Econometric Theory, 2019, Page 1 of 42.
doi:10.1017/S0266466619000112

NONPARAMETRIC DENSITY ESTIMATION BY B-SPLINE DUALITY

ZHENYU CUI
Stevens Institute of Technology

JUSTIN LARS KIRKBY
Georgia Institute of Technology

DUY NGUYEN
Marist College

In this article, we propose a new nonparametric density estimator derived from the theory of frames and Riesz bases. In particular, we propose the so-called *bi-orthogonal density estimator* based on the class of B-splines and derive its theoretical properties, including the asymptotically optimal choice of bandwidth. Detailed theoretical analysis and comparisons of our estimator with existing local basis and kernel density estimators are presented. The estimator is particularly well suited for high-frequency data analysis in financial and economic markets.

1. INTRODUCTION

Nonparametric density estimation is an enduring and ever-growing research area with vast applications across the physical and social sciences. Since the 1950's, several powerful approaches have emerged to improve the performance of probability density estimation, beyond the simple histogram representation. The first of these so-called *nonparametric* approaches can be found in Fix and Hodges (1951), Rosenblatt (1956), Parzen (1962), and Cencov (1962), and early reviews of the field are provided by Wegman (1972) and Izenman (1991). The literature on density estimation has steadfastly evolved in response to the growing applications of such techniques. While kernel density estimators prevail as the principal estimation approach, alternatives such as orthogonal sequence estimators have also received appreciable attention, including Schwartz (1967), Watson (1969), Hall (1981), Wahba (1981), and Hall (1987), and more recently Leitao, Oosterlee, Ortiz-Gracia, and Bohte (2018). While orthogonal sequence estimators generally rely on a global basis expansion of the (unknown) density, local density estimators using uniform B-splines and wavelets have also been considered in Redner (1999), Donoho, Johnstone, Kerkyacharian, and Picard (1996), Peter and Rangarajan (2008), Penev and Dechevsky (1997), and Huang (1999). Related

Address correspondence to J. Lars Kirkby, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30318, USA; e-mail: jkirkby3@gatech.edu.

techniques, such as smoothing splines in Gu and Qiu (1993) and Gu (1993), penalized B-splines (P-splines) in Eilers and Marx (1996), and logsplines in Kooperberg and Stone (1991) and Koo (1996), have also been considered. These techniques minimize the least square projection error onto a basis, subject to a smoothing/regularization penalty.

In nonparametric density estimation, a key challenge for many approaches is to determine the optimal bandwidth (or knot-positioning) to balance bias and variance.[1] Representative literature includes Hardle and Marron (1985), Sheather and Jones (1991), Jones, Marron, and Sheather (1996a). Cross-validation is commonly applied to determine the bandwidth, due to its generality (see Marron, 1987; Rudemo, 1982; Hall, 1987; Duin, 1976; Gu and Wang, 2003). A comparison of the merits of cross-validation and so-called *plug-in* methods is provided in Loader (1999). More recent progress in bandwidth selection includes Botev, Grotowski, and Kroese (2010), Figueroa-López and Li (2016). It has been observed that the appropriate bandwidth rule can vary depending on the application at hand and the availability of data, see Loader (1999), Zhang, Brooks, and King (2009), Carroll, Delaigle, and Hall (2013), Racine and Li (2017), so the availability of several complementary methods of selection is advantageous.

The application of nonparametric density estimators is widespread and growing in finance and financial econometrics. These estimators are often introduced to estimate transition densities (see Aït-Sahalia, 1996), the diffusion matrix (see Bandi and Moloche, 2017), state price densities (see Zhang et al., 2009; Song and Xiu, 2016; Beare and Schmidt, 2014), realized variance and stock volatilities (see Van Es, Spreij, and Zanten, 2003; Zu, 2015), implied volatilities (see Chen and Xu, 2014), as well as bond yields. In quantitative risk management, density estimates enable the calculation of risk measures, such as value-at-risk and expected shortfall (see Wang and Zhao, 2016; Cai and Wang, 2008; Opschoor, Dijk, and van der Wel, 2017). For a review of the literature pertaining to financial econometrics, please refer to Fan (2005) and references therein.

The present work extends the existing literature and methodology to include a new class of density estimators, the so-called *bi-orthogonal density estimators*, which utilize the duality theory of frames[2] to derive basis coefficient estimates. While our focus is on a class B-spline density estimator, the methodology holds in greater generality and can be applied more broadly to frames in a Hilbert space, where orthonormality of a complete sequence is replaced by a broader construct. Density approximation using B-splines has proven successful in the context of option pricing using parametric models for which the characteristic function is known analytically (see Kirkby, 2015; Cui, Kirkby, and Nguyen, 2017b; Kirkby, Nguyen, and Cui, 2017). The present work presents a generalization of this idea to

[1] For wavelets, the analogous choice is one of thresholding the wavelet coefficients. Similarly, for orthogonal sequences estimators, the number of basis elements in the approximation is of crucial importance.

[2] Please refer to the book Christensen (2003) and references therein for a thorough discussion of the subject on frames and Riesz bases.

the random sample nonparametric setting. In the current work, our attention is focused on the univariate case, with multivariate extensions left to future research.[3] We present the theoretical framework of bi-orthogonal density estimation, detail the computation of basis coefficients, and analyze theoretically the choice of the optimal bandwidth, which is of utmost importance for the performance of any nonparametric kernel-based estimator.

Compared with certain orthonormal basis estimators, such as wavelets, B-splines have several practical advantages due to their tractability. In particular, postestimation analysis is simplified by the fact that many closed-form estimates can be computed from the estimated density, such as expected shortfall measures and other functionals. The nonparametric CDF admits a closed-form using the B-spline basis, as well as its inverse, which can be used to easily calculate quantiles, estimate the value-at-risk, and simulate from the nonparametric density. Bi-orthogonality provides the advantages of an orthonormal basis, but with less complexity than traditional wavelets, which tend to be less tractable for postestimation analysis.

One of the disadvantages of (bounded support) kernel density estimators is that they scale the full data set at a rate $N\log(N)$, see Hayfield et al. (2008), thus rendering subsequent computational tasks more and more burdensome as the size of available data increases. In contrast, local bases can provide a data reduction that reduces the computational burden of inference and analysis after the estimation process. The proposed bi-orthogonal density estimators perform this reduction by encapsulating the sample information in a much smaller set of coefficients, the size of which grows at the rate of $N^{\frac{1}{2p+3}}$ for a pth order B-spline basis. With high-frequency data, as is commonly generated in financial markets, the data reduction is substantial, especially when several subsequent calculations are required involving the density estimate.

Across a broad range of tests, our use of bi-orthogonality to calculate the density estimator substantially improves the estimator's performance for local bases, as compared with the approach taken in Redner (1999), which differs in its computation of basis coefficients. It also outperforms the traditional and the cutting-edge kernel estimator proposed in Botev et al. (2010), as well as the log-spline approach of Kooperberg and Stone (1991). In Hall and Racine (2015), the authors prove that by increasing the polynomial order for local polynomial kernel estimation,[4] the optimal convergence rate N^{-1} can be obtained. In a similar vein, we prove that by increasing the B-spline order, our method can arbitrarily approach the optimal convergence rate of N^{-1}. In cases that are most relevant to empirically observed financial data, which are generally characterized by a peaked unimodal

[3] In particular, the methods presented here can be naturally extended to the multivariate case using tensor product bases.

[4] Thanks to the anonymous referee for raising this point, which motivates us to investigate this issue. Note that Hall and Racine (2015) establish the result for regressions, but mainly for local polynomial regressions. As for our case, the result is carried through with obvious modifications.

density, and also in challenging multimodal test cases, our method demonstrates consistently improved performance. Moreover, we provide a coefficient estimation method in the frequency space which facilitates the application of spectral filtering for variance reduction.

To summarize, the main contributions of the current article to the literature are two-fold:

1. We propose a novel nonparametric density estimator based on bi-orthogonality of Riesz bases. We derive the theoretical properties for the proposed estimator, as well as optimal bandwidth selection rules, and several competing approaches to bandwidth selection in the literature are compared in our setting.

2. We demonstrate through extensive numerical examples that our estimator substantially outperforms representative methods including Redner (1999) and Botev et al. (2010), especially as the sample size increases. It has the attractive features of *local coefficient estimation*, which provides particularly suitable tools for inference involving high-frequency data.

The article is organized as follows: Sections 1.1, 1.2, and 1.3 lay out the basics of delta sequence estimators and the projection theory, based on which we create our estimator. Section 2 proposes the *bi-orthogonal density estimator* and analyzes in detail its theoretical properties, including the coefficient estimation and estimates of the integrated squared bias and variance. We also derive the asymptotic mean integrated squared error (MISE), which is useful for plug-in style bandwidth selection and asymptotic error analysis. Section 3 discusses in detail the bandwidth selection for our proposed estimator and compares several approaches. Section 4 presents the empirical characteristic function (ECF) approach to estimate the projection coefficients and discusses the spectral filtering technique of the ECF. We also compare our estimator with existing estimators. Section 6 concludes the article with future research directions. Some technical and implementation details are collected in the Appendix.

1.1. Delta Sequence Estimators

Consider a set of i.i.d data, $\{X_n\}_{n=1}^N$, where X_n is governed by the probability density function f, which is unknown. Let $\delta_\lambda(x, y)$ be a bounded function of $x, y \in \mathbb{R}$, with a smoothing parameter $\lambda > 0$. We call $\{\delta_\lambda(x, y)\}_\lambda$ a *delta sequence* if, for every $g \in C^\infty(\mathbb{R})$, $\int_{-\infty}^{\infty} \delta_\lambda(x, y) g(y) dy \to g(x)$ as $\lambda \to \infty$. With the sample size N fixed, we formulate the corresponding *delta sequence estimator*, $\widetilde{f}_\lambda(x) = \frac{1}{N} \sum_{n=1}^N \delta_\lambda(x, X_n)$. The general theory of delta sequence estimators is investigated in the works of Woodroofe (1970), Walter and Blum (1979), and Terrell and Scott (1992). A special case is the traditional approach to nonparametric density estimation by means of a kernel function, where for some $K(v) : \mathbb{R} \to \mathbb{R}$ with $\int_{-\infty}^{\infty} K(v) dv = 1$, we define $\delta_h(x, X_n) := \frac{1}{h} K((x - X_n)/h)$. This yields the

kernel density estimator:

$$\widetilde{f_h}(x) = \frac{1}{hN}\sum_{n=1}^{N} K\left(\frac{x - X_n}{h}\right). \tag{1}$$

Common examples are the Gaussian kernel $K(\nu) = (2\pi)^{-1/2}e^{-\nu^2/2}$ and the polynomial kernels $K(\nu) = \kappa_{r,s}(1 - |\nu|^r)^s \mathbb{1}_{[|x|\le 1]}$ for a constant $\kappa_{r,s}$, $r > 0, s \ge 0$, discussed in equation (4.4) of Izenman (1991).

Alternatively, many authors have considered an *orthogonal series estimator*, which is defined in terms of an orthonormal basis $\{g_k\}_{k\in\mathbb{Z}}$ for $L^2(\mathbb{E})$, where[5] $\mathbb{E} \subset \mathbb{R}$. Let $\bar{z}$ denote the complex conjugate of the complex number z. If we define $\delta_M(x, X_n)$ by $\sum_{k=-M}^{M} g_k(x)\overline{g_k(X_n)}$, then we can formulate the estimator

$$\widetilde{f_M}(x) = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=-M}^{M} g_k(x)\overline{g_k(X_n)} = \sum_{k=-M}^{M} \alpha_k g_k(x), \tag{2}$$

where $\alpha_k := \frac{1}{N}\sum_{n=1}^{N}\overline{g_k(X_n)}$ is an unbiased estimate of the projection coefficient. The Hermite function approach of Schwartz (1967) is a prominent example with $\mathbb{E} = \mathbb{R}$. The basis is indexed by $\mathbb{N}$, with $g_k(x) := (2^k k\pi^{1/2})^{-1/2}e^{-x^2/2}H_k(x)$, where $H_k(x) := (-1)^k e^{x^2}(d^k/dx^k)(e^{-x^2})$. Similarly, Schwartz (1967), Watson (1969), and Leitao et al. (2018) study Harmonic series.

1.2. Basis Duality and Projection

While splines have been considered in Redner (1999) and Gehringer and Redner (1992), the present treatment is distinguished by our use of *frame duality* to obtain L^2 optimal projections of the target density. As we will show, this leads to superior performance compared with traditional approaches. In place of orthogonal sequences which yield the estimator in (2), we consider *Riesz sequences*, which admit *bi-orthogonal duals* to prescribe how to obtain the orthogonal projection of f. This framework encompasses a rich class of bases, including B-splines and even traditional orthogonal bases.

Given a function $\varphi \in L^2(\mathbb{R})$ named the *generator*[6] and a *resolution* $a > 0$ that defines the *bandwidth* $h := 1/a$, we construct a sequence on the support of $f \in L^2(\mathbb{R})$ as[7]

$$x_k = x_1 + (k-1)h, \quad k \in \mathbb{Z}, \tag{3}$$

where x_1 is a shift parameter determined below. We then form the sequence $\{\varphi_{a,k}(x)\}_{k\in\mathbb{Z}} := a^{1/2}\{\varphi(a(x - x_k))\}_{k\in\mathbb{Z}}$, which generates an approximation space

[5] If $\mathbb{E}$ is a strict subset of the support of f, then there will be a bias introduced from the truncation error.

[6] The generators considered in this article, $\varphi^{[p]}$, see Section 1.3, are compactly supported, but this is not a requirement in general. B-splines in particular offer a convenient basis with nice convergence properties, see Kirkby (2015, 2016).

[7] Note that if f is a density, with $\|f\|_\infty < \infty$, then $\|f\|_2^2 \le \|f\|_1\|f\|_\infty = \|f\|_\infty < \infty$, so $f \in L^2(\mathbb{R})$.

$\mathcal{M}_a := \overline{\text{span}}\{\varphi_{a,k}\}_{k\in\mathbb{Z}}$.[8] We require that φ is real valued, symmetric about the origin, and satisfies the frame bounds:

$$A\|g\|_2^2 \le \sum_{k\in\mathbb{Z}} |\langle g, \varphi_{a,k}\rangle|^2 \le B\|g\|_2^2, \quad \forall g \in L^2(\mathbb{R}), \tag{4}$$

for some $0 < A \le B$, where $\langle g_1, g_2\rangle := \int_{\mathbb{R}} g_1(x)\overline{g_2(x)}dx$, and $\|g\|_2^2 := \langle g, g\rangle$.

The duality theory of Riesz bases (see Christensen, 2003; Heil, 2011; Young, 1980) guarantees the existence of a *dual generator* $\widetilde{\varphi} \in L^2(\mathbb{R})$ such that the orthogonal projection of any $g \in L^2(\mathbb{R})$ onto $\mathcal{M}_a$, denoted by $P_{\mathcal{M}_a} g(y)$, satisfies

$$P_{\mathcal{M}_a} g(y) := \sum_{k\in\mathbb{Z}} \left(\int_{-\infty}^{\infty} g(x)\overline{\widetilde{\varphi}_{a,k}(x)}dx \right) \varphi_{a,k}(y) = \sum_{k\in\mathbb{Z}} \beta_{a,k} \cdot \varphi_{a,k}(y), \tag{5}$$

where[9] we can express $\beta_{a,k} := \mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)]$ if $g(x)$ is a density, and $\sum_{k\in\mathbb{Z}} \beta_{a,k}^2 < \infty$. Moreover, $\{\widetilde{\varphi}_{a,k}(x)\}_{k\in\mathbb{Z}} := \{a^{1/2}\widetilde{\varphi}(a(x - x_k))\}_{k\in\mathbb{Z}}$ is also a Riesz basis for $\mathcal{M}_a$, called the *bi-orthogonal basis* due to the fact that $\langle \varphi_{a,k}, \widetilde{\varphi}_{a,j}\rangle = \delta_{k,j}$. The orthogonal sequence estimators are actually a special case for which $\varphi \equiv \widetilde{\varphi}$; thus, our proposed framework can be seen as a generalization of the traditional (orthogonal) approaches to nonparametric density estimations.

