Shape-preserving wavelet-based multivariate density estimation

CARLOS AYA MORENO  GERY GEE NENS*  and  SPIRIDON PENEV
School of Mathematics and Statistics, UNSW Sydney, Australia

Abstract

Wavelet estimators for a probability density $f$ enjoy many good properties, however they are not ‘shape-

preserving’ in the sense that the final estimate may not be non-negative or integrate to unity. A solution
to negativity issues may be to estimate first the square-root of $f$ and then square this estimate up.
This paper proposes and investigates such an estimation scheme, generalising to higher dimensions some
previous constructions which are valid only in one dimension. The estimation is mainly based on nearest-
neighbour-balls. The theoretical properties of the proposed estimator are obtained, and it is shown to
reach the optimal rate of convergence uniformly over large classes of densities under mild conditions.
Simulations show that the new estimator performs as well in general as the classical wavelet estimator,
while automatically producing estimates which are bona fide densities.

1 Introduction

The mathematical theory of wavelets offers a powerful tool for approximating possibly irregular functions or
surfaces. It has been successfully applied in many different fields of applied mathematics and engineering,
see the classical references on the topic [Meyer 1992, Daubechies 1992], or Strang (1989, 1993) for shorter
reviews. In statistics, it provides a convenient framework for some nonparametric problems, in particular
density estimation and regression. As opposed to most of their competitors, such as kernels or splines,
wavelet-based estimators provide highly adaptive estimates by exploiting the localisation properties of the
wavelets. This translates into good global properties even when the estimated function presents sharp
features, such as acute peaks or abrupt changes. Indeed, wavelet estimators are (near-) optimal in some
sense over large classes of functions [Kerkyacharian and Picard 1993, Donoho et al 1996, 1995, Donoho

*Corresponding author: ggeenens@unsw.edu.au, School of Mathematics and Statistics, UNSW Sydney, NSW 2052 (Australia), tel +61 2 938 57032, fax +61 2 9385 7123
is called the ‘trend’ at level $j_0$, Chapter 2). This means that any square-integrable function $f \in L_2(\mathbb{R})$ can be expanded into that wavelet basis as

$$f(x) = \sum_{j \in \mathbb{Z}} \alpha_{j0,z}^* \varphi_{j0,z}(x) + \sum_{j=j_0}^{\infty} \sum_{z \in \mathbb{Z}} \beta_{j,z}^* \psi_{j,z}(x),$$

(1.1)

with $\forall j \in \mathbb{N}$ and $z \in \mathbb{Z}$, $\alpha_{j,z}^* = \int_{\mathbb{R}} \varphi_{j,z}(x)f(x) \, dx$ and $\beta_{j,z}^* = \int_{\mathbb{R}} \psi_{j,z}(x)f(x) \, dx$. The term $\sum_{z \in \mathbb{Z}} \alpha_{j0,z}^* \varphi_{j0,z}(x)$ is called the ‘trend’ at level $j_0$, while, for each level $j \geq j_0$, $\sum_{z \in \mathbb{Z}} \beta_{j,z}^* \psi_{j,z}(x)$ is the ‘detail’ at level $j$. A key feature of a multiresolution representation such as (1.1) is that, for any $j \geq j_0$, the trend at level $j+1$ coincides with the trend at level $j$ supplemented with the detail at level $j$. Specifically,

$$\sum_{z \in \mathbb{Z}} \alpha_{j+1,z}^* \varphi_{j+1,z}(x) = \sum_{z \in \mathbb{Z}} \alpha_{j,z}^* \varphi_{j,z}(x) + \sum_{z \in \mathbb{Z}} \beta_{j,z}^* \psi_{j,z}(x).$$

(1.2)

When $f$ in (1.1) is a probability density, noting that $\alpha_{j0,z}^* = \mathbb{E}(\varphi_{j0,z}(X))$ and $\beta_{j,z}^* = \mathbb{E}(\psi_{j,z}(X))$ paves the way for their estimation, upon observing a sample from $f$, by empirical averages, say $\hat{\alpha}_{j0,z}^*$ and $\hat{\beta}_{j,z}^*$. In addition, for any practical purpose the infinite expansion (1.1) needs to be truncated after a finite number of terms, say $J \geq j_0$ – in the wavelet jargon, one says that $f$ is approximated to the resolution level $J$. So, a wavelet estimator for $f$ writes

$$\hat{f}_J(x) = \sum_{z \in \mathbb{Z}} \hat{\alpha}_{j0,z}^* \varphi_{j0,z}(x) + \sum_{j=j_0}^{J} \sum_{z \in \mathbb{Z}} \hat{\beta}_{j,z}^* \psi_{j,z}(x),$$

which may ultimately include some thresholding of the estimated coefficients. Note that the sums over $z$ are essentially finite if the wavelets $\varphi$ and $\psi$ have compact support, as it is usually assumed.

Extending this framework to the multivariate case is conceptually straightforward. We assume that an orthogonal wavelet basis for $L_2(\mathbb{R}^d)$ is available – see [Meyer 1992, Section 3.6] for details about existence of such a basis. That is, there exist functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ and $\psi^{(q)} : \mathbb{R}^d \to \mathbb{R}$, $q \in Q = \{1, \ldots, 2^d - 1\}$, such that $\{\varphi_{j0,z}, \psi_{j,z}^{(q)} : j = j_0, \ldots, \infty, z \in \mathbb{Z}^d, q \in Q\}$ forms an orthonormal basis of $L_2(\mathbb{R}^d)$, with $\varphi_{j0,z}(x) = 2^{d/2} \varphi(2^{j0}x - z)$ and $\psi_{j,z}^{(q)}(x) = 2^{d/2} \psi^{(q)}(2^{j0}x - z)$. The functions $\psi^{(q)}$ are typically obtained via a tensor
product construction (Meyer, 1992, Sections 3.3-3.4). Then, any $d$-variate function $f \in L^2(\mathbb{R}^d)$ can be written

$$f(x) = \sum_{z \in \mathbb{Z}^d} \alpha^*_j(z) \varphi_j(x) + \sum_{j = j_0}^\infty \sum_{z \in \mathbb{Z}^d} \sum_{q \in Q} \beta^*_j(z) \psi_j(x),$$

(1.3)

where $\alpha^*_j(z) = \int_{\mathbb{R}^d} \varphi_j(x)f(x) \, dx$ and $\beta^*_j(z) = \int_{\mathbb{R}^d} \psi_j(x)f(x) \, dx$. When $f$ is a density, estimation of these coefficients, and hence of $f$ itself, follows in the same way as in one dimension.

One major drawback, though, of such wavelet-based estimators is that they are in general not ‘shape-preserving’. When estimating a probability density $f$, that means that the resulting estimator $\hat{f}$ may neither be non-negative, nor integrate to one (Dechevsky and Penev, 1997, 1998). Usually, simple rescaling solves the integrability issue, but overcoming the non-negativity issue requires caution. One way to address it is to first construct a wavelet estimator of $g = \sqrt{f}$ which, when squared up, would obviously produce an estimator of $f$ automatically satisfying the non-negativity constraint. Consider the univariate case. Clearly, $g \in L^2(\mathbb{R})$, as $\int_{\mathbb{R}} g^2(x) \, dx = \int_{\mathbb{R}} f(x) \, dx = 1$, hence we can write its expansion (1.1):

$$g(x) = \sum_{z \in \mathbb{Z}} \alpha_{j_0,z} \varphi_{j_0,z}(x) + \sum_{j = j_0}^\infty \sum_{z \in \mathbb{Z}} \beta_{j,z} \psi_{j,z}(x),$$

where

$$\alpha_{j,z} = \int_{\mathbb{R}} \varphi_{j,z}(x)g(x) \, dx = \int_{\mathbb{R}} \varphi_{j,z}(x)\sqrt{f}(x) \, dx \quad \text{and} \quad \beta_{j,z} = \int_{\mathbb{R}} \psi_{j,z}(x)g(x) \, dx = \int_{\mathbb{R}} \psi_{j,z}(x)\sqrt{f}(x) \, dx.$$ 

(1.4)

Difficulty in estimating these coefficients arises as $\alpha_{j,z} = E(\varphi_{j,z}(X)/\sqrt{f}(X))$ and $\beta_{j,z} = E(\psi_{j,z}(X)/\sqrt{f}(X))$ can no more be estimated directly by sample averages. Pinheiro and Vidakovic (1997) got around the presence of the unknown factor $1/\sqrt{f}$ in these expectations by plugging in a pilot estimator of $f$. Rather, Penev and Dechevsky (1997) suggested a more elegant construction based on order statistics and spacings. Unfortunately, direct application of their idea is limited to the univariate case, as spacings are not defined in more than one dimension. Yet, the need for a multivariate extension of the ‘Dechevsky-Penev’ construction was explicitly called for by McFadden (2003) in his Nobel Prize lecture. Cosma et al (2007) and Peter and Rangarajan (2008) attempted such extension but losing much of the initial flavour of the idea.

The aim of this paper is to suggest and study a wavelet estimator of $\sqrt{f}$ directly inspired by Penev and Dechevsky (1997)’s construction, hence keeping its simplicity and attractiveness, but available in any dimension. It will be shown in Section 2.1 that the volume of the smallest ball centred at $x$ and covering at least $k$ observations of the sample (for some $k \geq 1$), can act in some sense as a surrogate for a ‘multivariate
spacing’. The suggested estimator will then make use of \( k \)-nearest neighbour ideas, as will be formally defined in Section 2.2. Sections 3 and 4 respectively present the asymptotic properties of the proposed estimators of the wavelet coefficients and of the density estimator as a whole. Section 5 assesses the practical performance of the estimator through a simulation study and a real data application. Section 6 concludes and offers some perspectives of future research.

