CBMS-NSF Conference Series.

De Montricher, G.F., Tapia, R.A. and Thompson, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics*, **3(6),** 1329–1348.

Delyon, B. and Juditsky, A. (1993). Wavelet estimators, global error measures: Revisited. Technical Report. IRISA-INRIA. Available at http://www.irisa.fr.

Denby, L. and Pregibon, D. (1987). An example of the use of graphics in regression. *The American Statistician*, **41(1)**.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation, the L_1 view.* John Wiley & sons.

Donoho, D. (1993). Nonlinear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data. In *Proceedings of Symposia in Applied Mathematics*, volume 47, pp. 173–205. American Mathematical Society.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995b). Wavelet shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series* B, **57(2),** 301–369.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics*, **24(2),** 508–539.

Doukhan, P. and Leon, J. (1990). Deviation quadratique d'estimateur de densité par projection orthogonale. *Comptes Rendus Acad. Sciences Paris (A)*, **310,** 424–430.

Eubank, L.R. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker.

Good, I.J. and Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58(2),** 255–77.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Annals of Statistics*, **21(1),** 217–234.

Haar, A. (1910). Zur Theorie der orthogonalen Funktionen-Systeme. *Math. Ann.*, **69,** 331–371.

Hall, P., McKay, I. and Turlach, B.A. (1996). Performance of wavelet methods for functions with many discontinuities. *Annals of Statistics*, **24(6)**.

Hall, P. and Patil, P. (1995a). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Annals of Statistics*, **23(3),** 905–928.

Hall, P. and Patil, P. (1995b). On wavelet methods for estimating smooth functions. *Bernoulli*, **1,** 41–58.

Hall, P. and Patil, P. (1996). Effect of threshold rules on performance of wavelet-based curve estimators. *Statistica Sinica*, **6,** 331–345.

Härdle, W. (1991). *Smoothing Techniques with Implementation in S.* New York, Springer Verlag.

Izenman, A. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, **86,** 205–224.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statistics & Probability Letters*, **13,** 15–24.

Kerkyacharian, G. and Picard, D. (1993). Density estimation by kernel and wavelet methods: Optimality of Besov spaces. *Statistics & Probability Letters*, **18,** 327–336.

Leonard, T. (1973). A Bayesian method for histograms. *Biometrika*, **60(2),** 297–308.

Mallat, S.G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L_2(r)$. *Transactions of the American Mathematical Society*, **315(1),** 69–87.

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared error. *Annals of Statistics*, **20(2),** 712–736.

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

Müller, P. and Vidakovic, B. (1995). Bayesian inference with wavelets: Density estimations. Technical Report DP 95-33. ISDS, Duke University.

Neumann, M. (1996). Multivariate wavelet thresholding: a remedy against the curse of dimensionality. Technical Report. Humboldt University.

Pinheiro, A. and Vidakovic, B. (1997). Estimating the square root of a density via compactly supported wavelets. *Computational Statistics & Data Analysis*, **25,** 399–415.

Prakasa-Rao, B. (1983). *Nonparametric Functional Estimation.* Academic Press.

Rinehart, J.S. (1969). Thermal and seismic indications of Old Faithful Geyser's inner working. *J. Geophys. res.*, **74,** 566–573.

Scott, D.W. (1992). *Multivariate Density Estimation.* Wiley, New York.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, **10,** 795–810.

Tapia, R.A. and Thompson, J.R. (1978). *Nonparametric Probability Density Estimation.* The John Hopkins University Press, Baltimore and London.

Taswell, C. (1995). Wavbox 4: A software toolbox for wavelet transforms and adaptive wavelet packet decompositions. In *Wavelets and Statistics* (eds A. Antoniadis and G. Oppenheim). Springer Verlag.

Tribouley, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statistica Neerlandica*, **49(1),** 41–62.

Vannucci, M. (1996). On the application of wavelets in statistics. Ph.D. Thesis. Dipartimento di Statistica G.Parenti, University of Florence, Italy. In italian.

Vannucci, M. (1999). Density estimation and wavelets: Bayesian point of view. In *Bayesian Inference in Wavelet based Models* (eds B. Vidakovic and P. Müller). Springer Verlag.

Vannucci, M. and Corradi, F. (1997). Some findings on the covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective. Technical Report UKC/IMS/97/05. University of Kent at Canterbury. Under revision for Journal of the Royal Statistical Society, Series B.

Vannucci, M. and Vidakovic, B. (1995). Preventing the dirac disaster: wavelet based density estimation. Technical Report DP 95-27. ISDS, Duke University. Under revision for Journal of the Italian Staticatical Society.

Venables, W.N. and Ripley, B.D. (1992). *Modern Applied Statistics with S-Plus.* New York, Springer Verlag.

Walter, G.G. (1992). Approximation of Delta function by wavelets. *Journal of Approximation Theory*, **71,** 329–343.

Walter, G.G. (1994). *Wavelets and Others Orthogonal Systems with Applications.* CRC Press, Boca Raton, FL.

Walter, G.G. (1995). Estimation with wavelets and the curse of dimensionality. Technical Report. Department of Mathematical Sciences, University of Wisconsin-Milwaukee.

Walter, G.G. and Blum, J. (1979). Probability density estimation using Delta sequences. *Annals of Statistics*, **7(2),** 328–340.

Wu, D.W. (1994). Probability density estimation with wavelets. Ph.D. Thesis. University of Wisconsin-Milwaukee.

Wu, D.W. (1996). Asymptotic normality of the multiscale wavelet density estimator. *Communications in Statistics, Theory and Methods*, **25(9),** 1957–1970.