1.3. Compactly Supported Generators and B-splines

This article considers bi-orthogonal density projection onto bases formed by compactly supported generators, which can be characterized by their order of overlap, as follows.

DEFINITION 1. *For a symmetric, compactly supported generator $\varphi : \mathbb{R} \to \mathbb{R}$, we refer to φ as a p-th order generator if $p := \min\{k : supp(\varphi(\cdot)) \cap supp(\varphi(\cdot + k + 1)) = \emptyset\}$, where $supp(\varphi)$ is the interior of the support of φ.*

Of particular interests to our study are the B-spline generators (scaling functions), $\varphi^{[p]}$ for $p \ge 0$, where $\varphi^{[p]}$ will be defined by the relationship in (6) below. These are localized functions that are capable of capturing the fine details of a probability density function, and are especially useful when the density is characterized by an extreme peak as is common with high-frequency data, which is mostly unimodal.[10] Moreover, a local basis is very efficient to evaluate. When $p = 0$, the Haar generator is defined by $\varphi^{[0]}(y) := \mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(y)$, which produces density estimates with a step-like structure, similar to a histogram estimator. Moreover, the Haar basis is actually orthonormal and fits into the more traditional framework. Higher order B-spline generators are derived recursively by the convolution:

[8] Here, $\overline{\text{span}}\{\varphi_{a,k}\}_{k\in\mathbb{Z}}$ is the closure of the span of the basis elements $\varphi_{a,k}$, over $k \in \mathbb{Z}$ in $L^2(\mathbb{R})$.

[9] The conjugate has been dropped by assuming that φ is real-valued and symmetric.

[10] This framework easily accommodates alternative bases, such as those formed from standard window filter functions, which we leave as interesting extensions for future research.

$$\varphi^{[p]}(x) = \varphi^{[0]} \star \varphi^{[p-1]}(x) = \int_{-\infty}^{\infty} \varphi^{[p-1]}(y-x)\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}(y)dy. \tag{6}$$

For example, convoluting the Haar generator with itself yields the linear generator ($p = 1$):

$$\varphi^{[1]}(y) = \begin{cases} 1+y, & y \in [-1,0], \\ 1-y, & y \in [0,1], \end{cases}$$

and with the cubic scaling function obtained after two more convolutions. The orthogonal projection is given by (5), where the dual scaling function is guaranteed to exist for any B-spline basis.[11] A useful property of any B-spline generator is that φ provides a *partition of unity*, that is, $\sum_{k\in\mathbb{Z}} \varphi(x-k) = \sum_{k\in k(x)} \varphi(x-k) = 1$, where $k(x) := \{k \in \mathbb{Z} : \varphi(x-k) > 0\}$. Although this property is not strictly required, it simplifies some of the results that follow and will be assumed below. While our primary focus in this article is on B-spline generators, the theoretical results presented here apply to a larger class of generators.

2. PROJECTED DENSITY ESTIMATOR

In this section, we define the *bi-orthogonal density estimator* and investigate its theoretical properties under the following assumption.

Assumption 2. We assume throughout that $\varphi : \mathbb{R} \to \mathbb{R}$ is a symmetric, compactly supported generator of a Riesz basis, and a partition of unity, with a bounded dual generator, $\widetilde{\varphi} : \mathbb{R} \to \mathbb{R}$. Assume further that $f : \mathbb{R} \to \mathbb{R}$ is a density satisfying $\|f\|_\infty < \infty$.

We first suppose that an expression for the dual $\widetilde{\varphi}$ is known, and later present approximation methods that do not require an explicit representation of the dual.[12] The bi-orthogonal (projected) density estimator of the density function f, $\bar{f}^a(x; N)$, is defined by

$$\bar{f}^a(x;N) := \sum_{k\in\mathbb{Z}} \frac{1}{N} \sum_{1\le n\le N} \widetilde{\varphi}_{a,k}(X_n)\varphi_{a,k}(x) = \sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}(N)\varphi_{a,k}(x), \tag{7}$$

where

$$\bar{\beta}_{a,k}(N) := \frac{1}{N} \sum_{1\le n\le N} \widetilde{\varphi}_{a,k}(X_n), \tag{8}$$

[11] As the B-spline order increases, so does the smoothness of the density approximations. Higher order bases are able to capture a smooth density with great accuracy, even at low resolutions. On the other hand, the linear B-splines are very well suited for peaked densities, such as those for financial returns data over a very short time interval.

[12] As discussed in Section 1.2, the biorthgonal dual $\widetilde{\varphi}$ is guaranteed to exist for any Riesz basis, and in particular for those constructed with a B-spline generator, $\varphi = \varphi^{[p]}$.

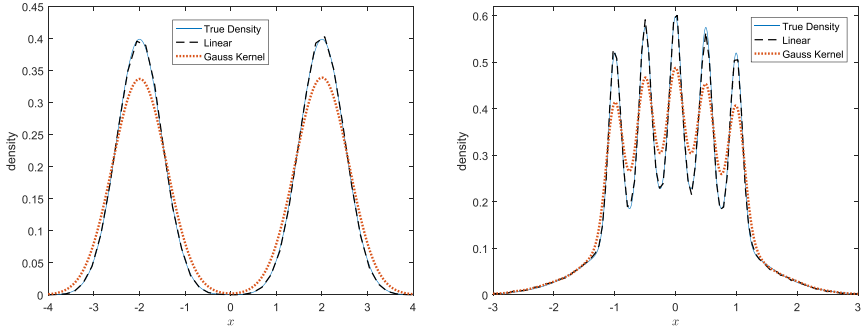

FIGURE 1. Linear bi-orthogonal estimator vs. Gaussian kernel. Comparison for separated bimodal (Left) and claw (Right) data from Table 5, with $N = 10^5$.

is an unbiased estimator of the coefficient $\beta_{a,k} := \mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)]$. Note that for k fixed, $\{\widetilde{\varphi}_{a,k}(X_n)\}_{n=1}^N$ is a sequence of i.i.d random variables. It immediately follows that

$$\mathbb{E}\big[\bar{f}^a(x;N)\big] = \sum_{k\in\mathbb{Z}} \beta_{a,k}\varphi_{a,k}(x) = P_{\mathcal{M}_a} f(x) \in L^2(\mathbb{R}),$$

so that $\bar{f}^a(x;N)$ is an unbiased estimator of the true orthogonal projection $P_{\mathcal{M}_a} f(x)$ in (5). We also note that $\bar{f}^a(x;N)$ can be expressed via the following reproducing kernel representation:

$$\bar{f}^a(x;N) = \frac{1}{N}\sum_{1\le n\le N}\left(\sum_{k\in\mathbb{Z}} \varphi_{a,k}(x)\widetilde{\varphi}_{a,k}(X_n)\right) = \frac{1}{hN}\sum_{1\le n\le N} K\left(\frac{x}{h}, \frac{X_n}{h}\right), \qquad \textbf{(9)}$$

which is reminiscent of (1), where[13]

$$K(x,y) := \sum_{k\in\mathbb{Z}} \varphi(x-k)\widetilde{\varphi}(y-k). \qquad \textbf{(10)}$$

It also holds that $\int \bar{f}^a(x;N)dx = 1$, which is established in Lemma A.3 in the Appendix.

Before diving into the theoretical development, we provide a motivating example of the linear bi-orthogonal density estimator in Figure 1. The left panel of Figure 1 illustrates a *separated* bi-modal mixture density, defined by $(1/2)\mathcal{N}(-2,1/4) + (1/2)\mathcal{N}(2,1/4)$, while the right panel illustrates a claw mixture, $\frac{1}{2}\mathcal{N}(0,1) + \sum_{k=0}^{4}\frac{1}{10}\mathcal{N}(\frac{k}{2}-1, (\frac{1}{10})^2)$. We have also included a Gaussian kernel density estimate, implemented with the standard Matlab package "ksdensity". The figures illustrate the outcome for one of 500 simulations, for which the

[13] For simplicity, in the theoretical treatment we will assume that $x_k = kh$, that is $x_1 = h$.

mean integrated squared error (MISE) of the linear bi-orthogonal estimator is a fraction of that of the kernel estimator (1/100 in the first case, and 1/26 in the second).[14]

This example serves to illustrate two main points. First, compared with the standard packages in many statistical libraries, there is often considerable room for improvement through experimentation. Second, bandwidth selection is a crucial component of density estimation. In particular, with a more careful bandwidth choice, the performance of the Gaussian kernel can be significantly improved, for example by employing the method of Botev et al. (2010), which is discussed in Section 3.2. We will also show that even with a more careful choice of bandwidth for the Gaussian kernel, the proposed bi-orthogonal estimator very often produces improved estimates, as measured by the MISE.

2.1. Theoretical Properties

Next, we establish some fundamental properties of $\bar{f}^a(x;N)$, and the convergence of $\bar{f}^a(x;N)$ to the true projection, $P_{\mathcal{M}_a} f(x)$, under mild assumptions. Several additional results are contained in the Appendix.

PROPOSITION 2.1. *Suppose that Assumption 2 holds for a generator φ and a density f, and let $\{X_i\}_{i=1}^N \overset{iid}{\sim} f$ be a sample. Then, for any $a = 1/h > 0$, the following statements hold*:

(*i*) *For any $x \in \mathbb{R}$, $\bar{f}^a(x;N) \xrightarrow{a.s.} P_{\mathcal{M}_a} f(x)$, and $\bar{f}^a(x;N) \xrightarrow{L^1} P_{\mathcal{M}_a} f(x)$, as $N \to \infty$.*

(*ii*) *For any $x \in \mathbb{R}$,*

$$Var(\bar{f}^a(x;N)) = \frac{1}{N}\sum_{k\in\mathbb{Z}} \varphi_{a,k}(x)\Big(\sum_{0\le|m|\le p} \varphi_{a,k-m}(x)\cdot \widetilde{C}_{a,k,k-m}\Big), \qquad (11)$$

where $\widetilde{C}_{a,j,k} := Cov(\widetilde{\varphi}_{a,j}(X_1), \widetilde{\varphi}_{a,k}(X_1))$, and $\sup_{j,k\in\mathbb{Z}}\{\widetilde{C}_{a,j,k}\} \le \|f\|_\infty \|\widetilde{\varphi}\|_2^2 < \infty$.

(*iii*) *The estimator variance is uniformly bounded,*

$$\sup_{x\in\mathbb{R}}\{Var(\bar{f}^a(x;N))\} \le \frac{R(\widetilde{\varphi})}{Nh}\|f\|_\infty,$$

where $R(\widetilde{\varphi}) := \int \widetilde{\varphi}^2(x)dx = \|\widetilde{\varphi}\|_2^2 < \infty$.

(*iv*) *Let $\varphi = \varphi^{[p]}$ be a pth order B-spline generator, and assume that $f \in C_b^{\bar{p}}(\mathbb{R})$, where $\bar{p} := \max\{p+1, 4\}$. Moreover, suppose that $N \to \infty, h \to$*

[14] For the bimodal data, we estimated a MISE of 2.1658e-04 and 0.0202 for the linear estimator and Gaussian kernel respectively, a 100 fold improvement for the linear estimator. Similarly, for the claw data, we estimated an MISE of 0.0011 and 0.0288, which represents a 26 fold improvement for the linear estimator. The bandwidth selection method used for the linear estimator is a simple heuristic discussed in Section 3.1.2.

0, *and* $Nh \to \infty$. *Then,*

$$\frac{\bar{f}^a(x;N) - f(x) - \mu_p(h)}{\sqrt{Var\left(\bar{f}^a(x;N)\right)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{12}$$

where the bias $\mu_p(h)$ *satisfies* $\mu_p(h) \le \lambda_p \|f^{(p+1)}\|_\infty \cdot \mathcal{O}(h^{p+1})$, *and with* $K(x,y)$ *defined in* (10),

$$\lambda_p := \frac{1}{(p+1)!} \sup_{x\in\mathbb{R}} \left(\int |x-y|^{p+1} |K(x,y)| dy \right) < \infty,$$

which is a constant depending only on φ, $\widetilde{\varphi}$, *and* $Var(\bar{f}_a(x;N))$ *is given in* (11).

Proof. See Appendix. ■

2.1.1. Coefficient Variance. From Proposition (2.1), the point estimate variance satisfies

$$\mathrm{Var}(\bar{f}^a(x;N)) = \sum_{k\in\mathbb{Z}} \varphi_{a,k}(x) \left(\mathrm{Var}(\bar{\beta}_{a,k}(N)) \varphi_{a,k}(x) + \sum_{1\le|m|\le p} \varphi_{a,k-m}(x) \cdot \frac{\widetilde{C}_{a,k,k-m}}{N} \right),$$

where $\widetilde{C}_{a,j,k} := \mathrm{Cov}(\widetilde{\varphi}_{a,j}(X_1), \widetilde{\varphi}_{a,k}(X_1))$. To derive an expression for the mean integrated squared error (MISE) of $\bar{f}^a(x;N)$ over $\mathbb{R}$, we first estimate the coefficient variance $\mathrm{Var}(\bar{\beta}_{a,k}(N))$ under the following assumption, which amounts to strengthening the Assumption 2.

Assumption 3. We assume that $f \in C_b^4(\mathbb{R})$ is an absolutely continuous density, where $C_b^n(\mathbb{R})$ is the set of nth order continuously differentiable bounded functions with bounded derivatives. Further assume that any generator φ is a symmetric, compactly supported Riesz basis generator with dual generator $\widetilde{\varphi}$ having finite moments $m_i(\widetilde{\varphi}^j) := \int \widetilde{\varphi}^j(x) x^i dx < \infty$, $0 \le i \le 4$, $j = 1, 2$.

LEMMA 1. *Let* $f \in C_b^4(\mathbb{R})$ *be a density from which a sample* $\{X_i\}_{i=1}^N$ *is observed, and let Assumption 3 hold. Define the coefficient variance by*

$$\bar{\sigma}_{a,k}^2 := Var(\bar{\beta}_{a,k}(N)) = \frac{1}{N} Var(\widetilde{\varphi}_{a,k}(X_1)), \quad k \in \mathbb{Z}, \tag{13}$$

where $\bar{\beta}_{a,k}(N)$ *is defined in* (8). *Then for* $h = 1/a > 0$,

$$\bar{\sigma}_{a,k}^2 = \frac{1}{N} \left\{ f(x_k) R(\widetilde{\varphi}) - h \cdot (f(x_k) m_0(\widetilde{\varphi}))^2 + \frac{h^2}{2} f^{(2)}(x_k) m_2(\widetilde{\varphi}^2) + \mathcal{O}(h^3) \right\}, \tag{14}$$

where $R(g) := \int g^2(x) dx$ *is the* roughness *of* $g \in L^2(\mathbb{R})$.

Proof. See Appendix. ■

2.2. Integrated Variance and Bias

This section derives some theoretical error bounds for the integrated variance and bias. These results will in turn help in the derivation of the asymptotic mean integrated squared error.

2.2.1. Integrated Variance. We next derive estimates for the integrated squared bias and variance of the estimator $\bar{f}^a(x;N)$. In particular, this provides a bound on the MISE, defined as

$$\begin{aligned}\text{MISE} &:= \mathbb{E}\left[\int \left(\bar{f}^a(x;N) - f(x)\right)^2 dx\right] \\ &= \int \mathbb{E}\left[\left(\bar{f}^a(x;N) - \mathbb{E}[\bar{f}^a(x;N)]\right)^2\right] dx + \int \left(\mathbb{E}[\bar{f}^a(x;N)] - f(x)\right)^2 dx, \end{aligned} \tag{15}$$

which is the sum of the integrated variance and the integrated squared bias. We can estimate each of these terms separately, starting with the first term in (15):

$$\int \text{Var}(\bar{f}^a(x;N))dx := \int \mathbb{E}\left[\left(\bar{f}^a(x;N) - \mathbb{E}[\bar{f}^a(x;N)]\right)^2\right] dx.$$

Assumption 4. For each of $j = 1, 2, 3$, we assume that $f^{(j)}$ is absolutely continuous, and moreover that $f^{(j)} \in L^1(\mathbb{R})$, so that $\|f^{(j)}\|_1 < \infty$.