2 Definition of the estimator

2.1 Motivation

Let \( \mathcal{X} = \{X_1, \ldots, X_n\} \) be a random sample from an unknown \( d \)-dimensional distribution \( F \) admitting a density \( f \) on \( \mathbb{R}^d \). Denote by \( X_{(k);i} \) the \( k \)th closest observation from \( X_i \) among the other points of \( \mathcal{X} \). Define \( R_{(k);i} = \|X_{(k);i} - X_i\| \) the Euclidean distance between \( X_i \) and \( X_{(k);i} \), and

\[
V_{(k);i} = c_0 R_{(k);i}^d \quad \text{where} \quad c_0 = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)},
\]

the volume of the ball of radius \( R_{(k);i} \) centred at \( X_i \) – hence it is the smallest ball centred at \( X_i \) containing at least \( k \) other observations from \( \mathcal{X} \). It is known (Ranneby et al. 2005, Proposition 2) that, conditionally on \( X_i \),

\[
n V_{(1);i} \xrightarrow{\mathcal{L}} \text{Exp} (f(X_i)) \quad \text{as} \ n \to \infty,
\]

meaning that (Johnson et al. 1994, Section 10.5)

\[
\sqrt{n} V_{(1);i} \xrightarrow{\mathcal{L}} \text{Rayleigh} \left( \frac{1}{\sqrt{2} \sqrt{f(X_i)}} \right) \quad \text{as} \ n \to \infty.
\]

Now, consider an arbitrary square-integrable function \( \phi : \mathbb{R}^d \to \mathbb{R} \), and define

\[
S_n = \frac{2}{\sqrt{\pi}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i) \sqrt{V_{(1);i}}.
\]

By the Law of Iterated Expectations, we have

\[
\mathbb{E}(S_n) = \mathbb{E} \left( \frac{2}{\sqrt{\pi}} \phi(X_i) \mathbb{E} \left( \sqrt{n} V_{(1);i} \mid X_i \right) \right).
\]
The expectation of a Rayleigh(\(\sigma\))-random variable is known to be \(\sigma \sqrt{\frac{\pi}{2}}\). If the convergence in law (2.2) implies the convergence of the moments (this is indeed the case here as will be formally derived later), then
\[
\mathbb{E}(S_n) \to \mathbb{E}\left(\frac{2}{\sqrt{\pi}} \phi(X_i) \frac{\sqrt{\pi}}{2\sqrt{f}(X_i)}\right) = \int_{\mathbb{R}^d} \phi(x) \sqrt{f(x)} f(x) \, dx = \int_{\mathbb{R}^d} \phi(x) \sqrt{f}(x) \, dx.
\]

Hence, \(S_n\) is an asymptotically unbiased estimator of \(\int_{\mathbb{R}^d} \phi(x) \sqrt{f(x)} \, dx\). This fact naturally suggests estimating the wavelet coefficients (1.4) by statistics of type (2.3), which is the idea formally investigated in this paper.

**2.2 Definition**

Let \(g = \sqrt{f}\), where \(f\) is the \(d\)-dimensional density to estimate. As \(g \in L_2(\mathbb{R}^d)\) always, we have, by (1.3),
\[
g(x) = \sum_{j=0}^{\infty} \sum_{z \in \mathbb{Z}^d} \alpha_{j_0, z} \varphi_{j_0, z}(x) + \sum_{j=j_0}^{\infty} \sum_{z \in \mathbb{Z}^d} \sum_{q \in Q} \beta_{j, z}^{(q)} \psi_{j, z}^{(q)}(x),
\]
with, for all \(j \in \mathbb{N}\), \(z \in \mathbb{Z}^d\) and \(q \in \mathbb{Q}\),
\[
\alpha_{j, z} = \int_{\mathbb{R}^d} \varphi_{j, z}(x) \sqrt{f}(x) \, dx \quad \text{and} \quad \beta_{j, z}^{(q)} = \int_{\mathbb{R}^d} \psi_{j, z}^{(q)}(x) \sqrt{f}(x) \, dx.
\]

The approximation of \(g\) to the resolution level \(J \geq j_0\) is
\[
g(J)(x) = \sum_{z \in \mathbb{Z}^d} \alpha_{j_0, z} \varphi_{j_0, z}(x) + \sum_{j=j_0}^{\infty} \sum_{z \in \mathbb{Z}^d} \sum_{q \in Q} \beta_{j, z}^{(q)} \psi_{j, z}^{(q)}(x) = \sum_{z \in \mathbb{Z}^d} \alpha_{J+1, z} \varphi_{J+1, z}(x),
\]
where the second equality follows by analogy with (1.2).

Now, motivated by the observations made in Section 2.1, we define the estimators of the wavelet coefficients \(\alpha_{j, z}\)’s and \(\beta_{j, z}^{(q)}\)’s in (2.4)-(2.5) as
\[
\hat{\alpha}_{j, z} = \frac{\Gamma(k)}{\Gamma(k+1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{j, z}(X_i) \sqrt{V(k); z}, \quad j \in \mathbb{N}; \ z \in \mathbb{Z}^d \quad (2.6)
\]
\[
\hat{\beta}_{j, z}^{(q)} = \frac{\Gamma(k)}{\Gamma(k+1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{j, z}^{(q)}(X_i) \sqrt{V(k); z}, \quad j \in \mathbb{N}; \ z \in \mathbb{Z}^d; q \in \mathbb{Q}, \quad (2.7)
\]
for some integer \(k \geq 1\). The coefficient \(\frac{\Gamma(k)}{\Gamma(k+1/2)}\) guarantees the consistency of these estimators, as will arise from the proof of Proposition 3.1 below. Note that, for \(k = 1\), \(\frac{\Gamma(1)}{\Gamma(3/2)} = \frac{2}{\sqrt{\pi}}\), as it was anticipated in
Section 2.1. Also, in the case \( d = 1 \), when the volume of a ball amounts to the width of an interval, (2.6) and (2.7) can easily be compared to Penev and Dechevsky (1997)'s estimators (their equations (3.2) and (3.3)). Although not identical, they definitely have the same flavour and are asymptotically equivalent.

Plugging (2.6) and (2.7) into the expansion (2.4) produces the estimator

\[
\hat{g}_J(x) = \sum_{z \in \mathbb{Z}^d} \hat{\alpha}_{j_0,z} \varphi_{j_0,z}(x) + \sum_{j=j_0}^{J} \sum_{z \in \mathbb{Z}^d} \sum_{q \in Q} \hat{\beta}_{j,z}^{(q)} \psi_{j,z}^{(q)}(x),
\]

which is also

\[
\hat{g}_J(x) = \sum_{z \in \mathbb{Z}^d} \hat{\alpha}_{j + 1,z} \varphi_{j + 1,z}(x)
\]

by (2.5) and the properties of multiresolution analysis. Squaring this up provides an estimator \( \hat{f}_J \) of \( f \).

As already noted in Penev and Dechevsky (1997), estimating \( f \) by squaring up an estimate of \( \sqrt{f} \) has the additional advantage of providing an easy way for normalising the density estimate. Specifically, enforcing the condition \( 1 = \int_{\mathbb{R}^d} \hat{f}(x) \, dx = \int_{\mathbb{R}^d} \hat{g}_J^2(x) \, dx \) amounts to imposing

\[
\sum_{z \in \mathbb{Z}^d} \hat{\alpha}_{j_0,z}^2 + \sum_{j=j_0}^{J} \sum_{z \in \mathbb{Z}^d} \sum_{q \in Q} \hat{\beta}_{j,z}^{(q)2} = 1,
\]

given that the wavelets are orthonormal. If this sum is not 1 after raw estimation of the coefficients by (2.6) and (2.7) but, say, another constant \( \kappa \), it is enough to divide each estimated coefficient by \( \sqrt{\kappa} \) for enforcing (2.10). Conventional wavelet estimators do not enjoy such a convenient way of normalising.

In the following section, the asymptotic properties of the coefficient estimators (2.6) and (2.7) are obtained. The asymptotic properties of the estimator (2.8)-(2.9) for \( \sqrt{f} \) and the ensuing estimator \( \hat{f}_J = \hat{g}_J^2 \) for \( f \) will be obtained in Section 4.

3 Asymptotic properties of the estimators of the wavelet coefficients

Throughout the paper we work under the following two standard assumptions.

**Assumption 3.1.** The sample \( \mathcal{X} = \{X_1, \ldots, X_n\} \) consists of i.i.d. replications of a random variable \( X \in \mathbb{R}^d \) whose distribution \( F \) admits a density \( f \).

**Assumption 3.2.** The functions \( \varphi \) and \( \psi^{(q)} \) \((q \in Q)\), have compact support on \( \mathbb{R}^d \) and are bounded. Defining \( \varphi_{j_0,z}(x) = 2^{d j_0/2} \varphi(2^{j_0} x - z) \) and \( \psi_{j,z}^{(q)}(x) = 2^{d j/2} \psi^{(q)}(2^j x - z) \), \( \{\varphi_{j_0,z}, \psi_{j,z}^{(q)}, j = j_0, \ldots, \infty, z \in \mathbb{Z}^d, q \in Q\} \) is an orthonormal basis of \( L_2(\mathbb{R}^d) \).
Now, the main ingredients in (2.6) and (2.7) are the $V(k)_i$’s, which are ‘$k$th-nearest-neighbour’-type of quantities whose behaviour has been extensively studied in the literature (Mack and Rosenblatt 1979, Hall 1983, Percus and Martin 1998, Evans et al. 2002, Evans 2008). Good properties for such quantities require the underlying density $f$ to be well-behaved in the following sense.

**Assumption 3.3.** The density $f$ has convex compact support $C \subset \mathbb{R}^d$, with $\sup_{x,y \in C} \|x-y\| = c_1 < \infty$. It is bounded and bounded away from 0 on $C$, i.e., there exist constants $a_1$ and $a_2$ such that $\inf_{x \in C} f(x) = a_1 > 0$ and $\sup_{x \in C} f(x) = a_2 < \infty$. In addition, $f$ is differentiable on $C$, with uniformly bounded partial derivatives of the first order.