PROPOSITION 2.2. *Let $f \in C_b^4(\mathbb{R})$ be a density, and φ a pth order generator, which satisfy Assumptions 3 and 4. Then for the estimator $\bar{f}^a(x;N)$ defined in* (7)*, the following hold*:

(*i*) *With $\bar{\sigma}_{a,k}^2 := Var(\bar{\beta}_{a,k}(N))$, the integrated variance can be bounded by*

$$\int Var(\bar{f}^a(x;N))dx \le \theta_p \cdot \sum_{k\in\mathbb{Z}} \bar{\sigma}_{a,k}^2, \tag{16}$$

where $\theta_p := \|\varphi\|_2^2 + 4\sum_{1\le m\le p}\int \varphi(x)\varphi(x-m)dx$. For example, for the pth order B-splines, $\theta_0 = 1, \theta_1 = 4/3$, and $\theta_p \le 2$ for $p \ge 2$. Equation (16) *holds as an equality when $p = 0$.*

(*ii*) *The sum of coefficient variances satisfies*

$$\begin{aligned}\sum_{k\in\mathbb{Z}} \bar{\sigma}_{a,k}^2 = \frac{R(\widetilde{\varphi})}{N}\frac{1}{h}\left(1 + \mathcal{O}\left(h\|f'\|_1\right)\right) - \frac{(m_0(\widetilde{\varphi}))^2}{N}\left(R(f) + \mathcal{O}(h\|f\cdot f'\|_1)\right) \\ + \frac{m_2(\widetilde{\varphi^2})h}{2N}\left(\int f^{(2)}(x)dx + \mathcal{O}\left(h\|f^{(3)}\|_1\right)\right) + \mathcal{O}(h^2/N),\end{aligned}$$

for small $h > 0$.

Proof. See Appendix. ■

2.2.2. Integrated Squared Bias. We now derive an upper bound on the integrated squared bias, which is the second term in (15):

$$\int \text{Bias}^2(\bar{f}^a(x;N))dx := \int \left(\mathbb{E}[\bar{f}^a(x;N)] - f(x)\right)^2 dx.$$

PROPOSITION 2.3. *For a p-th order B-spline basis, suppose that $f : \mathbb{R} \to \mathbb{R}$ is a density for which the roughness of $f^{(p+1)}$ is finite, i.e., $R(f^{(p+1)}) = \|f^{(p+1)}\|_2^2 < \infty$. The integrated squared bias can be bounded for $h > 0$ by*

$$\int Bias^2(\bar{f}^a(x;N))dx \le C_p \cdot \|f^{(p+1)}\|_2^2 \cdot h^{2(p+1)}, \tag{17}$$

where C_p is a constant depending only on the generator $\varphi = \varphi^{[p]}$.

Proof. See Appendix. ■

Remark 1. While Proposition 2.3 provides assurance that the integrated squared bias is well controlled, we are also interested in its asymptotic behavior as h approaches zero. From Proposition 5.2 of Unser and Daubechies (1997), it holds as $h \to 0$, $\lim_{h\to 0}(\|P_{\mathcal{M}_a} f - f\|_2 / h^{p+1}) = \bar{C}_p^{1/2} \|f^{(p+1)}\|_2$, for a constant $\bar{C}_p$ depending only on p. In particular, $\bar{C}_p^{1/2}$ is given for orders $p = 0, 1, 2, 3$ in Table 1. Hence, for a small $h > 0$, we have

$$\int \left(\mathbb{E}[\bar{f}^a(x;N)] - f(x)\right)^2 dx = \|P_{\mathcal{M}_a} f - f\|_2^2 = \bar{C}_p \cdot \|f^{(p+1)}\|_2^2 \cdot h^{2(p+1)}. \tag{18}$$

Note that Proposition 2.3 demonstrates that the bias vanishes as $h \downarrow 0$. However, from Proposition 2.2, the leading term $\frac{R(\tilde{\varphi})}{N}\frac{1}{h}\left(1 + \mathcal{O}(h\|f'\|_1)\right)$ in the variance expression will diverge as $h \downarrow 0$. The next section provides a choice of h which balances these competing forces.

2.2.3. Asymptotic MISE. We next combine the results of Propositions 2.2 and 2.3 to obtain a bound for the MISE, which can be minimized with respect to the

TABLE 1. B-spline constants and asymptotically optimal bandwidth for normal data

p	h_p^*	$\bar{C}_p^{1/2}$	$R(\tilde{\varphi})$	θ_p
0	$\sigma\left(\frac{2\sqrt{\pi}}{\bar{C}_0}\right)^{1/3} N^{-1/3}$	0.288675	1	1
1	$\sigma\left(\frac{2}{3}\frac{\theta_1\sqrt{\pi}}{\bar{C}_1}R(\tilde{\varphi})\right)^{1/5} N^{-1/5}$	3.72678×10^{-2}	1.73205	4/3
2	$\sigma\left(\frac{4}{45}\frac{\theta_2\sqrt{\pi}}{\bar{C}_2}R(\tilde{\varphi})\right)^{1/7} N^{-1/7}$	5.75055×10^{-3}	2.84217	2
3	$\sigma\left(\frac{4}{105}\frac{\theta_3\sqrt{\pi}}{\bar{C}_3}R(\tilde{\varphi})\right)^{1/9} N^{-1/9}$	9.09241×10^{-4}	4.96473	2

bandwidth parameter h, as h approaches zero. The resulting asymptotic MISE (AMISE) is summarized as follows.

PROPOSITION 2.4. *For a p-th order B-spline basis, suppose that the roughness of $f^{(p+1)}$ is finite, i.e., $R(f^{(p+1)}) = \|f^{(p+1)}\|_2^2 < \infty$, and Assumptions 3 and 4 are satisfied. The following hold*:
(*i*) *The AMISE satisfies the following bound for small $h > 0$:*

$$\mathbb{E}\left[\int \left(\bar{f}^a(x;N) - f(x)\right)^2 dx\right] \leq \theta_p \sum_{k\in\mathbb{Z}} \bar{\sigma}_{a,k}^2 + \bar{C}_p \cdot \|f^{(p+1)}\|_2^2 \cdot h^{2(p+1)}. \tag{19}$$

Moreover, as $h \to 0$, $Nh \to \infty$ and $N \to \infty$, then $\bar{f}^a(x;N) \xrightarrow{L^2} f(x)$, from which $\bar{f}^a(x;N)$ is a consistent estimator of $f(x)$.
(*ii*) *The asymptotically optimal bandwidth with respect to* (19) *satisfies*

$$h_p^* = \left(\frac{\theta_p}{2(p+1)\bar{C}_p} \cdot \frac{R(\widetilde{\varphi})}{\|f^{(p+1)}\|_2^2} \cdot \frac{1}{N}\right)^{\frac{1}{2p+3}}, \tag{20}$$

where $\theta_p \leq 2$ is defined in Proposition 2.2, and $\bar{C}_p$ is defined in Remark 1.
(*iii*) *The corresponding optimal asymptotic mean integrated squared error is given by*

$$AMISE_p^* = (\alpha_p + 1)R(\widetilde{\varphi}) \cdot \left(\frac{\bar{C}_p \|f^{(p+1)}\|_2^2}{\alpha_p R(\widetilde{\varphi})}\right)^{\frac{1}{2p+3}} \cdot N^{-\frac{2p+2}{2p+3}},$$

where $\alpha_p := \theta_p/2(p+1)$.

Proof. See Appendix. ■

Remark 2. The problem of boundary bias for density estimation has been studied in Masri and Redner (2005), and the recent work[15] of Cattaneo, Jansson, and Ma (2017), Calonico, Cattaneo, and Farrell (2018), for example. In the present case, when $f^{(k)}$ is discontinuous for some $k \leq p+1$, the bias in (18) no longer applies. This is the case for an estimator on a bounded or semi-infinite domain, such as $[l,u]$ or $[l,\infty)$, for which $f^{(k)}(b) \neq 0$ for the two boundary values $b \in \{l,u\}$. As long as $f^{(k)}(b) = 0$ for $0 \leq k \leq p+1$, regularity extends to the full domain $\mathbb{R}$, and Proposition 2.4 holds. We find that in many cases, the proposed bi-orthogonal estimator performs well on a (semi) bounded domain, as illustrated in Figure 2 for the log-normal data, and in the experiments that follow. However, improved performance is expected by explicitly accounting for boundary bias. For the general case, we leave a careful treatment of the boundary effects for future investigations.

[15] In these works, the authors design nonparametric density estimators which do not require pre-binning or any other transformation of data while maintaining fully adaptive boundaries.

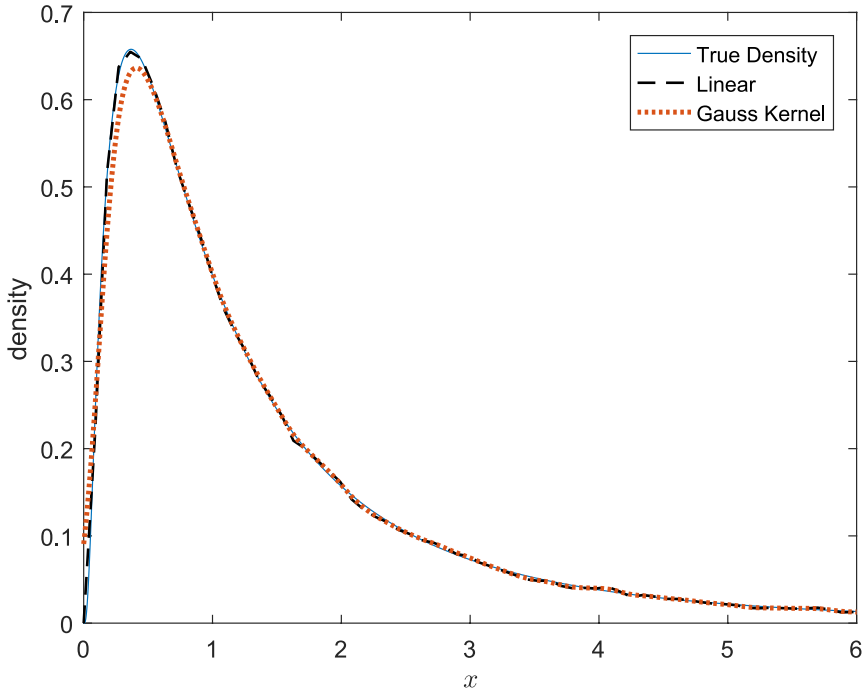

FIGURE 2. Linear bi-orthogonal estimator and Gaussian kernel with log-normal(0, 1) data, and $N = 10^5$.

Recall that for a standard histogram estimator, the AIMSE(h^*) converges at a rate of $\mathcal{O}(R(f')N^{-2/3})$ for a density $f \in C^1(\mathbb{R})$, while kernel density estimators achieve $\mathcal{O}(R(f'')N^{-4/5})$ with varying constants of convergence, for $f \in C^2(\mathbb{R})$. By comparison, Proposition 2.4 demonstrates a rate of $\mathcal{O}\left(R(f^{(p+1)})N^{-\frac{2p+2}{2p+3}}\right)$ for bi-orthogonal projection, again assuming sufficient regularity. For example, the linear basis with $p = 1$ converges at the same rate as kernel estimators, which is surpassed for $p \geq 2$. In contrast to the approach taken in Redner (1999) for B-splines (as well as in general for the kernel density estimators), for which the convergence rate is capped at $N^{-4/5}$, Proposition 2.4 demonstrates that for $f \in C^\infty(\mathbb{R})$, we can arbitrarily approach the optimal rate of $\mathcal{O}(N^{-1})$ by increasing the basis order, or equivalently the basis smoothness. This illustrates yet another advantage of the bi-orthogonal basis estimator.

Remark 3. We note that the bandwidth derived in Proposition 2.4 is based on θ_p in the bound $\int \mathrm{Var}(\bar{f}^a(x; N))dx \leq \theta_p \cdot \sum_{k\in\mathbb{Z}} \bar{\sigma}^2_{a,k}$. While the result is exactly optimal for the Haar basis ($p = 0$), we find that h^*_p tends to favor smoothness for $p \geq 1$, trading lower variance for higher bias. When supplemented with a spectral filter, as discussed in Section 4.1, in practice we may reduce the value of θ_p to

achieve more agile estimates with peaked densities, in order to capitalize on the locality of the B-spline bases.

To apply Proposition 2.4 in practice, note that $R(\widetilde{\varphi}) = \int \widetilde{\varphi}^2(x)dx = \frac{1}{2\pi}\int \widehat{\widetilde{\varphi}}^2(\xi)d\xi$, which can be computed analytically in certain cases, and numerically otherwise, since we know $\widehat{\widetilde{\varphi}}(\xi)$. Table 1 provides roughness values for the first four orders, and values for the other constants appearing in (20). Proposition 2.4 can also be used to help guide the selection of N. For example, given an error tolerance ϵ, we can choose $N = N(\epsilon)$ to make $AMISE_p^* < \epsilon$ by

$$N(\epsilon) > \left(\frac{(\alpha_p + 1)R(\widetilde{\varphi})}{\epsilon}\right)^{\frac{2p+3}{2p+2}} \left(\frac{\bar{C}_p \| f^{(p+1)} \|_2^2}{\alpha_p R(\widetilde{\varphi})}\right)^{\frac{1}{2p+2}}. \tag{21}$$

The only unknown term in (21) is $\| f^{(p+1)} \|_2^2$, and we next discuss three methods for estimating $\| f^{(p+1)} \|_2^2$. The first approach (Section 3.1) is based on a reference distribution, whereas the second approach is based on a kernel density estimate (Section 3.2).

3. BANDWIDTH SELECTION

The *proper* bandwidth selection is of utmost importance to the performance of the constructed nonparametric density estimator. In the literature, there have been significant research efforts in determining the *optimal* choice of the bandwidth. Early approaches can be found in Hardle and Marron (1985), Sheather and Jones (1991), Jones et al. (1996a), and references therein. In general, the appropriate selection rule is application dependent, and no single bandwidth approach is universally preferred.[16] In this section, we describe several bandwidth selection methods that can be utilized depending on the applications and data availability.

3.1. Reference Distribution Approach

3.1.1. Rule of Thumb for Normal Data. We begin by describing a simple plug-in rule based on the assumption of normality. To operationalize the bandwidth selection using h_p^* from (20), we require an estimate for $\| f^{(p+1)} \|_2^2$. Assuming normality, $\| f^{(p+1)} \|_2^2$ can be computed explicitly, resulting in a closed-form bandwidth as a function of the observed data. In this case, $f(x) \equiv \phi_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$, for $n \geq 0$, we can estimate $\| f^{(p+1)} \|_2^2 = \| \phi_\sigma^{(p+1)} \|_2^2$ using

[16] For example, Hall et al. (2005) highlight the influence on classification error when using nonparametric estimators to minimize the Bayes risk. For option Greeks determination, Liu and Hong (2009) and Liu and Hong (2011) document evidence of high sensitivity of the estimator to the choice of bandwidth. As demonstrated in Loader (1999), selection methods which over-smooth in some cases may also be susceptible to under-smoothing in others, and one should avoid blindly applying "asymptotically" optimal selection methods such as plug-in rules, even in large-sample settings.

$$\phi_\sigma^{(n)}(x) \equiv \frac{\partial^n \phi_\sigma}{\partial x^n}(x) = (-1)^n \frac{1}{(\sigma\sqrt{2})^n} H_n\left(\frac{x}{\sigma\sqrt{2}}\right)\phi_\sigma(x),$$

where $H_n(x)$ is the nth order Hermite polynomial given by $H_n(x) := (-1)^n e^{x^2} d^n/dx^n \left(e^{-x^2}\right)$.

We then have the following identity (see, for example, Wand and Jones, 1994, Fact C.1.12)

$$\int_{-\infty}^{\infty} \phi_\sigma^{(r)}(x-\mu)\phi_{\sigma'}^{(r')}(x-\mu')dx = (-1)^r \phi^{(r+r')}_{\sqrt{\sigma^2+(\sigma')^2}}(\mu-\mu').$$

By choosing $\sigma \equiv \sigma'$, $\mu = \mu' = 0$, we can compute $\|f^{(p+1)}\|_2^2$ using

$$\|f^{(p+1)}\|_2^2 = (-1)^{p+1}\phi^{2(p+1)}_{\sigma\sqrt{2}}(0) = (-1)^{3(p+1)}\frac{H_{2(p+1)}(0)}{\sqrt{\pi}\cdot(2\sigma)^{2p+3}}. \quad \textbf{(22)}$$

Table 1 summarizes the normal rule of thumb h^* for B-spline orders up to a cubic generator, which utilizes (22) as a plug-in for (20).

Figure 3 illustrates the convergence rate for the linear basis as a function of the sample size, N. The data is simulated from a normal distribution with several different values of σ, using the asymptotically optimal rule of thumb given in Table 1. The rate of convergence is estimated as the slope coefficient from a regression of $\log_{10}(RMSE)$ against $\log_{10}(N)$, for which the optimal slope is $-1/2$. For the linear basis, we observe a rate of about -0.48 (given in parentheses for each σ), which is nearly optimal. This is true despite the theoretical rate of -0.40 for the linear basis, a phenomenon which has been documented in related contexts (see Unser, 1996; Kirkby, 2015).

3.1.2. Robust Rule of Thumb and Heuristic. The normal rule of thumb bandwidth selection by itself is ill-suited for most practical applications, especially those arising in high-frequency financial applications. For unimodal data, Silverman (1986) proposed a *robust rule of thumb*, which replaces the standard deviation estimate $\hat{\sigma}$ by $\sigma^* := \min\left(\hat{\sigma}, \hat{R}/\zeta\right)$, where $\hat{R} = Q_3 - Q_1$ is the interquartile range, and $\zeta = 1.35$.[17] We will later illustrate a modified version of this approach, which defines $\zeta := 2.5 + (1 - \exp(-\kappa))$, where κ is the maximum of the excess kurtosis and zero. This modification not only protects against outliers but also performs relatively well for multimodal data, which is not the case for the original rule.[18] The new estimate σ^* is used together with (20) and (22) to determine the normal rule of thumb bandwidth, by replacing σ in Table 1 with σ^*.

Finally, we propose an additional heuristic which further protects against heavily leptokurtic distributions. In these cases, a common cause of failure for the

[17] For the normal distribution, $R = 1.35\sigma$.

[18] While it tends to produce smaller bandwidth estimates than the rule proposed by Silverman (1986), spectral filtering is utilized to placate the potentially increased variance, as discussed in Section 4.1.

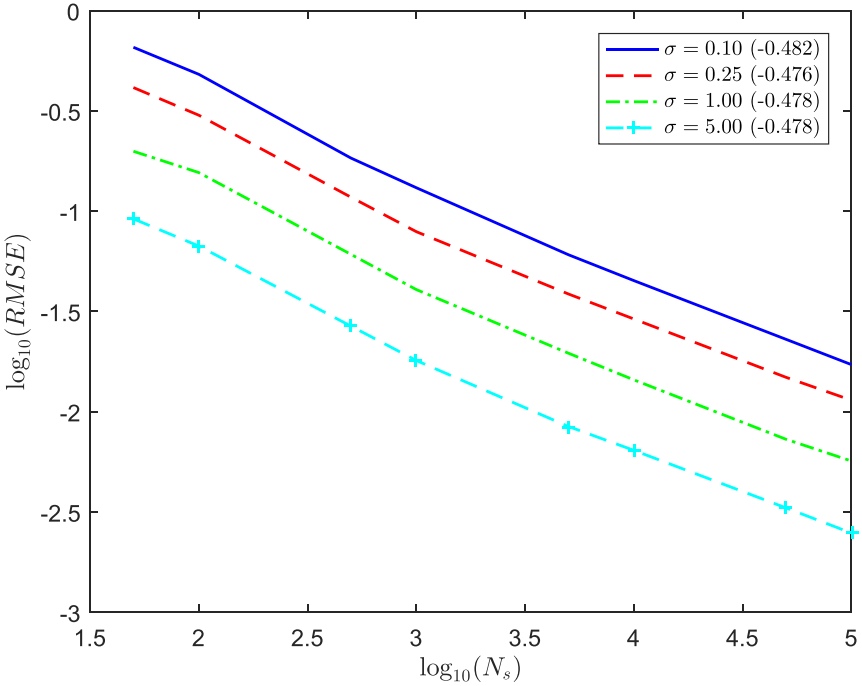

FIGURE 3. RMSE convergence for normal $\mathcal{N}(0, \sigma^2)$ data as function of sample size, $N_s = N$, with linear basis. Regression slope coefficient given in parentheses.

normal rule of thumb is that it places too much data in any one of the "bins" defined by $[x_k, x_{k+1}]$, such as those near a distribution's peak. Denote

$$p_k(h) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{[x_k, x_{k+1}]}(X_i), \quad p^*(h) := \max_{1 \le k \le N-1} \{p_k\},$$

with an initial bandwidth $h = x_{k+1} - x_k$ fixed. The maximal amount of mass of a normal $\mathcal{N}(\mu, \hat{\sigma}^2)$ density assignable to any bin of length h is given by a bin centered at μ, that is $[\mu - h/2, \mu + h/2]$. By symmetry (setting $\mu = 0$), under the normal rule of thumb, the corresponding maximum is therefore $p^{\hat{\sigma}}(h) := \Phi_{\hat{\sigma}}(h/2) - \Phi_{\hat{\sigma}}(-h/2)$, where $\Phi_{\hat{\sigma}}(x)$ is the CDF of a $\mathcal{N}(0, \hat{\sigma}^2)$ random variable. Prior to evaluating coefficients, if $p^*(h) > p^{\hat{\sigma}}(h)$, i.e., if we have allocated more mass to any bin than is theoretically prescribed under the normal rule of thumb, then we will decrease h to $h \leftarrow h \cdot (p^{\hat{\sigma}}(h)/p^*(h))$. This procedure is iterated until $p^*(h) \le p^{\hat{\sigma}}(h)$, that is until no bin $[x_k, x_{k+1}]$ contains a larger proportion of the sample data than would be maximally assigned. We find this approach to be a major improvement over the standard rule of thumb and provides an excellent initial

guess for the optimization-based cross-validation methods discussed in Section 3.3.

3.2. Plug-in Method

While the robust rule of thumb of Section 3.1.2 is simple and effective enough for many applications, a more careful estimate of $\|f^{(p)}\|_2^2$ can improve the estimation performance significantly, especially for multimodal data. These so-called *plug-in methods* often use an iterative approach to estimate $\|f^{(p)}\|_2^2$ (see Jones, Marron, and Sheather, 1996b; Sheather and Jones, 1991). A more recent approach is developed in Botev et al. (2010), which improves upon the ideas of Jones, Marron, and Park (1991), Jones et al. (1996b), and Marron (1985), and is discussed in the supplemental appendix. We illustrate the plug-in method combined with bi-orthogonal density estimation in the numerical experiments of Section 5.2, and find that it performs exceptionally well with multimodal data. For unimodal data, the rule of thumb heuristic from Section 3.1.2 performs similarly, and is less expensive to calculate.

3.3. Cross-validation and Comparison

Despite the simplistic allure of many plug-in rules, a careful application is advised especially in cases where the curvature of the true density is difficult to estimate, or when the smoothness conditions are likely violated. Moreover, while plug-in rules enjoy faster convergence they can be asymptotically inefficient when the assumptions are violated. A generally applicable approach to bandwidth selection is provided by cross-validation, which is detailed in Marron (1987), Hall (1987), Duin (1976), and Gu and Wang (2003). Several variations of this approach exist, some of which are tailored to particular estimators such as kernels. In the present context, likelihood-based cross-validation (see Duin, 1976 for more details) is ap-

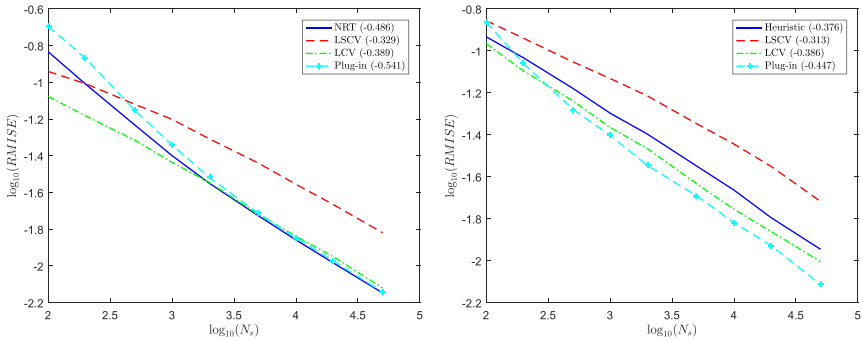

FIGURE 4. RMISE convergence comparison for four bandwidth selection methods as function of sample size, $N_S = N$, with linear basis. Left: normal $\mathcal{N}(0, 1)$ data. Right: skewed unimodal data. Regression slope coefficient given in parentheses.

TABLE 2. Test case densities for numerical experiments

Test case	Density	Test case	Density
Gaussian	$\mathcal{N}(0,1)$	Bimodal	$\frac{1}{2}\mathcal{N}(0,(\frac{1}{10})^2)+\frac{1}{2}\mathcal{N}(5,1)$
Student-t	t_ν, $\nu=6$	Separated bimodal	$\frac{1}{2}\mathcal{N}(-2,(\frac{1}{2})^2)+\frac{1}{2}\mathcal{N}(2,(\frac{1}{2})^2)$
Kurtotic unimodal	$\frac{2}{3}\mathcal{N}(0,1)+\frac{1}{3}\mathcal{N}(0,(\frac{1}{10})^2)$	Skewed bimodal	$\frac{3}{4}\mathcal{N}(0,1)+\frac{1}{4}\mathcal{N}(\frac{3}{2},(\frac{1}{3})^2)$
Skewed unimodal	$\frac{1}{5}\mathcal{N}(0,1)+\frac{1}{5}\mathcal{N}(\frac{1}{2},(\frac{2}{3})^2)+\frac{3}{5}\mathcal{N}(\frac{13}{12},(\frac{5}{9})^2)$	Trimodal	$\frac{1}{3}\sum_{k=0}^{2}\mathcal{N}(80k,(k+1)^4)$
Strongly skewed	$\sum_{k=0}^{7}\frac{1}{8}\mathcal{N}(3((\frac{2}{3})^k-1),(\frac{2}{3})^{2k})$	Claw	$\frac{1}{2}\mathcal{N}(0,1)+\sum_{k=0}^{4}\frac{1}{10}\mathcal{N}(\frac{k}{2}-1,(\frac{1}{10})^2)$
Outlier	$\frac{1}{10}\mathcal{N}(0,1)+\frac{9}{10}\mathcal{N}(0,(\frac{1}{10})^2)$	Nakagami	$\frac{2\mu^\mu x^{2\mu-1}}{\Gamma(\mu)\omega^\mu}\exp(-\frac{\mu}{\omega}x^2), x>0, \mu=\omega=2$
Gamma	$\frac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-x/\theta}, x>0$	Generalized	$\frac{1}{\sigma}\exp(-(1+k\frac{x-\mu}{\sigma})^{-\frac{1}{k}})(1+k\frac{x-\mu}{\sigma})^{-1-\frac{1}{k}}$
	$k=9, \theta=0.5$	Extreme value	$k=0$, $\mu=1$, $\sigma=0$, $x>0$ (type II)
Kou's double	$\sigma=0.04, \lambda=2, p_{up}=0.4$	Log-normal	$\frac{1}{x\sigma\sqrt{2\pi}}\exp(-(\ln(x)-\mu)^2/(2\sigma^2))$
Exponential	$\eta_1=3, \eta_2=5, \Delta_t=1/4$		$x>0$, $\mu=0$, $\sigma=1$
Merton's jump	$\sigma=0.08, \lambda=3, \mu_J=-0.01$	Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp(-(x/\lambda)^k)$
Diffusion	$\sigma_J=0.4, \Delta_t=1/4$		$x\geq 0$, $\lambda=1$, $k=5$
Smooth comb	$\sum_{k=0}^{5}\frac{2^{5-k}}{63}\mathcal{N}(\frac{65-96/2^k}{21},(\frac{32/63}{2^k})^2)$	Chi-square	$\chi^2_k, k=4, x>0$

plied by minimizing the average log-likelihood $\mathrm{LCV}(h) = -\frac{1}{N}\sum_{i=1}^{N} \log \bar{f}^{a}_{-i}(X_i)$, where $\bar{f}^{a}_{-i}(x)$ is the same as $\bar{f}^{a}(x;N)$ but with the observation X_i removed from the fitting process. This approach can be shown to minimize the Kullback-Leibler distance between the estimated density and the true density. A related strategy developed in Rudemo (1982) and Bowman (1984) is known as least squares cross-validation (LSCV), and is discussed in the supplemental appendix.

In Figure 4 we compare the convergence rate of several bandwidth selection methods for normal and skewed unimodal data, with test case densities and parameters defined in Table 2. For normal data (left), the normal rule of thumb (NRT) is given, while for the skewed unimodal data, the heuristic of Section 3.1.2 is employed, as the NRT is inappropriate in this setting. What we observe is that for small values of N, the cross validation methods LSCV and LCV provide competitive estimates, but often the plug-in approaches perform best in a large sampling setting. This finding is application dependent, and the best approach is to compare several competing bandwidth methods for the problem at hand.

4. EMPIRICAL CHARACTERISTIC FUNCTION APPROACH

In Section 2, we embarked under the assumption that $\widetilde{\varphi}$ is known explicitly. While this may be the case (see for example Appendix B.3), in general we can alternatively estimate the coefficients $\beta_{a,k}$ using the Fourier transform of $\widetilde{\varphi}$:

$$\widehat{\widetilde{\varphi}}(\xi) := \frac{\widehat{\varphi}(\xi)}{\mathbf{\Phi}(\xi)}, \qquad \mathbf{\Phi}(\xi) := \sum_{k\in\mathbb{Z}} |\widehat{\varphi}(\xi + 2\pi k)|^2, \quad \xi \in \mathbb{R}, \tag{23}$$

where $\widehat{\varphi}(\xi) = \mathcal{F}\varphi(\xi) = \int_{\mathbb{R}} \varphi(x) e^{ix\xi} dx$, and $\widehat{\widetilde{\varphi}}(\xi) = \mathcal{F}\widetilde{\varphi}(\xi)$. Transiting to the frequency space also facilitates the use of spectral filtering (see Section 4.1), offering greater control over the resulting density smoothness.[19]

Due to the compact support of φ, $\mathbf{\Phi}(\xi)$ is in fact a trigonometric polynomial, which admits a finite cosine series expansion. From Kirkby (2015, 2017), we can derive an expression for the transform of the p-th order dual generator $\widehat{\widetilde{\varphi}}^{[p]}(\xi) = \widehat{\varphi}^{[p]}(\xi)/\mathbf{\Phi}^{[p]}(\xi)$ by using

$$\widehat{\varphi}^{[p]}(\xi) = \left(\frac{\sin(\xi/2)}{(\xi/2)}\right)^{p+1},$$

and

$$\mathbf{\Phi}^{[p]}(\xi) = \int_{-\frac{p+1}{2}}^{\frac{p+1}{2}} \varphi^{[p]}(x)^2 dx + 2\sum_{k=1}^{p+1} \cos(k\xi) \int_{-\frac{p+1}{2}}^{\frac{p+1}{2}} \varphi^{[p]}(x)\varphi^{[p]}(x-k) dx.$$

Given an expression for the dual transform $\widehat{\widetilde{\varphi}} = \widehat{\widetilde{\varphi}}^{[p]}$, and the relation $\beta_{a,k} = \mathbb{E}[\widetilde{\varphi}_{a,k}(X)] = \mathbb{E}[a^{1/2}\widetilde{\varphi}(a(X - x_k))]$, the projection coefficients for the p-th order

[19] This enables a smoother basis representation without requiring a coarser spacing between basis elements.