We have then the following result.

**Proposition 3.1.** Under Assumptions 3.1-3.3, for all $j = j_0, \ldots, J_z \in Z^d$ and $q \in Q$, the estimators (2.6) and (2.7) are such that

$$
E(\hat{\alpha}_{j,z}) = \alpha_{j,z} + O(n^{-1/d}), \quad \text{Var}(\hat{\alpha}_{j,z}) = k^3 \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 O(n^{-1}),
$$

$$
E(\hat{\beta}_{j,z}^{(q)}) = \beta_{j,z}^{(q)} + O(n^{-1/d}), \quad \text{Var}(\hat{\beta}_{j,z}^{(q)}) = k^3 \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 O(n^{-1}),
$$

as $n \to \infty$. In particular, if $k$ is such that $k^{3/2} \frac{\Gamma(k)}{\Gamma(k + 1/2)} = o(n^{1/2})$, then

$$
E \left( (\hat{\alpha}_{j,z} - \alpha_{j,z})^2 \right) \to 0 \quad \text{and} \quad E \left( (\hat{\beta}_{j,z}^{(q)} - \beta_{j,z}^{(q)})^2 \right) \to 0
$$

as $n \to \infty$, and the estimators are consistent.

**Proof.** The proof makes use of an extension of Theorem 5.4 in Evans et al. (2002), and is given in Appendix. \hfill $\square$

The condition $k^{3/2} \frac{\Gamma(k)}{\Gamma(k + 1/2)} = o(n^{1/2})$ is obviously satisfied if $k$ keeps a fixed value. It also allows $k$ to grow along with $n$. As $k \to \infty$, $\Gamma(k) / \Gamma(k + 1/2) \sim k^{-1/2}$ and the condition is equivalent to $k = o(n^{1/2})$. It appears that the (first order) asymptotic bias of $\hat{\alpha}_{j,z}$ and $\hat{\beta}_{j,z}^{(q)}$ does not depend on $k$, while their (first order) asymptotic variance increases with it. This can be attributed to larger covariances among the $V(k)_i$’s as $k$ gets large, and suggests – at least at this level – to keep $k$ as small as possible, that is, to use $k = 1$ always. By contrast, consistency of nonparametric density estimators built on $k$-Nearest-Neighbours ideas usually requires $k \to \infty$ as $n \to \infty$ (Mack and Rosenblatt 1979, Hall 1983). The fact that it seems here advantageous to keep $k$ as small as possible is, therefore, noteworthy. Below, the results are presented both for $k = k_n$ satisfying $k^{3/2} \frac{\Gamma(k)}{\Gamma(k + 1/2)} = o(n^{1/2})$ and for $k = 1$. 7
4 Asymptotic properties of the estimators of $\sqrt{f}$ and $f$

4.1 Pointwise consistency

In this subsection, the estimator $\hat{g}_J(x)$ (2.8)-(2.9) is first shown to be pointwise consistent for $\sqrt{f}(x)$ at all $x$. This essentially follows from the results of Section 3 through the theory of approximating kernels, see Bochner (1955) for early developments, and Meyer (1992) and Härdle et al (1998) for the wavelet case.

From the father wavelet $\varphi$, let the approximating kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be

$$K(x, y) = \sum_{z \in \mathbb{Z}^d} \varphi(x - z) \varphi(y - z)$$

and its refinement at resolution $j \in \mathbb{N}$ be

$$K_j(x, y) = \sum_{z \in \mathbb{Z}^d} 2^{dj} \varphi(2^j x - z) \varphi(2^j y - z) = \sum_{z \in \mathbb{Z}^d} \varphi_{j,z}(x) \varphi_{j,z}(y).$$

Define the two associated operators:

$$K \phi(x) = \int_{\mathbb{R}^d} K(x, y) \phi(y) \, dy \quad \text{and} \quad K_j \phi(x) = \int_{\mathbb{R}^d} K_j(x, y) \phi(y) \, dy,$$

for all functions $\phi \in L_2(\mathbb{R}^d)$. Then we have the following result.

**Proposition 4.1.** Under Assumptions 3.1-3.3, the estimator (2.8)-(2.9) is such that, at all $x \in C$,

(i) $\mathbb{E}(\hat{g}_J(x)) = K_{J+1} \sqrt{f}(x) + O(n^{-1/d})$, 

(ii) $\left( \frac{\Gamma(k + 1/2)}{\Gamma(k)} \right)^2 \frac{n}{K^3} \text{Var}(\hat{g}_J(x)) \leq \kappa \int_{\mathbb{R}^d} K_{J+1}^2(x, y) \, dy + O(n^{-1/d}),$

for some constant $\kappa < \infty$, as $n \to \infty$. Moreover, the order of the remainder terms holds uniformly in $x \in C$.

**Proof.** See Appendix.

This result obviously implies the pointwise consistency of $\hat{g}_J(x)$ for $\sqrt{f}(x)$ at any fixed $x \in C$ provided that $k^{3/2} \frac{\Gamma(k)}{\Gamma(k+1/2)} = o(n^{1/2})$, in particular if $k$ is kept fixed.

4.2 Uniform $L_2$-consistency

Consistency in Mean Integrated Squared Error ($L_2$-consistency) of estimator (2.8)-(2.9) can now be established uniformly over large classes of functions, such as Sobolev classes. Call $W^{m,p}(\Omega)$ the Sobolev space
of functions defined on $\Omega \subset \mathbb{R}^d$ for which all mixed partial derivatives up to order $m \geq 0$ exist (in the weak sense) and belong to $L_p(\Omega)$, $1 \leq p \leq \infty$. Formally,

$$W^{m,p}(\Omega) = \left\{ \phi \in L^p(\Omega) : D^\alpha \phi \in L^p(\Omega) \, \forall \alpha \in \mathbb{N}^d : |\alpha| \leq m \right\},$$

where $D^\alpha$ is the $\alpha^{th}$ (multi-index notation) partial weak derivative operator, and $|\alpha| = \sum_{k=1}^d \alpha_k$. A norm on $W^{m,p}(\Omega)$ is classically defined as $\|\phi\|_{m,p} = \sum_{|\alpha| \leq m} \|D^\alpha \phi\|_p$ (Triebel 1992).

It follows from Assumption 3.3 that there exists an integer $m \geq 1$ such that $f \in W^{m,2}(C)$: $f$ has uniformly bounded partial derivatives on $C$, which implies $f \in W^{1,\infty}(C)$, and as $W^{1,\infty}(C) \subset W^{1,2}(C)$, at least $m = 1$. Of course, more regular (i.e. smoother) densities $f$ allow for a higher value of $m$. In addition, under Assumption 3.3, $\sqrt{f} \in W^{m,2}(C)$ as well. This appears clearly from the multivariate version of Faà di Bruno’s formula (see e.g. Hardy (2006)), which reads here, for all $\alpha \in \mathbb{N}^d$ such that $|\alpha| \leq m$:

$$D^\alpha \sqrt{f} = \sum_{\xi \in \Xi} f^{1/2-|\xi|} \prod_{\beta \in \xi} D^\beta f,$$

where $\Xi$ is the set of all partitions $\xi$ of the elements of $\alpha$ and the product is over all ‘blocks’ $\beta$ of the partition $\xi$. Then the $L^2$-norm of the second factor in each term is bounded because $|\beta| \leq m$ and $f \in W^{m,2}(C)$, and the first factor $f^{1/2-|\xi|}$ is uniformly bounded for all $0 \leq |\xi| \leq m$, because $f$ is both bounded from above (case $|\xi| = 0$) and bounded away from 0 (case $|\xi| \geq 1$). This also implies that, if $f \in B^{m,2}(L) = \{ \phi \in W^{m,2}(C) : \|\phi\|_{m,2} \leq L \}$ for some constant $0 \leq L < \infty$, i.e., a ball of radius $L$ in $W^{m,2}(C)$, then $\sqrt{f} \in B^{m,2}(L')$ for some other constant $0 \leq L' < \infty$.

Now, suppose that the father wavelet $\varphi$ introduced in Assumption 3.2 is such that the induced kernel (4.1) satisfies the following assumption.

**Assumption 4.1.** The kernel $K$ (4.1) is such that $|K(x,y)| \leq F(x-y)$, for some square integrable function $F : \mathbb{R}^d \to \mathbb{R}$ with $\int_{\mathbb{R}^d} |x|^{\nu} F(x) \, dx < \infty$ for all $\nu \in \mathbb{N}^d$ such that $|\nu| = m$. Moreover, for all $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} (y-x)^{\nu} K(x,y) \, dy = \delta_{0,\nu'}, \text{ for all } \nu' \in \mathbb{N}^d \text{ such that } |\nu'| \leq m - 1.$$

Here, for $x \in \mathbb{R}^d$ and $\nu \in \mathbb{N}^d$, $|x|^{\nu} = \prod_{k=1}^d |x_k|^{\nu_k}$, and $\delta_{\nu,\nu'}$ is the $d$-fold Kronecker delta, equal to 1 if $\nu_k = \nu'_k \forall k \in \{1, \ldots, d\}$ and 0 otherwise. Then, one can prove the following.

**Theorem 4.1.** Under Assumptions 3.1-3.3 and Assumption 4.1, the estimator (2.8)-(2.9) is such that

$$\sup_{f \in B^{m,2}(L)} \mathbb{E} \left( \|\hat{g}_J - \sqrt{f}\|_2^2 \right) \leq \kappa_1 2^{-2Jm} + \kappa_2 n^{-2/d} + \kappa_3 n^{-1} k^3 \left( \frac{\Gamma(k)}{\Gamma(k+1/2)} \right)^2 2^{dJ},$$

(4.2)
for some constants $\kappa_1, \kappa_2, \kappa_3' < \infty$ and $n$ large enough.

Proof. See Appendix.