B-spline generator satisfy

$$\begin{aligned}\beta_{a,k} &= \frac{a^{-1/2}}{2\pi}\mathbb{E}\left[\int_{-\infty}^{\infty}\exp\left(i\xi(X-x_k)\right)\cdot\widehat{\widehat{\varphi}}(-\xi/a)d\xi\right] \\ &= \frac{a^{-1/2}}{2\pi}\int_{-\infty}^{\infty}\mathbb{E}\left[\exp\left(i\xi(X-x_k)\right)\right]\cdot\widehat{\widehat{\varphi}}(-\xi/a)d\xi \\ &= \frac{a^{-1/2}}{\pi}\Re\left\{\int_0^{\infty}\exp(-ix_k\xi)\cdot\phi(\xi)\cdot\widehat{\widehat{\varphi}}(\xi/a)d\xi\right\}, \quad k\in\mathbb{Z}, \end{aligned} \tag{24}$$

where the characteristic function of $X_n \overset{d}{=} X$ is defined by

$$\phi(\xi) = \mathbb{E}[e^{iX\xi}] = \int e^{ix\xi}f(x)dx, \quad \xi\in\mathbb{R}.$$

Given a sample $\{X_n\}_{n=1}^N$, the *empirical characteristic function* (ECF) at ξ is the complex-valued sample statistic defined by

$$\phi_N(\xi) := \frac{1}{N}\sum_{n=1}^{N}\exp(iX_n\xi) = \frac{1}{N}\sum_{n=1}^{N}\{\cos(X_n\xi)+i\sin(X_n\xi)\}. \tag{25}$$

It follows that $\phi_N(\xi)$ is an unbiased estimator of $\phi(\xi)$, i.e., $\mathbb{E}[\phi_N(\xi)-\phi(\xi)]=0$. Hence, the estimator of $\beta_{a,k}$ formed by replacing ϕ with ϕ_N in equation (24) is also unbiased for $\beta_{a,k}$.

To implement the projected density method in practice, we require a truncated domain, by restricting to N_φ basis elements centered over the points:

$$x_k = x_1+(k-1)h, \quad k=1,\ldots,N_\varphi, \tag{26}$$

where x_1 is the leftmost grid point, and $h := 1/a$ is the step size (bandwidth). We will denote the truncated density support by $[l,u]=[x_1,x_{N_\varphi}]$, the choice of which is detailed in Appendix B.1. The density approximation becomes[20]

$$\widetilde{f}^a_{[l,u]}(x;N) = \sum_{k=1}^{N_\varphi}\widetilde{\beta}_{a,k}(N)\varphi_{a,k}(x), \tag{27}$$

where $\{\widetilde{\beta}_{a,k}\}_{k=1}^{N_\varphi}$ are the coefficients found upon discretizing (24).

Depending on the application, efficiency in determining the density estimator may be essential, in which case the fast Fourier transform (FFT) can be used to derive $\{\widetilde{\beta}_{a,k}\}_{k=1}^{N_\varphi}$ with complexity $\mathcal{O}(N_\varphi\log_2(N_\varphi))$. The tradeoff for efficiency (in this case a reduction from $\mathcal{O}(N_\varphi^2)$) is that the Nyquist frequency imposes a frequency grid spacing $\Delta_\xi = 2\pi a/N_\xi$ with $N_\xi = N_\varphi$. Details on the FFT-based implementation are given in Appendix 2.

[20] Note that, given the support of the basis elements centered over l and u, the density representation will extend slightly beyond $[l,u]$, for example, $[l-h,u+h]$ for the linear basis.

4.1. Spectral Filtering of the ECF

The use of spectral filters in option pricing was studied in Ruijter, Versteegh, and Oosterlee (2015), Cui, Kirkby, and Nguyen (2017a) to smooth extremely peaked densities which elicit Gibbs oscillations for Fourier methods. In the present context, we apply spectral filtering to the ECF in (25) to smooth the resulting density estimate, and reduce the sensitivity to the bandwidth selection. The advantage of the ECF-based approach is that in the frequency space, density smoothing (convolution) with a spectral filter reduces to the multiplication of Fourier transforms (characteristic functions).

A real, symmetric function $\Gamma(\xi)$ is a *filter* of order q if it satisfies: (i) $\Gamma(0) = 1$, $\Gamma^{(l)}(0) = 0$, $1 \le l \le q-1$; (ii) $\Gamma(\xi) = 0$ for $|\xi| \ge 1$; (iii) $\Gamma(\xi) \in C^{q-1}, \xi \in \mathbb{R}$, where in particular $\Gamma^{(l)}(\pm 1) = 0$ for $0 \le l \le q-1$. A spectrally filtered characteristic function truncated to $[-2\pi a, 2\pi a]$ takes the form $\Gamma_a(\xi)\phi(\xi)$, where $\Gamma_a(\xi) := \Gamma(\xi/(2\pi a))$, which owes to the property that convolution becomes multiplication in the frequency domain. Hence, the coefficients in (24) can be readily replaced by those of the spectrally filtered ECF:

$$\widetilde{\beta}_{a,k} \approx \frac{a^{-1/2}}{\pi} \Re\left[\int_0^{2\pi a} \exp(-i x_n \xi) \cdot \Gamma_a(\xi)\phi_N(\xi) \cdot \widehat{\widetilde{\varphi}}^{[p]}\left(\frac{\xi}{a}\right) d\xi\right]. \tag{28}$$

As an example, we consider the filtered coefficients using an exponential filter, given by $\Gamma(\xi) = \exp(-\tau\xi^q)$, where $\tau := \log\epsilon_m$, with ϵ_m defined as the machine precision epsilon.

Figure 5 illustrates the use of an exponential spectral filter ($q = 4$) for constructing a linear bi-orthogonal density estimator with normal data. To visualize the smoothing effect of filtering, the left figure considers the case of a relatively small sample, $N = 10^3$, combined with a bandwidth that is chosen slightly smaller than the optimal value prescribed by h^*. Filtering effectively reduces the variance

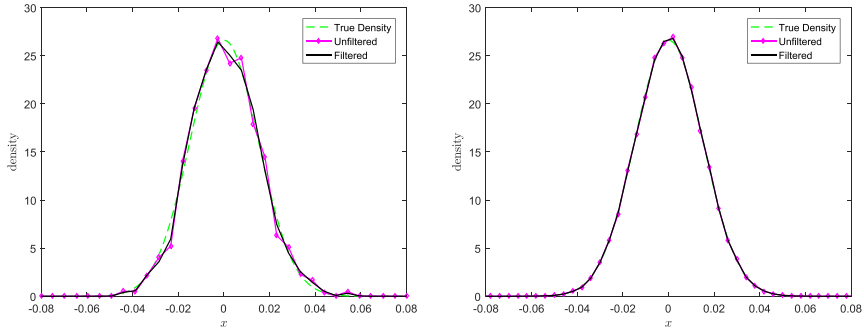

FIGURE 5. Linear bi-orthogonal estimation of normal $\mathcal{N}(0, \sigma^2\Delta_t)$ data for $\sigma = 0.15$, $\Delta_t = 1/100$; Left: $N = 10^3$ with bandwidth $h = 0.8 \cdot h^*$ chosen smaller than optimal. Right: $N = 10^5$, with $h = h^*$ chosen by optimal bandwidth.

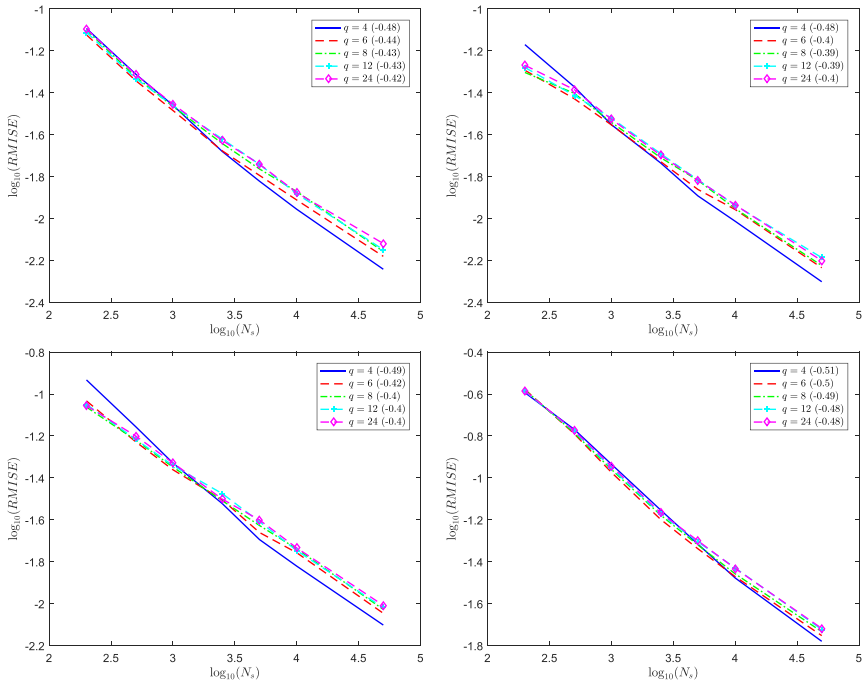

FIGURE 6. Spectral filter order convergence: linear bi-orthogonal estimation of Student-t (Top Left), gamma (Top Right), separated bimodal (Bottom Left), and kurtotic unimodal (Bottom Right), where parameters are given in Table 2. The rate of convergence (in parentheses) for each filter order q is estimated as the slope coefficient from a regression of $\log_{10}(RMSE)$ against $\log_{10}(N_s)$, where $N_s = N$ is the sample size.

of the estimator, and offers some protection against an aggressively chosen (small) bandwidth.

Figure 6 illustrates the effect of filter order on the RMISE convergence for several densities.[21] For each density, we plot the convergence as a function of N for $q \in \{4, 6, 8, 12, 24\}$, where $q = 24$ is numerically equivalent to no smoothing. A choice of $q \in \{4, 6, 8\}$ is found to produce good results, and the performance is not highly sensitive to the filter order within this range of values.[22] We note that MISE accounts for both bias and variance, and filtering trades extra bias for a reduction in variance. Hence, while a comparison of MISE demonstrates a modest improvement over no filtering, the smoothness of an individual estimate (rather than a measure taken on average) is often greatly improved, as demonstrated in the left panel of Figure 5, especially when the "optimal" bandwidth is underestimated.

[21] For these experiments, we apply the plug-in bandwidth method described in Section 3.2.

[22] In practice, we use $p = 6$ for $N \leq 1,600$, and $p = 4$ otherwise.

TABLE 3. B-spline estimator comparison: bi-orthogonal estimator vs. Redner (1999). Comparison of $R(h_i) = \text{Ratio}(h_i) := \text{MISE}_{Biorth}/\text{MISE}_{Redner}$ for two plugin bandwidth methods. Ratios $R(h_1)$, $R(h_2)$ are computed by averaging 1,000 Monte Carlo replications, where h_1 denotes the average bandwidth computed using the plug-in method from Section 3.2, and h_2 using a Normal rule of thumb

Case	N	h_1	h_2	$R(h_1)$	$R(h_2)$	Case	N	h_1	h_2	$R(h_1)$	$R(h_2)$
	50	1.83	1.49	0.84	0.60		50	2.57	2.22	0.84	0.72
	100	1.51	1.30	0.55	0.51		100	2.09	1.94	0.60	0.51
Gaussian	200	1.26	1.13	0.43	0.43	Gamma	200	1.75	1.69	0.39	0.37
	400	1.07	0.98	0.36	0.40		400	1.46	1.47	0.38	0.36
	800	0.91	0.86	0.31	0.35		800	1.24	1.28	0.38	0.32
	50	1.96	1.80	0.84	0.75		50	1.86	1.63	0.84	0.85
	100	1.59	1.59	0.65	0.61		100	1.39	1.42	0.86	0.94
Student-t	200	1.32	1.38	0.51	0.50	Skewed	200	1.03	1.23	0.90	0.95
	400	1.09	1.21	0.44	0.41	Bimodal	400	0.79	1.08	0.82	0.72
	800	0.93	1.05	0.36	0.32		800	0.62	0.94	0.63	0.55
	50	0.59	1.54	0.96	0.94		50	1.45	1.29	0.95	0.87
	100	0.38	1.34	0.94	0.95		100	1.10	1.13	0.91	0.89
Strongly	200	0.27	1.18	0.88	0.97	Claw	200	0.79	0.98	0.97	0.93
Skewed	400	0.20	1.02	0.82	0.87		400	0.26	0.85	0.90	0.96
	800	0.16	0.89	0.75	0.96		800	0.19	0.74	0.59	0.91

5. NUMERICAL EXPERIMENTS

5.1. Comparison with Existing Spline Estimators

The first comparison is made to the B-spline estimator of Redner (1999), with the goal of isolating the improvement due to the use of duality for coefficient calculation, by fixing the same plug-in bandwidth for each of the two approaches. Table 3 provides a comparison with two bandwidths chosen for each experiment, and used by both estimators, with sample sizes $N \in \{50, 100, 200, 400, 800\}$. The MISE ratio is given for each of six test cases from Table 2, ranging from benign normal data to heavy tailed and multimodal distributions. In each case with either bandwidth rule, the bi-orthogonal estimator outperforms the method of Redner (1999), often substantially. In the large sample comparisons made in Section 5.2, the improvement is even more pronounced, as the bi-orthogonal estimator explicitly targets reduced bias, which dominates the asymptotic performance.

We next compare the bi-orthogonal estimator with the log-spline approach of Kooperberg and Stone (1991), implemented in R as the package "logspline". As demonstrated in Table 4, the bi-orthogonal estimator compares favorably,[23] and outperforms in the majority of cases. As the bi-orthogonal estimator targets a reduction in bias, the benefit of the proposed method is realized especially for larger values of N, which is further demonstrated in Section 5.2.

[23] For these experiments, we use the plug-in method of Section 3.2, and spectral filtering order $p = 4$.

TABLE 4. B-spline estimator vs. log-spline. Provided are the MISE of each estimator, and Ratio $= \text{MISE}_{Biorth}/\text{MISE}_{LogSpline}$

Case	N	50	100	200	400	800	1,600	3,200
Gaussian	B-spline	2.9e-02	1.1e-02	4.8e-03	2.2e-03	1.1e-03	6.7e-04	3.3e-04
	Log-spline	2.8e-02	1.6e-02	7.7e-03	5.0e-03	2.0e-03	1.1e-03	4.5e-04
	Ratio	1.02	0.66	0.62	0.45	0.56	0.61	0.73
Student-t	B-spline	3.1e-02	1.4e-02	5.9e-03	2.7e-03	1.3e-03	7.3e-04	4.1e-04
	Log-spline	1.9e-02	1.0e-02	5.2e-03	3.0e-03	1.5e-03	8.2e-04	5.3e-04
	Ratio	1.59	1.33	1.14	0.89	0.90	0.89	0.78
Gamma	B-spline	2.0e-02	7.8e-03	3.0e-03	1.7e-03	1.0e-03	5.5e-04	2.9e-04
	Log-spline	2.0e-02	1.2e-02	5.7e-03	3.0e-03	1.4e-03	6.3e-04	3.4e-04
	Ratio	1.00	0.66	0.53	0.57	0.73	0.86	0.84
Skewed-bimodal	B-spline	2.7e-02	2.0e-02	1.3e-02	7.7e-03	3.6e-03	1.9e-03	8.6e-04
	Log-spline	3.8e-02	2.0e-02	8.8e-03	5.6e-03	3.0e-03	1.7e-03	1.0e-03
	Ratio	0.71	0.99	1.42	1.38	1.19	1.13	0.85
Claw	B-spline	7.7e-02	5.7e-02	5.0e-02	3.1e-02	1.2e-02	5.6e-03	2.9e-03
	Log-spline	8.9e-02	6.9e-02	5.3e-02	4.1e-02	2.9e-02	1.9e-02	6.4e-03
	Ratio	0.86	0.83	0.94	0.76	0.40	0.29	0.46

5.2. Asymptotic Experiments

This section illustrates the performance of the linear bi-orthogonal estimator in the asymptotic (high-frequency) case that is relevant for applications in finance and economics, for example.[24] Table 5 provides a wide range of test cases considered in Marron and Wand (1992) and Botev et al. (2010), as well as several additional cases, which are categorized as follows: unimodal on $\mathbb{R}$, unimodal on the semi-infinite domain $\mathbb{R}_+$, and multimodal on $\mathbb{R}$. The MISE of the bi-orthogonal estimator is presented,[25] as well as the ratio $\text{MISE}_{Biorth}/\text{MISE}_{Botev}$ with the Gaussian kernel method of Botev et al. (2010), and the ratio $\text{MISE}_{Biorth}/\text{MISE}_{Redner}$ with the method of Redner (1999). The bandwidth of these two methods is made using the same plug-in estimate of $\|f^{(2)}\|_2^2$ for each method, described in Section 3.2. We also provide a comparison with the Matlab kernel density estimation package "ksdensity", for the Gaussian and Epanechnikov kernels. For roughly symmetric unimodal data, the standard package implementation of the Gaussian and Epanechnikov kernels performs similarly to the implementation developed in Botev et al. (2010). By comparison, for multimodal and skewed data, the difference is often staggering and highlights the importance of careful bandwidth selection. However, even with a more careful bandwidth approach for the Gaussian kernel, we often observe superior performance of the bi-orthogonal estimator for a wide range of test cases, although the method of Botev et al. (2010) outperforms in some cases in Table 5.

[24] Each method below is estimated using the plug-in bandwidth method described in Section 3.2, and the bi-orthogonal estimator uses a spectral filter order of $q = 6$ for all experiments.

[25] To calculate the MISE, we fix a fine mesh $\{y_k\}_{k=1}^K$ and average the ISE for each of 50 samples. The ISE is defined by $\text{ISE} := \Delta_y \sum_{k=1}^K (\bar{f}(y_k; N) - f(y_k))^2$, where $f(y)$ is the true density, and $\bar{f}(y; N)$ is the estimate.

TABLE 5. MISE comparison with Gaussian kernel method of Botev et al. (2010) and spline estimator of Redner (1999). MISE computed as average ISE over 1,000 trials for each case, and reported for the linear bi-orthogonal estimator. Ratios of MISE for the linear bi-orthogonal estimator with respect to other four methods are reported: $\text{MISE}_{Biorth}/\text{MISE}_{other}$, for the methods of Redner (1999), Botev et al. (2010), and the standard Matlab kernel packages with a Gaussian (Gauss.) and Epanechnikov (Epan.) kernel

N	Case	MISE	Botev	Redner	Gauss.	Epan.	Case	MISE	Botev	Redner	Gauss.	Epan.
10^4	Gaussian	1.1e-04	0.65	0.24	0.67	0.70	Weibull	5.8e-04	0.68	0.25	0.69	0.72
10^5		2.0e-05	0.65	0.25	0.65	0.68		9.4e-05	0.29	0.24	0.64	0.67
10^4	Student-t	1.2e-04	0.69	0.25	0.69	0.72	Generalized	3.7e-04	0.98	0.44	0.38	0.35
10^5		1.9e-05	0.64	0.24	0.64	0.67	Extr. value	4.5e-05	0.35	0.30	0.24	0.23
10^4	Kou's double	6.0e-03	0.74	0.28	0.51	0.51	Bimodal	1.2e-03	0.78	0.30	0.00	0.00
10^5	Exponential	8.7e-04	0.64	0.27	0.47	0.48		1.6e-04	0.64	0.25	0.00	0.00
10^4	Merton jump	2.9e-03	0.79	0.31	0.29	0.28	Separated	2.4e-04	0.71	0.27	0.01	0.01
10^5	Diffusion	4.1e-04	0.69	0.28	0.24	0.23	Bimodal	3.4e-05	0.64	0.25	0.01	0.01
10^4	Kurtotic	1.1e-03	0.88	0.34	0.09	0.08	Skewed	3.5e-04	1.02	0.40	0.30	0.27
10^5	Unimodal	1.4e-04	0.70	0.27	0.05	0.05	bimodal	3.9e-05	0.73	0.28	0.16	0.16
10^4	Skewed	2.2e-04	0.71	0.26	0.71	0.74	Trimodal	1.2e-04	0.87	0.35	0.00	0.00
10^5	Unimodal	2.8e-05	0.63	0.24	0.62	0.65		1.5e-05	0.59	0.27	0.00	0.00
10^4	Strongly	1.9e-03	1.28	0.54	0.07	0.06	Claw	1.0e-03	0.92	0.37	0.04	0.03
10^5	skewed	2.0e-04	0.81	0.33	0.02	0.02		1.4e-04	0.72	0.28	0.01	0.01
10^4	Outlier	1.2e-03	0.68	0.24	0.69	0.71	Smooth	5.2e-03	0.02	0.96	0.12	0.11
10^5		1.9e-04	0.62	0.25	0.63	0.65	Comb	3.8e-03	0.01	0.97	0.14	0.12
10^4	Gamma	9.2e-05	0.72	0.26	0.72	0.75	Nakagami	2.9e-04	0.69	0.27	0.70	0.73
10^5		1.5e-05	0.68	0.26	0.68	0.71		4.1e-05	0.64	0.24	0.64	0.67
10^4	Log-normal	3.6e-04	1.08	0.68	0.19	0.16	Chi-square	1.4e-04	0.79	0.68	0.36	0.33
10^5		6.7e-05	0.61	0.56	0.12	0.10		2.5e-05	0.75	0.63	0.24	0.21

For high-frequency data, the advantage of the bi-orthogonal density estimator is observed most notably in cases of heavy-tailed data with peaked densities, resulting from the locality of the basis representation. For example, we consider two distributions which are commonly taken as log return distributions in financial models, and belong to the class of *jump diffusions* (see related econometric studies of jump diffusions in Yu, 2007; Filipović, Mayerhofer, and Schneider, 2013; Li and Chen, 2016). The log return in a jump diffusion model for an asset S_t over a time increment Δ_t can be written as

$$X(\Delta_t) = \ln(S(t+\Delta_t)/S(t)) = \theta\Delta_t + \sigma W(\Delta_t) + \sum_{k=1}^{N(\Delta_t)} J_k, \tag{29}$$

where θ is the drift, $W(\Delta_t) \sim \sqrt{\Delta_t} \cdot \mathcal{N}(0,1)$ is the increment of a Brownian motion, $N(\Delta_t)$ is a Poisson process with rate λ, and J_k is the jump size (the distribution of which determines the particular type of the jump diffusion model). In Merton's jump diffusion (MJD) (Merton, 1976), $J_k \sim \mathcal{N}(\mu_J, \sigma_J)$. In Kou's double exponential jump diffusion (Kou, 2002), J_k has a density given by

$$f_J(y) = p_{up} \cdot \eta_1 e^{-\eta_1 y} \mathbb{1}[y \geq 0] + (1 - p_{up}) \cdot \eta_2 e^{\eta_2 y} \mathbb{1}[y < 0],$$

where $\eta_1 > 1$, $\eta_2 > 0$, and $p_{up} \geq 0$ the probability of an upward jump.

As seen in Table 5, the bi-orthogonal estimator consistently outperforms the competing methods. In addition, the bi-orthogonal estimator has computational advantages over the kernel estimator, as it performs a substantial data reduction. While the kernel estimator retains the full data set, the bi-orthogonal estimator maintains a set of basis coefficients which grow slowly at a rate of $N^{\frac{1}{2p+3}}$ (recall (20)). The method of Redner (1999) performs an analogous reduction, but is considerably less accurate than the bi-orthogonal estimator which differs in the calculation of basis coefficients. In general, the bi-orthogonal estimator reduces MISE by a factor of 3–4 times compared with the estimator of Redner (1999).

6. CONCLUSIONS AND FUTURE RESEARCH

In this article, we have introduced a new method of nonparametric density estimation, namely the bi-orthogonal density estimator. As prominent members of this class, we considered the B-spline density estimators, which are locally supported and ideal for capturing fine features of a density. We find that, in addition to several computational advantages of the bi-orthogonal estimators over kernel density estimators, they provide accurate approximations even at coarse resolutions. In particular, they outperform other representative competing methods such as the local basis method of Redner (1999) and the kernel method of Botev et al. (2010). Given the local nature of B-splines, they are ideal for capturing the peaked densities that are commonly observed with high-frequency data.

As for future research directions, extension to nonparametric regressions (see recent work of Wang and Hong, 2018; Adusumilli and Otsu, 2018) is a promising venue. As for methodological extensions, the case of higher dimensions (e.g., stochastic volatility settings) using tensor bases, and a more refined characterization of the optimal choices of basis functions is of interest to explore.

REFERENCES

Adusumilli, K. & T. Otsu (2018) Nonparametric instrumental regression with errors in variables. *Econometric Theory* 36(6), 1256–1280.

Aït-Sahalia, Y. (1996) Nonparametric pricing of interest rate derivative securities. *Econometrica* 64(3), 527–560.

Bandi, F.M. & G. Moloche (2017) On the functional estimation of multivariate diffusion processes. *Econometric Theory* 34(4), 896–946.

Beare, B. & L. Schmidt (2014) An empirical test of pricing kernel monotonicity. *Journal of Applied Econometrics* 31(2), 338–356.

Botev, Z.I., J.F. Grotowski, & D.P. Kroese (2010) Kernel density estimation via diffusion. *Annals of Statistics* 38(5), 2916–2957.

Bowman, A. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71, 353–360.

Cai, Z. & X. Wang (2008) Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics* 147(1), 120–130.

Calonico, S., M.D. Cattaneo, & M.H. Farrell (2018) On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association* 113(522), 767–779.

Carroll, R., A. Delaigle, & P. Hall (2013) Unexpected properties of bandwidth choice when smoothing discrete data from construction a functional data classifier. *Annals of Statistics* 41(6), 2739–2767.

Cattaneo, M.D., M. Jansson, & X. Ma (2017) Simple Local Polynomial Density Estimators. Technical report, Working paper. Retrieved July 22, 2017 from http://wwwpersonal.umich.edu/˜cattaneo/papers/Cattaneo-Jansson-Ma_2017_LocPolDensity.pdf.

Cencov, N. (1962) Evaluation of an unknown distribution density from observations. *Soviet Mathematics* 3, 1559–1562.

Chen, S. & Z. Xu (2014) On implied volatility for options-some reasons to smile and more to correct. *Journal of Econometrics* 179(1), 1–15.

Christensen, O. (2003) *An Introduction to Frames and Riesz Bases*. Birkhauser Boston.

Cui, Z., J. Kirkby, & D. Nguyen (2017a) Equity-linked annuity pricing with cliquet-style guarantees in regime-switching and stochastic volatility models with jumps. *Insurance*: *Mathematics and Economics* 74, 46–62.

Cui, Z., J. Kirkby, & D. Nguyen (2017b) A general framework for discretely sampled realized variance derivatives in stochastic volatility models with jumps. *European Journal of Operational Research* 262(1), 381–400.

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

Duin, R. (1976). On the choice of smoothing parameters of Parzen estimators of probability density functions. *IEEE Transactions on Computers* C-25, 1175–1179.

Eilers, P. & B. Marx (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2), 89–121.

Fan, J. (2005) A selective overview of nonparametric methods in financial econometrics. *Statistical Science* 317–337.

Figueroa-López, J.E. & C. Li (2016) Optimal kernel estimation of spot volatility of stochastic differential equations, arXiv preprint, arXiv:1612.04507.

Filipović, D., E. Mayerhofer, & P. Schneider (2013) Density approximations for multivariate affine jump-diffusion processes. *Journal of Econometrics* 176(2), 93–111.

Fix, E. & J. Hodges (1951) Nonparametric discrimination: Consistency properties. Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, February.

Gehringer, K.R. & R.A. Redner (1992) Nonparametric probability density estimation using normalized b–splines. *Communications in Statistics-Simulation and Computation* 21(3), 849–878.

Gu, C. (1993) Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association* 88(422), 495–504.

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

Gu, C. & J. Wang (2003) Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica* 13, 811–826.

Hall, P. (1981) On trigonometric series estimates of densities. *Annals of Statistics* 9(3), 683–685.

Hall, P. (1987) Cross-validation and the smoothing of orthogonal series density estimators. *Journal of Multivariate Analysis* 21(2), 189–206.

Hall, P., K.-H. Kang (2005) Bandwidth choice for nonparametric classification. *The Annals of Statistics* 33(1), 284–306.

Hall, P.G. & J.S. Racine (2015) Infinite order cross-validated local polynomial regression. *Journal of Econometrics* 185(2), 510–525.

Hardle, W. & J.S. Marron (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics* 13(4), 1465–1481.

Hayfield, T., J.S. Racine (2008) Nonparametric econometrics: The NP package. *Journal of Statistical Software* 27(5), 1–32.

Heil, C. (2011) *A Basis Theory Primer, Expanded Edition*. Birkhauser.

Huang, S.-Y. (1999) Density estimation by wavelet-based reproducing kernels. *Statistica Sinica* 9(1), 137–151.

Izenman, A. (1991) Recent developments in nonparametric density estimation. *Journal of the American Statistical Association* 86(413), 205–223.

Jones, M., J.S. Marron, & B.U. Park (1991) A simple root *n* bandwidth selector. *Annals of Statistics* 19(4), 1919–1932.

Jones, M.C., J.S. Marron, & S.J. Sheather (1996a) A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 91(433), 401–407.

Jones, M.C., J.S. Marron, & S.J. Sheather (1996b) Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* 11(3), 337–381.

Kirkby, J. (2015) Efficient option pricing by frame duality with the fast Fourier transform. *SIAM Journal on Financial Mathematics* 6(1), 713–747.

Kirkby, J. (2016). An efficient transform method for Asian option pricing. *SIAM Journal on Financial Mathematics* 7(1), 845–892.

Kirkby, J. (2017) Robust option pricing with characteristic functions and the B-spline order of density projection. *Journal of Computational Finance* 21(2), 101–127.

Kirkby, J. & S. Deng (2019) Static hedging and pricing of exotic options with payoff frames. *Mathematical Finance,* 29(2), 407–693.

Kirkby, J., D. Nguyen, & Z. Cui (2017) A unified approach to Bermudan and barrier options under stochastic volatility models with jumps. *Journal of Economic Dynamics and Control* 80, 75–100.

Koo, J. (1996) Bivariate B-splines for tensor logspline density estimation. *Computational Statistics & Data Analysis* 21, 31–42.

Kooperberg, C. & C. Stone (1991) A study of logspline density estimation. *Computational Statistics & Data Analysis* 12, 327–347.

Kou, S.G. (2002) A jump-diffusion model for option pricing. *Management Science* 48(8), 1086–1101.

Leitao, A., C. Oosterlee, L. Ortiz-Gracia, & S. Bohte (2018) On the data-driven COS method. *Applied Mathematics and Computation* 317, 68–84.

Li, C. & D. Chen (2016) Estimating jump–diffusions using closed-form likelihood expansions. *Journal of Econometrics* 195(1), 51–70.
Liu, G. & L.J. Hong (2009) Kernel estimation of quantile sensitivities. *Naval Research Logistics* 56(6), 511–525.
Liu, G. & L.J. Hong (2011) Kernel estimation of the Greeks for options with discontinuous payoffs. *Operations Research* 59(1), 96–108.
Loader, C. (1999) Bandwidth selection: Classical or plug-in? *Annals of Statistics* 27(2), 415–438.
Marron, J. (1987) A comparison of cross-validation techniques in density estimation. *Annals of Statistics* 15(1), 152–162.
Marron, J. & M. Wand (1992) Exact mean integrated squared error. *Annals of Statistics* 20(2), 712–736.
Marron, J.S. (1985) An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Annals of Statistics* 13(3), 1011–1023.
Masri, R. & R. Redner (2005) Convergence rates for uniform B-spline density estimators on bounded and semi-infinite domains. *Nonparametric Statistics* 17(5), 555–582.
Merton, R. (1976) Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* 3, 125–144.
Opschoor, A., D. Dijk, & M. van der Wel (2017) Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics* 32(7), 1298–1313.
Parzen, E. (1962) On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33, 1065–1076.
Penev, S. & L. Dechevsky (1997) On nonnegative wavelet-based density estimators. *Journal of Nonparametric Statistics* 7, 365–394.
Peter, A. & A. Rangarajan (2008) Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE Transactions on Image Processing* 17(4), 458–468.
Racine, J. & K. Li (2017) Nonparametric conditional quantile estimation: A locally weighted quantile kernel approach. *Journal of Econometrics* 201(1), 72–94.
Redner, R. (1999) Convergence rates for uniform B-spline density estimators part I: One dimension. *SIAM Journal on Scientific Computing* 20(6), 1929–1953.
Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* 27, 832–837.
Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* 9, 65–78.
Ruijter, M., M. Versteegh, & C. Oosterlee (2015) On the application of spectral filters in a Fourier option pricing technique. *Journal of Computational Finance* 19(1), 76–106.
Schwartz, S. (1967) Estimation of a probability density by an orthogonal series. *Annals of Mathematical Statistics* 38, 1261–1265.
Sheather, S.J. & M.C. Jones (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B* (*Methodological*) 53(3), 683–690.
Silverman, B. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.
Song, Z. & D. Xiu (2016) A tale of two option markets: Pricing kernels and volatility risk. *Journal of Econometrics* 190(1), 176–196.
Terrell, G. & D. Scott (1992) Variable kernel density estimation. *Annals of Statistics* 20(3), 1236–1265.
Unser, M. (1996) Vanishing moments and the approximation power of wavelet expansions. In P. Delogne (ed.), *Image Processing, 1996. Proceedings, International Conference on*, vol. 1, pp. 629–632. IEEE.
Unser, M. & I. Daubechies (1997) On the approximation power of convolution-based least squares versus interpolation. *IEEE Transactions on Signal Processing* 45(7), 1697–1711.
Van Es, B., P. Spreij, & H. Zanten (2003) Nonparametric volatility density estimation. *Bernoulli* 9(3), 451–465.