Clearly, the bound in the right-hand side of (4.2) is a non-decreasing function of $k$, which suggests to take $k = 1$ as it was already noted below Proposition 3.1. For that choice, we have directly:

**Corollary 4.1.** Under Assumptions 3.1-3.3 and Assumption 4.1, the estimator (2.8)-(2.9) with $k = 1$ in (2.6)-(2.7) is such that

$$
\sup_{f \in B^{m, 2}(L)} \mathbb{E} \left( \| \hat{g}_J - \sqrt{f} \|_2^2 \right) \leq \kappa_1 2^{-2Jm} + \kappa_2 n^{-2/d} + \kappa_3 \frac{2^{dJ}}{n},
$$

for some constants $\kappa_1, \kappa_2, \kappa_3' < \infty$ and $n$ large enough.

The terms depending on $J$ are balanced for $2^J \propto n^{\frac{1}{2m+2}}$, in which case

$$
\sup_{f \in B^{m, 2}(L)} \mathbb{E} \left( \| \hat{g}_J - \sqrt{f} \|_2^2 \right) \leq \kappa' n^{-\frac{2m}{2m+d}} + \kappa'' n^{-2/d},
$$

for two constants $\kappa', \kappa'' < \infty$. Finally, by the Cauchy-Schwartz inequality,

$$
\| \hat{g}_J^2 - f \|_2^2 = \|(\hat{g}_J - \sqrt{f})(\hat{g}_J + \sqrt{f})\|_2^2 \leq \| \hat{g}_J - \sqrt{f} \|_2^2 \times \| \hat{g}_J + \sqrt{f} \|_2^2.
$$

Assumptions 3.2 and 3.3 ensure that the second factor is bounded, whereby we have the following result about $\hat{g}_J^2$ as an estimator of the density $f$.

**Theorem 4.2.** Under Assumptions 3.1-3.3 and Assumption 4.1, the estimator $\hat{g}_J^2$ with $k = 1$ in (2.6)-(2.7) and $2^J \propto n^{\frac{1}{2m+d}}$ is a uniformly $L_2$-consistent estimator of $f$, such that

$$
\sup_{f \in B^{m, 2}(L)} \mathbb{E} \left( \| \hat{g}_J^2 - f \|_2^2 \right) \leq \kappa' n^{-\frac{2m}{2m+d}} + \kappa'' n^{-2/d},
$$

(4.3)

for some constants $\kappa', \kappa'' < \infty$.

Note that the first term in the right-hand side of (4.3) is the optimal nonparametric rate of convergence in this situation, as per Stone (1982)'s classical results. That term is dominated by the second one only for $d > \frac{2m}{m-1}$. Hence we have the following corollary.

**Corollary 4.2.** Under Assumptions 3.1-3.3 and Assumption 4.1, the estimator $\hat{g}_J^2$ with $k = 1$ in (2.6)-(2.7) and $2^J \propto n^{\frac{1}{2m+d}}$ is asymptotically optimal for $f$ uniformly over $B^{m, 2}(L) \subset W^{m, 2}(C)$, in the sense
that
\[
\sup_{f \in B^{m,2}(L)} \mathbb{E} \left( \| \hat{g}_J^2 - f \|_2^2 \right) \leq \kappa' n^{-2m/(2m+2)} ,
\]
for \( d \leq \frac{2m}{m-1} \).

As \( \frac{2m}{m-1} > 2 \), the estimator is always optimal in one and two dimensions. Under the classical mild smoothness assumption \( m = 2 \), it is optimal for \( 1 \leq d \leq 4 \) – this probably covers most of the cases of practical interest, given that the optimal rate of convergence itself becomes very poor in higher dimensions (Curse of Dimensionality, Geenens [2011]). In any case, for ‘rough’ densities \( f \) \((m = 1)\), the estimator reaches the optimal rate in all dimensions.

5 Numerical experiments

5.1 Simulation study

In this section the practical performance of the shape-preserving estimator \( \hat{g}_J^2 \) based on (2.8)-(2.9) is compared to that of the classical wavelet estimator. Three bivariate \((d = 2)\) Gaussian mixtures were considered: (a) two components, showing two peaks with very different covariance structures (Figure 5.1a); (b) two components, showing two similar peaks (Figure 5.1b), and (c) a bivariate version of Marron and Wand [1992]’s ‘smooth comb’, showing 4 peaks of decreasing spread (Figure 5.1c). Those were scaled and truncated to the unit square \([0,1]^2\), in order to satisfy Assumption 3.3. Note that mixtures (a) and (c) exhibit peaks of different spread and orientation, features known to cause difficulties in density estimation.

![Figure 5.1: Bivariate densities used in the simulation study.](image)

For each density, \( M = 500 \) random samples of size \( n = 2^\ell \), for \( \ell \in \{7, \ldots, 12\} \), that is, from \( n = 128 \) up to \( n = 4096 \), were generated, and our procedure was used on each of them for estimating \( f \). Proper

\footnote{The exact expressions are available from the authors upon request.}
\footnote{Sample sizes as powers of 2 are customary in the wavelet framework due to their suitability when resorting to the Fast Wavelet Transform, however the estimator described in Section 2.2 remains obviously valid for any arbitrary sample size \( n \).}
normalisation of all estimates was enforced through \( (2.10) \). The accuracy of a given estimate \( \hat{f} \) was measured by the Integrated Squared Error (ISE) \( \int_{[0,1]^2} \left( \hat{f}(x) - f(x) \right)^2 \, dx \), approximated by Riemannian summing on a fine regular partition of \([0,1]^2\). The Mean Integrated Squared Error (MISE) of an estimator was then approximated by averaging the ISE’s over the \( M = 500 \) Monte-Carlo replications, see Table 5.1.

Estimators \((2.6)-(2.7)\) were computed with bivariate wavelets \( \varphi_{j,z} \) and \( \psi_{j,z}^{(q)} \) obtained by tensor products of univariate Daubechies wavelets with 6 vanishing moments (Daubechies, 1992). In agreement with the asymptotic results, the value \( k = 1 \) in \((2.6)-(2.7)\) was given primary focus, but \( k = 2, 4, 8, \ldots, \sqrt{n} \) were also tested to investigate the effect of \( k \) in finite samples. For the three densities and all sample sizes, the choice \( k = 1 \) always lead to the final estimator with the smallest MISE, or within statistical significance (given \( M = 500 \) Monte-Carlo replications) to the estimator with the smallest MISE. Hence in Table 5.1 only the results for \( k = 1 \) are reported. In \((2.8)\), the baseline resolution was taken \( j_0 = 0 \) and the resolution levels \( J \in \{-1, 0, 1, 2, 3\} \) were considered – the case \( J = -1 \) is here defined as the estimator with the trend at baseline level \( j_0 = 0 \) only. For comparison, the density \( f \) was also estimated on each sample by the classical wavelet estimator described in Härdele et al (1998), whose MISE was approximated in the exact same way as above.

The whole procedure was developed in Python, using the BallTree \( k \)-Nearest neighbour algorithm (Omhundro, 1989) and the PyWavelets library that supports a number of orthogonal and biorthogonal wavelet families. It is available as open source in a github repository along with an implementation of the classic wavelet estimator. Note that, despite only the case \( d = 2 \) is reported here, the estimator can handle potentially any number of dimensions.

Analysing Table 5.1 reveals that neither estimator seems to have an absolute edge over the other, and the observed differences in MISE are low. For small sample sizes, the classical estimator is usually doing slightly better (although not always). This can be understood as it is based on simple averages which typically behave better than nearest-neighbour distances when the number of observations is not large. On the other hand, for larger samples, the Shape-Preserving (SP) estimator does usually better (although not always). It profits from the fact that it makes proper use of the probability mass that the classical one loses below zero in the low-density areas. This is illustrated by Figure 5.2 which shows typical estimates for the shape-preserving estimator and the classical one for sample size \( n = 4096 \) \((k = 1, j_0 = 0\) and \( J = 3 \)). Note how the classic estimator loses mass in areas of low density, even for this large sample. Therefore, although the results in Table 5.1 indicate that the classical estimator might be slightly more accurate \textit{sensu stricto} \((\text{MISE} = 0.090 \text{ for classical, MISE} = 0.134 \text{ for SP})\), it seems that the SP estimator may be preferable: the