Wahba, G. (1981) Data-based optimal smoothing of orthogonal series density estimates. *Annals of Statistics* 9(1), 146–156.
Walter, G. & J. Blum (1979) Probability density estimation using delta sequences. *Annals of Statistics* 7(2), 328–340.
Wand, M.P. & M.C. Jones (1994) *Kernel Smoothing*. CRC Press.
Wang, C.-S. & Z. Zhao (2016) Conditional value-at-risk: Semiparametric estimation and inference. *Journal of Econometrics* 195(1), 86–103.
Wang, X. & Y. Hong (2018) Characteristic function based testing for conditional independence: A nonparametric regression approach. *Econometric Theory* 34(4), 815–849.
Watson, G. (1969) Density estimation by orthogonal series. *The Annals of Mathematical Statistics* 38, 1262–1265.
Wegman, E. (1972) Nonparametric probability density estimation: A summary of available methods. *Technometrics* 14(3), 533–546.
Woodroofe, M. (1970) On choosing a delta sequence. *Annals of Mathematical Statistics* 41, 1665–1671.
Young, R. (1980) *An Introduction to Nonharmonic Fourier Series*, Revised ed. Academic Press.
Yu, J. (2007) Closed-form likelihood approximation and estimation of jump-diffusions with an application to the realignment risk of the Chinese Yuan. *Journal of Econometrics* 141(2), 1245–1280.
Zhang, X., R. Brooks, & M. King (2009) A Bayesian approach to bandwidth selection for multivariate kernel regression with an application to state-price density estimation. *Journal of Econometrics* 153(1), 21–32.
Zu, Y. (2015) Nonparametric specification tests for stochastic volatility models based on volatility density. *Journal of Econometrics* 187(1), 323–344.

APPENDIX A: Proofs

LEMMA A.2. *Suppose that Assumption 2 holds for a generator φ and density f, and let $\{X_i\}_{i=1}^N \overset{iid}{\sim} f$ be a sample. Then for any $a = 1/h > 0$, the following hold*:

(*i*) *For any $k \in \mathbb{Z}$, $\bar{\beta}_{a,k}(N) \xrightarrow{a.s.} \beta_{a,k}$, and $\sup_{k\in\mathbb{Z}} |\bar{\beta}_{a,k}(N)| \leq a^{1/2}\|\widetilde{\varphi}\|_\infty < \infty$ (uniformly in N)*

(*ii*) *For any $k \in \mathbb{Z}$, $\bar{\beta}_{a,k}(N) \xrightarrow{L^1} \beta_{a,k}$, that is $\mathbb{E}[|\bar{\beta}_{a,k}(N) - \beta_{a,k}|] \to 0$ as $N \to 0$.*

(*iii*) *For any $k \in \mathbb{Z}$, for any $x \in \mathbb{R}$,*

$$\sqrt{N} \cdot \left(\left(\bar{\beta}_{a,k}(N) - \beta_{a,k} \right) \varphi_{a,k}(x) \right) \xrightarrow{d} \varphi_{a,k}(x) \cdot \mathcal{N}(0, Var(\widetilde{\varphi}_{a,k}(X_1))), \tag{A.1}$$

where $\sup_{k\in\mathbb{Z}} \left\{ Var(\widetilde{\varphi}_{a,k}(X_1)) \right\} \leq a\|\widetilde{\varphi}\|_\infty^2 < \infty$, uniformly in N.

Proof. From (8), we have

$$\bar{\beta}_{a,k}(N) - \beta_{a,k} = \frac{1}{N} \sum_{1\leq n\leq N} \widetilde{\varphi}_{a,k}(X_n) - \mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)] = \frac{1}{N} \sum_{1\leq n\leq N} Y_n - \mathbb{E}[Y_1],$$

where $Y_n \overset{iid}{\sim} \widetilde{\varphi}_{a,k}(X_1)$. Since $\widetilde{\varphi}_{a,k}(x)$ is bounded, we have, $\mathbb{E}[|Y_n|] \leq a^{1/2}\|\widetilde{\varphi}\|_\infty < \infty$, so that $\sup_{k\in\mathbb{Z}} |\bar{\beta}_{a,k}(N)| < \infty$, and the strong law of large numbers (SLLN) yields

$\bar{\beta}_{a,k}(N) \xrightarrow{a.s.} \beta_{a,k}$, which proves (i). Part (ii) then follows by dominated convergence. Since $\widetilde{\varphi}_{a,k}(x)$ is bounded, we have by Popoviciu's inequality that

$$\mathrm{Var}(Y_n) \le \frac{1}{4}\Big(\sup_x \widetilde{\varphi}_{a,k}(x) - \inf_x \widetilde{\varphi}_{a,k}(x)\Big)^2 \le \frac{a}{4}(2\|\widetilde{\varphi}\|_\infty)^2 = a\|\widetilde{\varphi}\|_\infty^2,$$

so $\sup_{k\in\mathbb{Z}}\big\{\mathrm{Var}(\widetilde{\varphi}_{a,k}(X_1))\big\} < \infty$. By the central limit theorem (CLT), as $N \to \infty$,

$$\sqrt{N}\Big(\frac{1}{N}\sum_{1\le n\le N}\widetilde{\varphi}_{a,k}(X_n) - \beta_{a,k}\Big) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\widetilde{\varphi}_{a,k}(X_1))),$$

from which (A.1) follows, and (iii) is proved. This completes the proof. ■

LEMMA A.3. *Suppose that Assumption 2 holds for a generator φ and density f, and let $\{X_i\}_{i=1}^N \overset{iid}{\sim} f$ be a sample. Then for any $a = 1/h > 0$, the following hold*:

(*i*) $\int \bar{f}^a(x;N)dx = a^{-1/2}\sum_{k\in\mathbb{Z}}\bar{\beta}_{a,k}(N) = 1, a.s.$
(*ii*) *Regularity*: $\|\bar{f}^a(\cdot;N)\|_\infty < \infty$, $\|\bar{f}^a(\cdot;N)\|_1 < \infty$, $\|\bar{f}^a(\cdot;N)\|_2 < \infty$.

Proof. From standard results, $\overline{\mathrm{span}}\{\varphi_{a,k}\}_{k\in\mathbb{Z}} = \overline{\mathrm{span}}\{\widetilde{\varphi}_{a,k}\}_{k\in\mathbb{Z}}$. Hence by the partition of unity assumption (in which case $\int\varphi(x)dx = 1$) and duality of the basis, for any fixed $x \in \mathbb{R}$,

$$1 = a^{-1/2}\sum_{k\in\mathbb{Z}}\varphi_{a,k}(x) = \sum_{k\in\mathbb{Z}}\varphi_{a,k}(x)\langle 1, \widetilde{\varphi}_{a,k}\rangle = \sum_{k\in\mathbb{Z}}\widetilde{\varphi}_{a,k}(x)\langle 1, \varphi_{a,k}\rangle = a^{1/2}\sum_{k\in\mathbb{Z}}\widetilde{\varphi}_{a,k}(x), \tag{A.2}$$

so $\widetilde{\varphi}$ also generates a partition of unity. Recalling the kernel representation in (9), we have

$$\frac{1}{h}\int K\left(\frac{x}{h}, \frac{X_n}{h}\right)dx = \frac{1}{h}\sum_{k\in\mathbb{Z}}\widetilde{\varphi}_{a,k}(X_n)\int\varphi_{a,k}(x)dx = a^{1/2}\sum_{k\in\mathbb{Z}}\widetilde{\varphi}_{a,k}(X_n) = 1,$$

where the final equality follows from (A.2), which holds for any $x \in \mathbb{R}$. Hence

$$\int \bar{f}^a(x;N)dx = \frac{1}{hN}\sum_{1\le n\le N}\int K\left(\frac{x}{h}, \frac{X_n}{h}\right)dx = 1.$$

Now, for any $x \in \mathbb{R}$,

$$|\bar{f}^a(x;N)| \le \sum_{k\in\mathbb{Z}}\varphi_{a,k}(x)|\bar{\beta}_{a,k}(N)| \le \sup_{k\in\mathbb{Z}}|\bar{\beta}_{a,k}(N)|\sum_{k\in\mathbb{Z}}\varphi_{a,k}(x) = a^{1/2}\sup_{k\in\mathbb{Z}}|\bar{\beta}_{a,k}(N)| < \infty,$$

which follows from Lemma A.2 and the partition of unity property. Hence, $\|\bar{f}^a(\cdot;N)\|_\infty < \infty$. Since $\int\bar{f}^a(x;N)dx = 1$, $\|\bar{f}^a(\cdot;N)\|_1 < \infty$, and we have

$$\int(\bar{f}^a(\cdot;N))^2dx \le \|\bar{f}^a(\cdot;N)\|_\infty\|\bar{f}^a(\cdot;N)\|_1 < \infty,$$

which completes the proof. ■

Proof of Proposition 2.1. First note that $\|f\|_2^2 = \int f^2(x)dx \le \|f\|_\infty \int f(x)dx = \|f\|_\infty$ so $\|f\|_\infty < \infty$ ensures $f \in L^2(\mathbb{R})$. As a result, the orthogonal projection $P_{\mathcal{M}_a} f(y)$ defined in (5) is valid. Almost sure convergence as $N \to \infty$ follows immediately from Lemma A.2 (i), and

$$\mathbb{E}[|\bar{f}^a(x;N)|] \le a^{1/2} \sum_{k \in k(x)} \mathbb{E}|\bar{\beta}_{a,k}(N)| < \infty,$$

where $k(x) := \{k : \varphi_{a,k}(x-k) > 0\}$ (which is finite because for any fixed x, only finitely many basis elements are nonzero at x, for any $a > 0$, due to the compact support of the generator), and $\mathbb{E}|\bar{\beta}_{a,k}(N)| \le \mathbb{E}|\widetilde{\varphi}_{a,k}(X_1)| < \infty$ from previous arguments. Moreover, we have that

$$\begin{aligned}\mathbb{E}[|\bar{f}^a(x;N) - P_{\mathcal{M}_a} f(x)|] &= \mathbb{E}\Big[\Big|\sum_{k\in\mathbb{Z}} (\bar{\beta}_{a,k}(N) - \beta_{a,k})\varphi_{a,k}(x)\Big|\Big] \\ &\le a^{1/2} \sum_{k\in k(x)} \mathbb{E}|\bar{\beta}_{a,k}(N) - \beta_{a,k}| \overset{N\to\infty}{\longrightarrow} 0,\end{aligned}$$

since $\bar{\beta}_{a,k}(N) \overset{L^1}{\longrightarrow} \beta_{a,k}$ from Lemma A.2 (ii), and $k(x)$ is a finite set (from which we can apply the triangle inequality). Hence L^1 convergence holds, which proves (i).

For (ii), we have that

$$\begin{aligned}\mathrm{Var}(\bar{f}^a(x;N)) &= \mathrm{Var}\Big(\sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}(N)\varphi_{a,k}(x)\Big) \\ &= \sum_{k\in\mathbb{Z}} \sum_{\substack{j\in\mathbb{Z} \\ |k-j|\le p}} \varphi_{a,k}(x)\varphi_{a,j}(x)\mathrm{Cov}\left(\bar{\beta}_{a,k}(N), \bar{\beta}_{a,j}(N)\right),\end{aligned}$$

since $\varphi_{a,k}(x)\varphi_{a,j}(x) = 0$ for $|k-j| > p$, where p is defined by Definition 1. (For B-spline generators, p is the B-spline order.) Moreover,

$$\begin{aligned}\mathrm{Cov}(\bar{\beta}_{a,j}(N), \bar{\beta}_{a,k}(N)) &= \frac{1}{N^2} \sum_{1\le n\le N} \sum_{1\le m\le N} \mathrm{Cov}(\widetilde{\varphi}_{a,j}(X_n), \widetilde{\varphi}_{a,k}(X_m)) \\ &= \frac{1}{N^2} \sum_{1\le n\le N} \mathrm{Cov}(\widetilde{\varphi}_{a,j}(X_n), \widetilde{\varphi}_{a,k}(X_n)) \\ &= \frac{1}{N}\mathrm{Cov}(\widetilde{\varphi}_{a,j}(X_1), \widetilde{\varphi}_{a,k}(X_1)), \qquad \textbf{(A.3)}\end{aligned}$$

where the second equality holds by independence of the sample. It then follows that

$$\begin{aligned}\mathrm{Var}(\bar{f}^a(x;N)) &= \sum_{k\in\mathbb{Z}} \sum_{0\le|m|\le p} \varphi_{a,k}(x)\varphi_{a,k-m}(x)\mathrm{Cov}(\bar{\beta}_{a,k}(N), \bar{\beta}_{a,k-m}(N)) \\ &= \frac{1}{N} \sum_{k\in\mathbb{Z}} \sum_{0\le|m|\le p} \varphi_{a,k}(x)\varphi_{a,k-m}(x)\widetilde{C}_{a,k,k-m}, \qquad \textbf{(A.4)}\end{aligned}$$

where $\widetilde{C}_{a,j,k} := \mathrm{Cov}(\widetilde{\varphi}_{a,j}(X_1), \widetilde{\varphi}_{a,k}(X_1))$. Moreover, by the Cauchy-Schwartz inequality

$$|\widetilde{C}_{a,j,k}| = \left| \int \widetilde{\varphi}_{a,j}(x)\widetilde{\varphi}_{a,k}(x) f(x) dx \right| \\ \leq \|f\|_\infty \left(\int |\widetilde{\varphi}_{a,j}(x)|^2 dx \right)^{1/2} \left(\int |\widetilde{\varphi}_{a,k}(x)|^2 dx \right)^{1/2} = \|f\|_\infty \|\widetilde{\varphi}\|_2^2, \qquad \textbf{(A.5)}$$

independently of a, and we conclude that

$$\sup_{j,k\in\mathbb{Z}} \left\{ \mathrm{Cov}(\widetilde{\varphi}_{a,j}(X_1)), \mathrm{Cov}(\widetilde{\varphi}_{a,k}(X_1)) \right\} \leq \|f\|_\infty \|\widetilde{\varphi}\|_2^2 < \infty.$$

Similarly from (A.3) we have $\sup_{j,k\in\mathbb{Z}} \left\{ \mathrm{Cov}(\bar{\beta}_{a,j}(N), \bar{\beta}_{a,k}(N)) \right\} < \infty$. By the compact support of φ, only finitely many of the terms in (A.4) are nonzero for any $x \in \mathbb{R}$, and they are uniformly bounded. Hence we conclude that $\mathrm{Var}(\bar{f}^a(x; N)) < \infty$. Moreover, we have $\sup\{x : \mathrm{Var}(\bar{f}^a(x; N))\} < \infty$, by taking the supremum over the uniform bounds, and noting that $\|\varphi_{a,k}\|_\infty < a^{1/2}$, uniformly in $k \in \mathbb{Z}$. (We are using the fact that finite summations of a fixed size of uniformly bounded numbers are also uniformly bounded. In particular, the supremum is over combinations of $\widetilde{C}_{a,k,k-m}$ of a fixed length $2p+1$, since $\|\varphi_{a,k}\varphi_{a,k-m}\|_\infty < a$.)