\[3\text{https://github.com/carlosayam/PyWDE}\]
| $n$ | $J + 1$ | SP  | Class. | $n$ | $J + 1$ | SP  | Class. | $n$ | $J + 1$ | SP  | Class. |
|-----|--------|-----|--------|-----|--------|-----|--------|-----|--------|-----|--------|
| 128 |    0   | 3.490 | 3.686  | 128 |    0   | 4.315 | 4.383  | 128 |    0   | 8.320 | 8.324  |
|     |   1    | 2.907 | 3.097  |     |   1    | 3.269 | 3.436  |     |   1    | 6.800 | 6.818  |
|     |   2    | 1.358 | 1.086  |     |   2    | 0.836 | 1.225  |     |   2    | 4.565 | 5.142  |
|     |   3    | 1.199 | 0.862  |     |   3    | 0.912 | 0.540  |     |   3    | 1.972 | 2.082  |
|     |   4    | 4.697 | 1.964  |     |   4    | 4.337 | 1.909  |     |   4    | 3.743 | 2.159  |
| 256 |    0   | 3.491 | 3.686  | 256 |    0   | 4.314 | 4.384  | 256 |    0   | 8.319 | 8.323  |
|     |   1    | 2.891 | 3.092  |     |   1    | 3.253 | 3.430  |     |   1    | 6.793 | 6.813  |
|     |   2    | 1.286 | 1.043  |     |   2    | 0.747 | 1.184  |     |   2    | 4.480 | 5.096  |
|     |   3    | 0.778 | 0.634  |     |   3    | 0.471 | 0.326  |     |   3    | 1.561 | 1.861  |
|     |   4    | 2.351 | 0.995  |     |   4    | 2.143 | 0.989  |     |   4    | 1.973 | 1.213  |
| 512 |    0   | 3.491 | 3.686  | 512 |    0   | 4.314 | 4.383  | 512 |    0   | 8.319 | 8.323  |
|     |   1    | 2.880 | 3.090  |     |   1    | 3.247 | 3.427  |     |   1    | 6.790 | 6.811  |
|     |   2    | 1.235 | 1.022  |     |   2    | 0.714 | 1.165  |     |   2    | 4.425 | 5.074  |
|     |   3    | 0.543 | 0.523  |     |   3    | 0.245 | 0.210  |     |   3    | 1.332 | 1.756  |
|     |   4    | 1.093 | 0.518  |     |   4    | 1.048 | 0.496  |     |   4    | 1.087 | 0.745  |
| 1024|    0   | 3.491 | 3.686  | 1024|    0   | 4.314 | 4.383  | 1024|    0   | 8.320 | 8.323  |
|     |   1    | 2.873 | 3.088  |     |   1    | 3.243 | 3.425  |     |   1    | 6.790 | 6.809  |
|     |   2    | 1.211 | 1.012  |     |   2    | 0.690 | 1.152  |     |   2    | 4.395 | 5.064  |
|     |   3    | 0.406 | 0.468  |     |   3    | 0.137 | 0.151  |     |   3    | 1.203 | 1.700  |
|     |   4    | 0.529 | 0.274  |     |   4    | 0.519 | 0.245  |     |   4    | 0.585 | 0.499  |
| 2048|    0   | 3.491 | 3.686  | 2048|    0   | 4.314 | 4.383  | 2048|    0   | 8.320 | 8.323  |
|     |   1    | 2.872 | 3.087  |     |   1    | 3.243 | 3.424  |     |   1    | 6.789 | 6.809  |
|     |   2    | 1.190 | 1.007  |     |   2    | 0.680 | 1.147  |     |   2    | 4.380 | 5.058  |
|     |   3    | 0.343 | 0.439  |     |   3    | 0.081 | 0.123  |     |   3    | 1.141 | 1.673  |
|     |   4    | 0.267 | 0.149  |     |   4    | 0.261 | 0.123  |     |   4    | 0.349 | 0.378  |
| 4096|    0   | 3.492 | 3.686  | 4096|    0   | 4.313 | 4.383  | 4096|    0   | 8.320 | 8.323  |
|     |   1    | 2.871 | 3.087  |     |   1    | 3.241 | 3.424  |     |   1    | 6.787 | 6.809  |
|     |   2    | 1.180 | 1.004  |     |   2    | 0.676 | 1.145  |     |   2    | 4.370 | 5.055  |
|     |   3    | 0.312 | 0.425  |     |   3    | 0.051 | 0.110  |     |   3    | 1.103 | 1.659  |
|     |   4    | 0.134 | 0.090  |     |   4    | 0.128 | 0.061  |     |   4    | 0.225 | 0.317  |

Table 5.1: (Approximated) MISE of the shape-preserving estimator (SP) and the classical wavelet estimator (Class.) for different sample sizes and different values of $J + 1$ ($j_0 = 0$). The smallest MISE is highlighted for each sample size.

‘price to pay’ (in terms of MISE) for getting estimates which are automatically proper densities is quite low.

Figure 5.2 also reveals how challenging it is, for both estimators, to re-construct two peaks of such different spread. In that respect, the introduction of a thresholding scheme would be helpful to allow a higher resolution to be selected while killing out any unwarranted noise. The shape-preserving estimator is expected to profit more from the introduction of such thresholding, as it is noted from Table 5.1 that the classical estimator sometimes allows a higher resolution, already. More on this in Section 6.
Figure 5.2: Comparisons of estimates for Gaussian Mixture (a), $n = 4096$, $k = 1$ and $j_0 = 0$ and $J = 3$. 
5.2 Real data: Old Faithful geyser

Old Faithful geyser is a very active geyser in the Yellowstone National Park, Wyoming, USA. Data on eruption times and waiting times (both in minutes) between eruptions of Old Faithful form a well-known bivariate data set of $n = 272$ observations. In particular, it was used for illustration in Vannucci (1995), in a review of different types of wavelet density estimators. The shape-preserving estimator was computed on these data using Daubechies wavelets with 7 vanishing moments (as in Vannucci (1995)). The best results were obtained with $j_0 = 0$ and $J = 2$, producing the estimate shown in Figure 5.3. As opposed to Figure 6 in Vannucci (1995), the shape-preserving estimator shows some small bumps of potential interest near the main peaks. In view of the raw data (scatter plot, left panel) and other available kernel-based density estimates (Silverman 1986, Hyndman 1996), this seems legitimate.

6 Conclusions and future work

Penev and Dechevsky (1997) suggested an elegant construction of a wavelet estimator of the square-root of a univariate probability density in order to deal with negativity issues in an automatic way. Based on spacings, their idea could not be easily generalised beyond the univariate case, though. This paper provides such an extension, essentially making use of nearest-neighbour-balls, the “probabilistic counterpart to univariate spacings” (Ranneby et al. 2005) in higher dimensions. The asymptotic properties of the estimator were obtained. It always attains the optimal rate of convergence in Mean Integrated Square Error in $d = 1$ and $d = 2$ dimensions, in dimensions up to $d = 4$ for reasonably smooth densities, and in all

\footnote{see www.geyserstudy.org/geyser.aspx?GeyserNo=OLDFAITHFUL.}
dimensions for ‘rough’ densities. In practice, the estimator was seen to be on par with the classical wavelet estimator, while automatically producing estimates which are always bona fide densities.

Continuation of this research includes the introduction of a thresholding scheme. It is well-known that thresholding wavelet coefficients in the classical case gives better estimates in general Besov spaces (Donoho et al, 1996, Donoho and Johnstone, 1998). For a set of coefficients \( \{c_z; z \in \mathbb{Z}^d\} \) essentially defining a particular wavelet family, the father wavelet satisfies \( \phi(x) = \sum_{z \in \mathbb{Z}^d} c_z \phi(2x - z) \) (and similar for the functions \( \psi^{(q)} \)'s); see Daubechies (1992). This implies that \( \phi_{j,z}(x) = \sum_{z' \in \mathbb{Z}^d} c_{z'} \phi_{j+1,z'-2z}(x) \), which, in turn, carries over to the wavelet coefficients, viz. \( \alpha_{j,z} = \sum_{z' \in \mathbb{Z}^d} c_{z'} \alpha_{j,z-2z'} \) (and similar for the \( \beta \)'s). This dilation equation is often used for motivating and justifying thresholding in the conventional wavelet setting.

Now, substituting in (2.6) yields

\[
\hat{\alpha}_{j,z} = \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{j,z}(X_i) \sqrt{V(k);i} = \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \sum_{z' \in \mathbb{Z}^d} c_{z'} \varphi_{j+1,z'-2z}(X_i) \right\} \sqrt{V(k);i} = \sum_{z' \in \mathbb{Z}^d} c_{z'} \left\{ \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{j+1,z'-2z}(X_i) \sqrt{V(k);i} \right\} = \sum_{z' \in \mathbb{Z}^d} c_{z'} \hat{\alpha}_{j+1,z'-2z},
\]

and similar for the \( \hat{\beta}_{j,z}^{(q)} \)'s from (2.7). Hence, although the wavelet estimator developed in this paper is different in nature, the dilation equation applies to the estimated coefficients as it does in the conventional case. This suggests to carry on with thresholding for the shape-preserving estimator as well.

Some numerical experiments were carried out and, indeed, it was seen that improvements could be obtained. Figure 6.1b shows the shape-preserving estimator without thresholding on a typical sample of size \( n = 256 \) from the Gaussian mixture (a) (see Section 5) using Daubechies wavelets with 6 vanishing moments, \( k = 1, j_0 = 0 \) and \( J = 3 \). This resolution is of course too high at this sample size (see Table 5.1), and the estimate is highly undersmoothed. Then soft thresholding was applied in (2.8) on those estimated coefficients \( \hat{\beta}_{j,z}^{(q)} \) such that \( |\beta_{j,k}| < C \sqrt{j} + 1 / \sqrt{n} \), for an appropriate \( C \) (Delyon and Juditsky, 1996). The improvement is visually obvious (Figure 6.1c). The formal theoretical study of such a thresholding scheme is beyond the scope of this paper, though, and will be investigated in a follow-up paper.

Our theoretical results also provide some avenue for dealing with generalisations of sums like (2.6). For instance, Agarwal et al (2017) consider estimating the Fourier transform of the square-root of a probability
Figure 6.1: Shape-preserving estimates without (b) and with (c) thresholding for Gaussian mixture (a), $n = 256, k = 1, j_0 = 0, J = 3$.

density, viz. $\mathcal{F}\{\sqrt{f}\}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega x} \sqrt{f}(x) \, dx$, where $i = \sqrt{-1}$. Lemma A.1 ensures that

$$\frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{j=1}^{n} e^{-i\omega X_j} \sqrt{V(k);j}$$

is an asymptotically unbiased estimate of $\mathcal{F}\{\sqrt{f}\}(\omega)$ for all $\omega$’s, which could be used in several frameworks.

Finally, the estimator proposed here may provide interesting benefits in more applied settings as well, for instance for image and shape recognition, in the spirit of Peter and Rangarajan (2008) and Peter et al (2017).

Acknowledgements

This research involved computations using the Linux computational cluster Katana supported by the Faculty of Science, UNSW Sydney. The work of Gery Geenens was supported by a Faculty Research Grant from the Faculty of Science, UNSW Sydney. The work of Spiridon Penev was partially supported by the Australian Government through the Australian Research Council’s Discovery Projects funding scheme (project DP160103489).

A Appendix

Preliminaries

First some preliminary concepts and technical results are presented.

For any convex and compact $C \subset \mathbb{R}^d$, let $\partial C$ denote its boundary. For $\eta > 0$, define the $\eta$-belt of $C$ as

$$C(<\eta) = \left\{ x \in C : \inf_{y \in \partial C} \| y - x \| < \eta \right\},$$
the set of points in $C$ within Euclidean distance $\eta$ or less from $\partial C$. Also, we call $C(>\eta) = C \setminus C(<\eta)$ the $\eta$-interior of $C$.

Fix $x \in C$, call $B_x(r)$ the ball of radius $r$ centred at $x$ and $\mu(B_x(r)) = c_0 r^d$ its volume ($\mu$ is the Lebesgue measure on $\mathbb{R}^d$, $c_0 = \frac{\pi^{d/2}}{\Gamma(d/2+1)}$). Results in [Percus and Martin (1998) and Evans et al (2002)] show that the following two properties hold for any compact and convex set $C \subset \mathbb{R}^d$:

C1. There exists $c_2 > 0$, independent of $x \in C$, such that for $r < \sup_{x,y \in C} \|x-y\|$, $\mu(B_x(r) \cap C) \geq c_2 r^d$;

C2. There exist constants $\lambda > 0$ and $c_3 > 0$ such that for all $0 < \eta < \lambda$, $\mu(C(\eta)) < c_3 \eta$.

The following technical lemma will be used repeatedly in the proofs below.

**Lemma A.1.** Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ be a random sample from a distribution $F$ admitting a density $f$ supported on $C \subset \mathbb{R}^d$ satisfying Assumption 3.3. Let $R_{(k)}$ be the distance between $X_i$ and its $k$th nearest neighbour in the sample, as defined in Section 2.1. Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be bounded on $C$ and $a > 0$ such that $\int_{\mathbb{R}^d} \phi(x) f(x)^{1-a} \, dx < \infty$. Then, $\forall i \in \{1, \ldots, n\}$, as $n \to \infty$,

$$
\mathbb{E} \left( \phi(X_i) R_{(k):i}^{ad} \right) = \mathbb{E} \left( \phi(X_i) \mathbb{E} \left( R_{(k):i}^{ad} | X_i \right) \right) = \frac{1}{n^a} \frac{\Gamma(k+a)}{\Gamma(k)} \frac{1}{c_0^d} \left( \int_{\mathbb{R}^d} \phi(x) f(x)^{1-a} \, dx + O \left( n^{-1/d} \right) \right).
$$

**Proof.** Call

$$
\omega_x(r) = \int_{B_x(r)} f(z) \, dz,
$$

the probability that the random variable $X \sim F$ falls in $B_x(r)$, and set $\omega_i(r) = \omega_{X_i}(r)$ when referring to the ball centred at one particular observation $X_i$ from the sample. Let $F_{(k):i}$ be the distribution function of $R_{(k):i}$ for fixed $X_i$, that is, $F_{(k):i}(r) = \mathbb{P}(R_{(k):i} \leq r | X_i)$. With $X_i$ fixed, Lemma 4.1 in [Evans et al (2002)] writes

$$
dF_{(k):i}(r) = k \left( \frac{n-1}{k} \right) \omega_i(r)^{k-1} (1 - \omega_i(r))^{n-k-1} \, d\omega_i(r).
$$

Hence

$$
\mathbb{E} \left( R_{(k):i}^{ad} | X_i \right) = k \left( \frac{n-1}{k} \right) \int_0^{c_1} r^{ad} \omega_i(r)^{k-1} (1 - \omega_i(r))^{n-k-1} \, d\omega_i(r).
$$

Since $f$ is positive on $C$ and $C$ is convex, $\omega_i(r)$ is strictly increasing for $0 \leq r \leq r_0$ for some $r_0$, and $\omega_i(r) \equiv 1$ for $r_0 \leq r$. Writing $h_i(\omega)$ for the inverse function $\omega_i^{-1}$ (where it exists), a change of variable yields

$$
\mathbb{E} \left( R_{(k):i}^{ad} | X_i \right) = k \left( \frac{n-1}{k} \right) \int_0^{1} h_i(\omega)^{ad} \omega^{k-1} (1 - \omega)^{n-k-1} \, d\omega.
$$
Define $\delta_n \doteq n^{-1/d}$, and break this expectation down into

$$
\mathbb{E} \left( R_{(k);i}^{ad} | X_i \right) = k \left( \frac{n-1}{k} \right) \int_0^{\omega(\delta_n)} h_i(\omega) \omega^{d-1} (1 - \omega)^{n-k-1} \, d\omega + k \left( \frac{n-1}{k} \right) \int_{\omega(\delta_n)}^{1} h_i(\omega) \omega^{d-1} (1 - \omega)^{n-k-1} \, d\omega
$$

$$
= k \left( \frac{n-1}{k} \right) \int_0^{\omega(\delta_n)} h_i(\omega) \omega^{d-1} (1 - \omega)^{n-k-1} \, d\omega + O(n^{-b})
$$

for all $b > 0$, uniformly in $X_i$, as per Lemma 5.3 of [Evans et al. 2002].

Now, with $h_x = \omega_x^{-1}$, see that

$$
\mathbb{E} \left( \phi(X_i) \mathbb{E} \left( R_{(k);i}^{ad} | X_i \right) \right) = \int_C \phi(x) \left\{ k \left( \frac{n-1}{k} \right) \int_0^{1} h_x(\omega) \omega^{d-1} (1 - \omega)^{n-k-1} \, d\omega \right\} f(x) \, dx \quad \text{(A.1)}
$$

$$
= \int_C \phi(x) \left\{ k \left( \frac{n-1}{k} \right) \int_0^{\omega_x(\delta_n)} h_x(\omega) \omega^{d-1} (1 - \omega)^{n-k-1} \, d\omega \right\} f(x) \, dx + O(n^{-b}),
$$

as $\phi$ and $f$ are bounded on the compact $C$. As $b$ can be taken arbitrarily large, the remainder term can be neglected in front of any term tending to 0 polynomially fast. Hence, (asymptotically) all contribution to the inner integral in (A.1) comes from the set $\omega \in (0, \omega_x(\delta_n))$, that is, when $R_{(k);i}$ is smaller than $\delta_n$.

Now, write (A.1) as

$$
\int_C \ldots \, dx = \int_{C(>\delta_n)} \ldots \, dx + \int_{C(<\delta_n)} \ldots \, dx \doteq (I) + (II)
$$

with $C(>\delta_n)$ and $C(<\delta_n)$ the $\delta_n$-interior and $\delta_n$-belt of $C$ as defined above.

**Integral (I):** $\int_{C(>\delta_n)} \ldots \, dx$, hence $x \in \delta_n$-interior and the distance from $x$ to $\partial C$ is at least $\delta_n$. Hence for all $r \leq \delta_n$, $B_x(r) \cap C = B_x(r)$. The first mean value theorem for definite integrals establishes the existence of $\xi_1 \in B_x(r) \subset C$ such that

$$
\omega_x(r) = \int_{B_x(r)} f(z) \, dz = f(\xi_1) \mu(B_x(r)) = f(\xi_1)c_0 r^d. \quad \text{(A.2)}
$$

By the mean value theorem, there is a $\xi_2$ between $x$ and $\xi_1$, hence $\xi_2 \in B_x(r) \subset C$, such that $f(\xi_1) = f(x) + \nabla f(\xi_2)^t(x - \xi_1)$. Because $\xi_1 \in B_x(\delta_n)$ and $\|\nabla f(\xi_2)\| < M$ for an absolute constant $M$ (the partial derivatives of $f$ are uniformly bounded on $C$ by Assumption 3.3), we have $|f(\xi_1) - f(x)| < \delta_n M$ and hence $f(\xi_1) = f(x) + O(\delta_n)$. Substitution in (A.2) gives $\omega_x(r) = (f(x) + O(\delta_n)) c_0 r^d$. As $f$ is bounded from below, this means that, as $n \to \infty$,

$$
h_x(\omega) = \left( \frac{\omega}{c_0 f(x)} \right)^{1/d} (1 + O(\delta_n)),
$$

where the $O(\delta_n)$-term holds uniformly in $x$ and $\omega$. This can be substituted in the inner integral of (A.1),
and we obtain

\[
\int_{C(>\delta_n)} \phi(x) \left\{ k \left( \frac{n-1}{k} \right) \int_0^1 h_x(\omega)^{\alpha d} \omega^{k-1}(1-\omega)^{n-k-1} \, d\omega \right\} f(x) \, dx
\]

\[
= \frac{\Gamma(n)}{\Gamma(k)\Gamma(n-k)} \left( 1 + O(\delta_n) \right) \int_{C(>\delta_n)} \phi(x) \frac{f^{1-a}(x)}{c_0^a} \, dx \int_0^1 \omega^{a+k-1}(1-\omega)^{n-k-1} \, d\omega
\]

\[
= \frac{\Gamma(n)}{\Gamma(k)\Gamma(n-k)} \frac{\Gamma(k+a)\Gamma(n-k)}{\Gamma(n+a)} \left( 1 + O(\delta_n) \right) \int_{C(>\delta_n)} \phi(x) \frac{f^{1-a}(x)}{c_0^a} \, dx
\]

Now, given that \( f \) is bounded from below and above on \( C \), \( f(x)^{1-a} \leq a_3 = \max\{(1/a_1)^{1-a}, a_2^{1-a}\} \), and by C2 above, \( \mu(C(<\delta_n)) < c_3 \delta_n \) for \( n \) large enough. So,

\[
\left| \int_{C(<\delta_n)} \phi(x) f^{1-a}(x) \, dx \right| \leq \sup_{x \in C(<\delta_n)} |\phi(x)| a_3 c_3 \delta_n = O(\delta_n),
\]

as \( n \to \infty \). Therefore,

\[
\int_{C(>\delta_n)} \phi(x) f^{1-a}(x) \, dx = \int_{C(>\delta_n)} \phi(x) f^{1-a}(x) \, dx + O(\delta_n) = \int_{\mathbb{R}^d} \phi(x) f^{1-a}(x) \, dx + O(\delta_n).
\]

Noting that \( \Gamma(n)/\Gamma(n+a) = n^{-a}(1 + O(n^{-1})) = n^{-a}(1 + O(\delta_n)) \), we finally get

\[
\int_{C(>\delta_n)} \phi(x) \left\{ k \left( \frac{n-1}{k} \right) \int_0^1 h_x(\omega)^{\alpha d} \omega^{k-1}(1-\omega)^{n-k-1} \, d\omega \right\} f(x) \, dx
\]

\[
= \frac{1}{n^a} \frac{\Gamma(k+a)}{\Gamma(k)} \frac{1}{c_0^a} \left( \int_{\mathbb{R}^d} \phi(x) f(x)^{1-a} \, dx + O \left( n^{-1/d} \right) \right). \quad (A.3)
\]

**Integral (II):** \( \int_{C(\leq \delta_n)} \ldots \, dx \), hence we can no more assume that \( B_x(r) \subset C \). However, as \( \sup_{x \in C(\leq \delta_n)} f(x) \leq \sup_{x \in C} f(x) \leq a_2 \) and \( \mu(B_x(r) \cap C) < \mu(B_x(r) = c_0 d \), it holds \( \omega_x(r) < a_2 c_0 d \). An upper bound for its inverse is thus \( h_x(\omega) \leq (a_2 c_0)^{-1/d} \omega^{1/d} \). Hence,

\[
\left| (II) \right| \leq \int_{C(\leq \delta_n)} |\phi(x)| \left\{ k \left( \frac{n-1}{k} \right) \int_0^1 h_x(\omega)^{\alpha d} \omega^{k-1}(1-\omega)^{n-k-1} \, d\omega \right\} f(x) \, dx
\]

\[
\leq \int_{C(\leq \delta_n)} |\phi(x)| \left\{ k \left( \frac{n-1}{k} \right) \int_0^1 (a_2 c_0)^{-a} \omega^{a+k-1}(1-\omega)^{n-k-1} \, d\omega \right\} f(x) \, dx
\]

\[
= \frac{\Gamma(a+k)}{\Gamma(k)} \frac{\Gamma(n)}{\Gamma(n+a)} (a_2 c_0)^{-a} \int_{C(\leq \delta_n)} |\phi(x)| f(x) \, dx
\]

\[
\leq \frac{\Gamma(a+k)}{\Gamma(k)} \frac{\Gamma(n)}{\Gamma(n+a)} (a_2 c_0)^{-a} \sup_{x \in C} |\phi(x)| a_2 c_3 \delta_n,
\]

20
by C2 above. Thus,

\[ |(II)| \leq \frac{\Gamma(k + a)}{\Gamma(k)} O(n^{-a} \delta_n) = \frac{\Gamma(k + a)}{\Gamma(k)} O(n^{-a-1/d}) \quad \text{as} \quad n \to \infty. \quad \text{(A.4)} \]

Putting together (A.3) and (A.4) in (A.1), it follows

\[ E \left( \phi(X_i) R_{(k);i}^{ad} \right) = \frac{1}{n^a} \frac{\Gamma(k + a)}{\Gamma(k)} \frac{1}{\sqrt{c_0}} \left( \int_{\mathbb{R}^d} \phi(x) f(x)^{1-a} dx + O \left( n^{-1/d} \right) \right), \]

as announced.

**Proof of Proposition 3.1**

The proof is given for the coefficients \( \hat{\alpha}_{j,z} \). The proof for the coefficients \( \hat{\beta}^{(q)}_{j,z} \) is identical.

**Bias:** From (2.6), we have with (2.1)

\[ E(\hat{\alpha}_{j,z}) = E \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi_{j,z}(X_i) \sqrt{V_{(k);i}} \right) = n^{1/2} \frac{\Gamma(k)}{\Gamma(k + 1/2)} \sqrt{c_0} E \left( \varphi_{j,z}(X_1) R_{(k);1}^{d/2} \right). \]

Applying Lemma A.1 with \( \phi = \varphi_{j,z} \) and \( a = 1/2 \) yields

\[ E \left( \varphi_{j,z}(X_1) R_{(k);1}^{d/2} \right) = n^{-1/2} \frac{\Gamma(k + 1/2)}{\Gamma(k)} \frac{1}{\sqrt{c_0}} \left( \int_{\mathbb{R}^d} \varphi_{j,z}(x) \sqrt{f(x)} dx + O(n^{-1/d}) \right), \]

which gives

\[ E(\hat{\alpha}_{j,z}) = \int_{\mathbb{R}^d} \varphi_{j,z}(x) \sqrt{f(x)} dx + O(n^{-1/d}) = \alpha_{j,z} + O(n^{-1/d}). \]

**Variance:** Lemma 4.6(ii) of [Evans (2008)] gives an upper bound on the variance of statistics of type

\[ S_n = \sum_{i=1}^n h_{i,n}(\mathcal{X}), \quad \text{where} \quad h_{i,n}(\mathcal{X}) \quad \text{is an arbitrary (measurable) function of the sample point} \quad X_i \quad \text{and its} \quad k-\text{nearest neighbours among the sample} \quad \mathcal{X}. \]

Take here

\[ h_{i,n}(\mathcal{X}) = \varphi_{j,z}(X_i) \sqrt{V_{(k);i}} \]

and see that \( \hat{\alpha}_{j,z} = \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} S_n. \) Lemma 4.6(ii) of [Evans (2008)] reads

\[ \text{Var}(S_n) \leq 2(n + 1)(3 + 8k^2 c_0) E \left( h_{i,n}^2(\mathcal{X}) \right), \quad \text{(A.5)} \]
for $n \geq 16k$. Here,
\[
E(h_{i,n}^2(\mathcal{X})) = E(\varphi_{j,z}^2(X_i) V_{(k);i})
\]
\[
= c_0 E(\varphi_{j,z}^2(X_i) R_{(k);i}^d)
\]
\[
= \frac{k}{n} \left( \int_{\mathbb{R}^d} \varphi_{j,z}^2(x) \, dx + O(n^{-1/d}) \right),
\]
from Lemma A.1 with $\phi = \varphi_{j,z}^2$ and $a = 1$. By definition, $\int_{\mathbb{R}^d} \varphi_{j,z}^2(x) \, dx = 1$ (orthonormal wavelet basis, Assumption 3.2), hence $E(h_{i,n}^2(\mathcal{X})) = \frac{k}{n}(1 + O(n^{-1/d}))$. From this and (A.5), we obtain
\[
\mathbb{V}ar(\hat{\alpha}_{j,z}) \leq \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 \frac{1}{n} \frac{2(n + 1)(3 + 8k^2d_c)}{n^2} \left( \frac{k}{n} \right) = k^3 \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 O(n^{-1}),
\]
as $n \to \infty$.

\[\square\]

**Proof of Proposition 4.1**

From (2.9) we have
\[
\hat{g}_J(x) = \sum_{z \in \mathbb{Z}^d} \hat{\alpha}_{J+1,z} \varphi_{J+1,z}(x)
\]
\[
= \sum_{z \in \mathbb{Z}^d} \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{J+1,z}(X_i) \sqrt{V_{(k);i}} \varphi_{J+1,z}(x)
\]
\[
= \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_{(k);i}^{d/2} \sum_{z \in \mathbb{Z}^d} \varphi_{J+1,z}(X_i) \varphi_{J+1,z}(x)
\]
\[
= \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_{(k);i}^{d/2} K_{J+1}(x, X_i),
\]
(A.6)

hence
\[
E(\hat{g}_J(x)) = \frac{\Gamma(k)}{\Gamma(k + 1/2)} \frac{1}{\sqrt{n}} \sqrt{c_0} E \left( K_{J+1}(x, X_1) R_{(k);1}^{d/2} \right).
\]

Lemma A.1 with $\phi = K_{J+1}(x, \cdot)$ and $a = 1/2$ establishes that
\[
E \left( K_{J+1}(x, X_1) R_{(k);1}^{d/2} \right) = \frac{1}{\sqrt{n}} \frac{\Gamma(k + 1/2)}{\Gamma(k)} \frac{1}{\sqrt{c_0}} \left( \int_{\mathbb{R}^d} K_{J+1}(x, y) \sqrt{f(y)} \, dy + O(n^{-1/d}) \right),
\]
and inspection of the proof of Lemma A.1 reveals that the $O(n^{-1/d})$ term holds uniformly in $x \in C$. This means that
\[
E(\hat{g}_J(x)) = \int_{\mathbb{R}^d} K_{J+1}(x, y) \sqrt{f(y)} \, dy + O(n^{-1/d}) = K_{J+1} \sqrt{f}(x) + O(n^{-1/d}),
\]

\[22\]
as \( n \to \infty \), uniformly in \( x \in C \), proving (i).

It follows from (A.6) as well that

\[
\text{Var} \left( \frac{\Gamma(k + 1/2)}{\Gamma(k)} \sqrt{\frac{n}{K^3}} \hat{g}_J(x) \right) = \frac{c_0}{k^3} \text{Var} \left( \sum_{i=1}^{n} h_{i,n}(X) \right)
\]

where here \( h_{i,n}(X) \equiv K_{J+1}(x, X_i)R_{(k)}^{d/2} \). Lemma A.1 with \( a = 1 \) and \( \phi = K_{J+1}^2(x, \cdot) \) yields

\[
\mathbb{E} \left( h_{i,n}^2(X) \right) = \frac{k}{c_0 n} \left( \int_{\mathbb{R}^d} K_{J+1}^2(x, y) \, dy + O(n^{-1/d}) \right)
\]

(with again the \( O(n^{-1/d}) \)-term holding uniformly in \( x \in C \)). Hence, for \( n \geq 16k \), Lemma 4.6(ii) of Evans (2008) gives

\[
\text{Var} \left( \sum_{i=1}^{n} h_{i,n}(X) \right) \leq 2(n + 1)(3 + 8k^2dc_0) \frac{k}{c_0 n} \left( \int_{\mathbb{R}^d} K_{J+1}^2(x, y) \, dy + O(n^{-1/d}) \right),
\]

whereby

\[
\text{Var} \left( \frac{\Gamma(k + 1/2)}{\Gamma(k)} \sqrt{\frac{n}{K^3}} \hat{g}_J(x) \right) \leq \text{constant} \times \int_{\mathbb{R}^d} K_{J+1}^2(x, y) \, dy + O(n^{-1/d}).
\]

This establishes (ii). \( \square \)

**Proof of Theorem 4.1**

The Mean Integrated Squared Error (MISE) \( \mathbb{E} \left( \| \hat{g}_J - \sqrt{f} \|_2^2 \right) \) can classically be decomposed into the integrated squared bias and the integrated variance:

\[
\mathbb{E} \left( \| \hat{g}_J - \sqrt{f} \|_2^2 \right) = \| \mathbb{E}(\hat{g}_J) - \sqrt{f} \|_2^2 + \mathbb{E} \left( \| \hat{g}_J - \mathbb{E}(\hat{g}_J) \|_2^2 \right).
\]

For the bias term, it follows from Proposition 4.1(i) that

\[
\| \mathbb{E}(\hat{g}_J) - \sqrt{f} \|_2 \leq \| K_{J+1} \sqrt{f} - \sqrt{f} \|_2 + O(n^{-1/d}).
\]

As \( f \in B^{m,2}(L) \) implies \( \sqrt{f} \in B^{m,2}(L') \) for some \( 0 \leq L' < \infty \), one can call (multivariate versions of) Theorem 8.1(ii) and Corollary 10.1 of Hardle et al. (1998) to obtain

\[
\sup_{f \in B^{m,2}(L)} \| K_{J+1} \sqrt{f} - \sqrt{f} \|_2 \leq \kappa_1 2^{-J_m},
\]

23
for some constant $\kappa_1$. Hence, for $n$ large enough,

$$\sup_{f \in B_n^{2,L}(L)} \|f \sqrt{\hat{g}_J} - \sqrt{f}\|_2 \leq \kappa_1 2^{-J} + \kappa_2 n^{-1/d},$$  \hfill (A.8)

for constants $\kappa_1, \kappa_2 < \infty$.

To evaluate $\int_{\mathbb{R}^d} K_{J+1}^2(x,y) \, dy$ in the right-hand side of Proposition 4.1 (ii) we use that

$$\int_{\mathbb{R}^d} K_{J+1}^2(x,y) \, dy = \int_{\mathbb{R}^d} 2^{2d(J+1)} K^2(2^{J+1}x, 2^{J+1}y) \, dy \leq 2^{2d(J+1)} \int_{\mathbb{R}^d} F^2(2^{J+1}(x - y)) \, dy = 2^{(d+1)J} \int_{\mathbb{R}^d} F^2(v) \, dv,$$

where Assumption 4.1 justifies the inequality. It follows

$$\text{Var}(\hat{g}_J(x)) \leq \text{constant} \times n^{-1} k_3 \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 \left( 2^{dJ} \int_{\mathbb{R}^d} F^2(v) \, dv + O(n^{-1/d}) \right),$$

which can be integrated over the compact $C$:

$$\mathbb{E}(\|\hat{g}_J - \mathbb{E}(\hat{g}_J)\|_2^2) = \int_{\mathbb{R}^d} \text{Var}(\hat{g}_J(x)) \, dx \leq \text{constant} \times n^{-1} k_3 \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 \left( 2^{dJ} \int_{\mathbb{R}^d} F^2(v) \, dv + O(n^{-1/d}) \right).$$

Hence, for $n$ large enough, there exists a constant $\kappa_3' < \infty$ such that

$$\mathbb{E}(\|\hat{g}_J - \mathbb{E}(\hat{g}_J)\|_2^2) < \kappa_3' n^{-1} k_3 \left( \frac{\Gamma(k)}{\Gamma(k + 1/2)} \right)^2 2^{dJ}.$$  \hfill (A.9)

Plugging (A.8) and (A.9) in (A.7) yields the result. \hfill \qed

References

Agarwal, R., Chen, Z. and Sarma, S.V. (2017), A novel nonparametric maximum likelihood estimator for probability density functions, IEEE Trans. Pattern Anal. Mach. Intell., 39, 1294-1308.

Bochner, S., Harmonic analysis and the theory of probability, University of California Press, 1955.

Cosma, A., Scaillet, O. and von Sachs, R. (2007), Multivariate wavelet-based shape-preserving estimation for dependent observations, Bernoulli, 13, 301-329.

Daubechies, I., Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.
Dechevsky, L. and Penev, S. (1997), On shape preserving probabilistic wavelet approximators, Stochastic Anal. Appl., 15, 187-215.

Dechevsky, L. and Penev, S. (1998), On shape preserving wavelet estimators of cumulative distribution functions and densities, Stochastic Anal. Appl., 16, 423-462.

Delyon, B. and Juditsky, A. (1996), On minimax wavelet estimators, Appl. Comput. Harmon. Anal., 3, 215-228.

Donoho, D.L. and Johnstone, I.M. (1994), Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81, 425-455.

Donoho, D.L. and Johnstone, I.M. (1995), Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc., 90, 1200-1224.

Donoho, D.L. and Johnstone, I.M. (1996), Neo-classical minimax problems, thresholding, and adaptive function estimation, Bernoulli, 2, 39-62.

Donoho, D.L. and Johnstone, I.M. (1998), Minimax estimation via wavelet shrinkage, Ann. Statist., 26, 879-921.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995), Wavelet shrinkage: Asymptopia? J. R. Stat. Soc. Ser. B Stat. Methodol., 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.

Evans, D. (2008), A law of large numbers for nearest neighbour statistics, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 464, 3175-3192.

Evans, D., Jones, A.J. and Schmidt, W.M. (2002), Asymptotic moments of near-neighbour distance distributions, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 458, 2839-2849.

Fan, J., Hall, P., Martin, M. and Patil, P. (1996), On the local smoothing of nonparametric curve estimators, J. Amer. Statist. Assoc., 91, 258-266.

Geenens, G. (2011), Curse of Dimensionality and related issues in nonparametric functional regression, Statistics Surveys, 5, 30-43.

Hall, P. (1983), On near neighbour estimates of a multivariate density, J. Multivariate Anal., 13, 24-39.
Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A., Wavelets, Approximation and Statistical Applications, Lecture Notes in Statistics, Springer, New York, 1998.

Hardy, M. (2006), Combinatorics of partial derivatives, Electron. J. Combin., 13, Research paper R1.

Hyndman, R.J. (1996), Computing and graphing highest density regions, Amer. Statist., 50, 120-126.

Johnson, N.L., Kotz, S. and Balakrishnan, N., Continuous Univariate Distributions (Volume 1), John Wiley and Sons, New York, 1994.

Kerkyacharian, G. and Picard, D. (1993), Density estimation by kernel and wavelet methods: optimality of Besov spaces, Statist. Probab. Lett., 18, 327-336.

Mack, Y.P. and Rosenblatt, M. (1979), Multivariate k-nearest neighbor density estimates, J. Multivariate Anal., 9, 1-15.

Marron, S. and Wand, M. (1992), Exact Mean Integrated Squared Error, Ann. Statist., 20, 712-736.

McFadden, D. (2003), Economic choices. In T. Persson (ed.), Nobel Lectures in Economic Sciences 1996-2000, pp. 330-365, World Scientific, Singapore.

Meyer, Y., Wavelets and Operators, Cambridge University Press, 1992.

Nason, G., Wavelet methods in statistics with R, Use R! Series, Springer Science & Business Media, 2008.

Omohundro, S.M., Five balltree construction algorithms, Berkeley: International Computer Science Institute, 1989.

Penev, S. and Dechevsky, L. (1997), On non-negative wavelet-based density estimators, J. Nonparametr. Stat., 7, 365-394.

Percus, A.G. and Martin, O.C. (1998), Scaling universalities of kth nearest neighbor distances on closed manifolds, Adv. in Appl. Math., 21, 424-436.

Peter, A.M. and Rangarajan, A. (2008), Maximum Likelihood Wavelet Density Estimation With Applications to Image and Shape Matching, IEEE Trans. Image Process., 17, 458-468.

Peter, A.M., Rangarajan, A. and Moyou, M. (2017), The Geometry of Orthogonal-Series, Square-Root Density Estimators: Applications in Computer Vision and Model Selection, In: Computational Information Geometry for Image and Signal Processing, Nielsen, F., Critchley, F. and Dodson, C.T.J. (Eds), Springer International Publishing, pp. 175-215.
Pinheiro, A. and Vidakovic, B. (1997), Estimating the square root of a density via compactly supported wavelets, Comput. Statist. Data Anal., 25, 399-415.

Ranneby, B., Jammalamadaka, S.R. and Teterukovskiy, A. (2005), The maximum spacing estimation for multivariate observations, J. Statist. Plann. Inference, 129, 427-446.

Silverman, B.W., Density Estimation for Statistics and Data Analysis, Chapman and Hall/CRC, 1986.

Stone, C.J. (1982), Optimal global rates of convergence for nonparametric regression, Ann. Statist., 10, 1040-1053.

Strang, G. (1989), Wavelets and dilation equations: a brief introduction, SIAM Review, 31, 614-627.

Strang, G. (1993), Wavelet transforms versus Fourier transforms, Bulletin Amer. Math. Soc., 28, 288-305.

Triebel, H. Theory of Function Spaces II, Birkhuser Verlag, Basel, 1992.

Vannucci, M. (1995), Nonparametric density estimation using wavelets, Discussion Paper 95-26, Institute of Statistics and Decision Sciences, Duke University.

Vidakovic, B., Statistical Modeling by Wavelets, Wiley, New York, 1999.