To prove (iii), from (11) and (A.5),

$$|\mathrm{Var}(\bar{f}^a(x; N))| \leq \frac{1}{N} \sum_{k\in\mathbb{Z}} \varphi_{a,k}(x) \Big(\sum_{0\leq|m|\leq p} \varphi_{a,k-m}(x) \cdot |\widetilde{C}_{a,k,k-m}| \Big) \\ \leq \frac{\|f\|_\infty \|\widetilde{\varphi}\|_2^2}{N} \sum_{k\in\mathbb{Z}} \varphi_{a,k}(x) \Big(\sum_{0\leq|m|\leq p} \varphi_{a,k-m}(x) \Big) \\ \leq \frac{\|f\|_\infty \|\widetilde{\varphi}\|_2^2}{N} \cdot a^{1/2} \cdot a^{1/2} = \frac{\|f\|_\infty \|\widetilde{\varphi}\|_2^2}{Nh},$$

where the final equality follows from the partition of unity property, and here $\sum_{0\leq|m|\leq p} \varphi_{a,k-m}(x) \leq a^{1/2}$ for any fixed $k \in \mathbb{Z}$ and $x \in \mathbb{R}$, and $\sum_{k\in\mathbb{Z}} \varphi_{a,k}(x) = a^{1/2}$ for any $x \in \mathbb{R}$. Also note that $\|\widetilde{\varphi}\|_2^2 < \infty$, since $\widetilde{\varphi} \in L^2(\mathbb{R})$.

To prove (iv), note from (9) that we can write $\bar{f}^a(x; N) = \frac{1}{N} \sum_{1\leq n\leq N} Z_h(x, X_n)$, where with $K(x, y)$ defined in (10), $Z_h(x, X_n) := \frac{1}{h} K\left(\frac{x}{h}, \frac{X_n}{h}\right)$ are i.i.d. with finite variance, $\mathrm{Var}(Z_h(x, X_1)) = N \cdot \mathrm{Var}(\bar{f}^a(x; N)) < \infty$ from (ii). Hence by the CLT, for any fixed $x \in \mathbb{R}$,

$$\sqrt{N}\left(\bar{f}^a(x; N) - P_{\mathcal{M}_a} f(x)\right) = \sqrt{N} \left(\frac{1}{N} \sum_{1\leq n\leq N} Z_h(x, X_n) - P_{\mathcal{M}_a} f(x) \right) \\ = \sqrt{N} \left(\sum_{k\in\mathbb{Z}} (\bar{\beta}_{a,k}(N) - \beta_{a,k}) \varphi_{a,k}(x) \right) \\ \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(Z_h(x, X_1))),$$

which gives

$$\left(\bar{f}^a(x;N) - P_{\mathcal{M}_a} f(x)\right) / \sqrt{\mathrm{Var}\left(\bar{f}^a(x;N)\right)} \xrightarrow{d} \mathcal{N}(0,1). \tag{A.6}$$

Next, note that

$$\bar{f}^a(x;N) - f(x) = \left(\bar{f}^a(x;N) - P_{\mathcal{M}_a} f(x)\right) + \left(P_{\mathcal{M}_a} f(x) - f(x)\right),$$

where $\left(P_{\mathcal{M}_a} f(x) - f(x)\right)$ is deterministic. From (A.6) we have asymptotically (as $N \to \infty$)

$$\bar{f}^a(x;N) - P_{\mathcal{M}_a} f(x) \overset{d}{\sim} \mathcal{N}\left(0, \mathrm{Var}\left(\bar{f}^a(x;N)\right)\right). \tag{A.7}$$

From Proposition 3.3 of Unser and Daubechies (1997), for $x \in \mathbb{R}$,

$$|P_{\mathcal{M}_a} f(x) - f(x)| \le \|P_{\mathcal{M}_a} f - f\|_\infty \le \lambda_p \cdot \|f^{(p+1)}\|_\infty \cdot h^{p+1}, \tag{A.8}$$

where $\lambda_p < \infty$ is a constant independent of f, defined above. The result then follows immediately. This completes the proof. ■

Proof of Lemma 1. Given $f \in C_b^4(\mathbb{R})$, we have (To simplify notations, the domain of integration is understood as $\mathbb{R}$ when unspecified.) by Taylor's theorem with $n = 2$ that

$$f(x_k + hy)\,dy = \sum_{j=0}^{n+1} f^{(j)}(x_k)(hy)^j + \frac{f^{(n+2)}(\xi_k(y))}{(n+2)!}(hy)^{(n+2)},$$

for any fixed x_k and h, and some $\xi_k(y) \in [x_k, x_k + hy]$. Since $\widetilde{\varphi}_{a,k}(x)$ is bounded, $\mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)^2] < \infty$, and we have

$$\begin{aligned}
\mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)^2] &= \int (a^{1/2}\widetilde{\varphi}(a(y - x_k)))^2 f(y)dy \\
&= \int \widetilde{\varphi}^2(y) f\left(x_k + \frac{y}{a}\right) dy \\
&= \int \widetilde{\varphi}^2(y) \left(\sum_{j=0}^{n+1} f^{(j)}(x_k)(hy)^j + \frac{f^{(n+2)}(\xi_k(y))}{(n+2)!}(hy)^{(m+2)} \right) dy \\
&= \sum_{j=0}^{n+1} \frac{h^j f^{(j)}(x_k)}{j!} \int \widetilde{\varphi}^2(y) y^j dy + \int \widetilde{\varphi}^2(y) \frac{f^{(n+2)}(\xi_k(y))}{(n+2)!}(hy)^{(n+2)} dy \\
&= \sum_{j=0}^{n/2} \frac{h^{(2j)} f^{(2j)}(x_k)}{(2j)!} m_{2j}(\widetilde{\varphi}^2) + \zeta_k(hy),
\end{aligned} \tag{A.9}$$

where $m_j(f)$ is the jth moment of f. The final equality follows from the symmetry of $\widetilde{\varphi}$, and hence the symmetry of its square (whereby all odd moments $m_{2j+1}(\widetilde{\varphi}^2)$ vanish). In particular

$$\begin{aligned}
|\zeta_k(hy)| &\le \frac{h^{n+2}}{(n+2)!} \left| \int \widetilde{\varphi}^2(y) f^{(n+2)}(\xi_k(y)) \cdot y^{(n+2)} dy \right| \\
&\le h^{n+2} \frac{\|f^{(m+2)}\|_\infty}{(n+2)!} \int \widetilde{\varphi}^2(y) \cdot y^{(n+2)} dy = h^{n+2} \left(\frac{\|f^{(m+2)}\|_\infty}{(n+2)!} \cdot m_{n+2}(\widetilde{\varphi}^2) \right).
\end{aligned}$$

Hence, $\zeta_k(hy) = \mathcal{O}(h^{n+2}) = \mathcal{O}(h^4)$, uniformly in k and y. Simplifying the expression in (A.9), we have

$$\mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)^2] = f(x_k)R(\widetilde{\varphi}) + \frac{h^2}{2} f^{(2)}(x_k) m_2(\widetilde{\varphi}^2) + \mathcal{O}(h^4). \tag{A.10}$$

For our purpose, a second order expansion will suffice, which we combine with the unbiasedness of $\bar{\beta}_{a,k}(N)$ to conclude from (A.9) that

$$\begin{aligned}\bar{\sigma}^2_{a,k} &= \mathrm{Var}(\widetilde{\varphi}_{a,k}(X_1)) = \frac{1}{N}\left[\mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)^2] - \beta^2_{a,k}\right] \\ &= \frac{1}{N}\left[f(x_k)R(\widetilde{\varphi}) + \frac{h^2}{2} f^{(2)}(x_k) m_2(\widetilde{\varphi}^2) - \beta^2_{a,k} + \mathcal{O}(h^4)\right],\end{aligned} \tag{A.11}$$

where we have replaced $m_0(\widetilde{\varphi}^2)$ by the *roughness*, $R(\widetilde{\varphi}) := \int \widetilde{\varphi}^2(x)dx = \|\widetilde{\varphi}\|_2^2$, in order to maintain consistency with the literature.

Finally, note that $\beta_{a,k} = \mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)]$ has an identical expansion as $\mathbb{E}[\widetilde{\varphi}_{a,k}(X_1)^2]$, but with $m_{2j}(\widetilde{\varphi})$ in place of $m_{2j}(\widetilde{\varphi}^2)$, and with an overall multiplier of $h^{1/2}$. In particular, from this expansion for $\beta_{a,k}$, we find that

$$\beta^2_{a,k} = h \cdot (f(x_k)m_0(\widetilde{\varphi}))^2 + h^3 f^{(2)}(x_k) m_2(\widetilde{\varphi}) \cdot f(x_k) m_0(\widetilde{\varphi}) + \mathcal{O}(h^5). \tag{A.12}$$

Plugging (A.12) into (A.11), an expression for $\bar{\sigma}^2_{a,k}$ up to order $\mathcal{O}(h^3)$ is

$$\bar{\sigma}^2_{a,k} = \frac{1}{N}\left\{ f(x_k)R(\widetilde{\varphi}) - h \cdot (f(x_k)m_0(\widetilde{\varphi}))^2 + \frac{h^2}{2} f^{(2)}(x_k) m_2(\widetilde{\varphi}^2) + \mathcal{O}(h^3)\right\}.$$

This completes the proof. ■

Proof of Proposition 2.2. (i) First note that

$$\begin{aligned}\mathrm{Var}(\bar{f}^a(x;N)) &= \mathrm{Var}\Big(\sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}(N)\varphi_{a,k}(x)\Big) \\ &= \sum_{k\in\mathbb{Z}} \sum_{\substack{j\in\mathbb{Z} \\ |k-j|\le p}} \varphi_{a,k}(x)\varphi_{a,j}(x)\mathrm{Cov}\left(\bar{\beta}_{a,k}(N), \bar{\beta}_{a,j}(N)\right),\end{aligned}$$

since $\varphi_{a,k}(x)\varphi_{a,j}(x) = 0$ for $|k-j| > p$. For $p \ge 0$, with $\bar{\sigma}^2_{a,k} := \mathrm{Var}(\bar{\beta}_{a,k}(N))$, and $C_{a,j,k} := \mathrm{Cov}\left(\bar{\beta}_{a,k}(N), \bar{\beta}_{a,j}(N)\right)$, it holds that

$$\begin{aligned}\int \mathrm{Var}(\bar{f}^a(x;N))dx &= \int \sum_{k\in\mathbb{Z}} \sum_{0\le|k-j|\le p} \varphi_{a,k}(x)\varphi_{a,j}(x)C_{a,j,k}dx \\ &= \sum_{k\in\mathbb{Z}} \sum_{0\le|k-j|\le p} C_{a,j,k}\int \varphi_{a,k}(x)\varphi_{a,j}(x)dx \\ &= \sum_{0\le|m|\le p} \sum_{k\in\mathbb{Z}} C_{a,k,k-m}\int \varphi(x)\varphi(x-m)dx \\ &= \sum_{0\le|m|\le p} \int \varphi(x)\varphi(x-m)dx \sum_{k\in\mathbb{Z}} C_{a,k,k-m}.\end{aligned}$$

Using the bound

$$|\mathrm{Cov}(X,Y)| \le |\mathrm{Corr}(X,Y)|\sigma_X\sigma_Y \le \sigma_X\sigma_Y \le \max\{\sigma_X^2,\sigma_Y^2\} \le \sigma_X^2+\sigma_Y^2, \tag{A.13}$$

we obtain

$$\begin{aligned}
\int \mathrm{Var}(\bar{f}^a(x;N))dx &\le \sum_{0\le|m|\le p}\int \varphi(x)\varphi(x-m)dx \sum_{k\in\mathbb{Z}}\left(\bar{\sigma}_{a,k}^2+\bar{\sigma}_{a,k-m}^2\right)\\
&= 2\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2 \sum_{0\le|m|\le p}\int \varphi(x)\varphi(x-m)dx\\
&= 2\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2 \int \varphi(x)\Big(\sum_{0\le|m|\le p}\varphi(x-m)\Big)dx\\
&\le 2\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2 \int \varphi(x)dx = 2\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2,
\end{aligned}$$

since $\sum_{0\le|m|\le p}\varphi(x-m)\le 1$ on the domain of $\varphi(x)$ (which follows from the partition of unity property of φ). While this holds in general for $p\ge 0$, we can obtain a tighter bound in the special cases as follows. With $\delta_{m,0}=1$ if $m=0$, we have

$$\begin{aligned}
\int \mathrm{Var}(\bar{f}^a(x;N))dx &\le \sum_{0\le|m|\le p}\int \varphi(x)\varphi(x-m)dx \sum_{k\in\mathbb{Z}}\left(\bar{\sigma}_{a,k}^2+(1-\delta_{m,0})\bar{\sigma}_{a,k-m}^2\right)\\
&\le \sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2 \sum_{0\le|m|\le p}\int \varphi(x)\varphi(x-m)dx + \sum_{k\in\mathbb{Z}}\sum_{1\le|m|\le p}\bar{\sigma}_{a,k-m}^2\int \varphi(x)\varphi(x-m)dx\\
&= (\|\varphi\|_2^2+2\xi_p)\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2 + 2\sum_{1\le m\le p}\int \varphi(x)\varphi(x-m)dx\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k-m}^2\\
&= (\|\varphi\|_2^2+4\xi_p)\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2,
\end{aligned}$$

where $\xi_p := \sum_{1\le m\le p}\int\varphi(x)\varphi(x-m)dx$ is easily derived for any B-spline basis. For example, for the Haar basis, $\left(\|\varphi\|_2^2+4\xi_p\right)=(1+0)=1$, while for the linear basis, $\left(\|\varphi\|_2^2+4\xi\right)=(2/3+4/6)=4/3$.

(ii) For the second claim, note from (14) that

$$\begin{aligned}
\sum_{k\in\mathbb{Z}}\bar{\sigma}_{a,k}^2 &= \sum_{k\in\mathbb{Z}}\frac{1}{N}\Big\{f(x_k)R(\widetilde{\varphi})-h(f(x_k)m_0(\widetilde{\varphi}))^2+\frac{h^2}{2}f^{(2)}(x_k)m_2(\widetilde{\varphi}^2)\\
&\qquad +m_0(\widetilde{\varphi})m_2(\widetilde{\varphi})h^3f^{(2)}(x_k)f(x_k)+\mathcal{O}(h^4)\Big\},\\
&= \frac{R(\widetilde{\varphi})}{N}\frac{1}{h}\sum_{k\in\mathbb{Z}}f(x_k)h-\frac{(m_0(\widetilde{\varphi}))^2}{N}\sum_{k\in\mathbb{Z}}(f(x_k))^2h+\frac{m_2(\widetilde{\varphi}^2)h}{2N}\sum_{k\in\mathbb{Z}}f^{(2)}(x_k)h\\
&\qquad +\frac{m_0(\widetilde{\varphi})m_2(\widetilde{\varphi})h^2}{N}\sum_{k\in\mathbb{Z}}f^{(2)}(x_k)f(x_k)h+\mathcal{O}(h^3/N).
\end{aligned}$$

We next use the relation (see Redner, 1999, Lemma 6)

$$\sum_{k\in\mathbb{Z}}g(c_k)h=\int_{-\infty}^{\infty}g(x)dx+\mathcal{O}(h\|g'\|_1),$$

for g satisfying $g, g' \in L_1(\mathbb{R})$ and c_k in the kth interval $[x_k, x_{k+1}]$. It follows that

$$\sum_{k\in\mathbb{Z}} \bar{\sigma}_{a,k}^2 = \frac{R(\widetilde{\varphi})}{N}\frac{1}{h}\left(1+\mathcal{O}(h\|f'\|_1)\right) - \frac{(m_0(\widetilde{\varphi}))^2}{N}(R(f)+\mathcal{O}(h\|f\cdot f'\|_1)) + \frac{m_2(\widetilde{\varphi}^2)h}{2N}\left(\int f^{(2)}(x)dx + \mathcal{O}\left(h\|f^{(3)}\|_1\right)\right) + \mathcal{O}(h^2/N),$$

where we note that each of the integrals is finite by Assumption 4. In particular, $\|f\cdot f'\|_1 \le \|f\|_\infty \|f'\|_1$, so $f\cdot f' \in L^1$, and the result follows. This completes the proof. ■

Proof of Proposition 2.3. By definition,

$$\mathbb{E}[\bar{f}^a(x;N)] = \sum_{k\in\mathbb{Z}} \mathbb{E}[\bar{\beta}_{a,k}(N)]\varphi_{a,k}(x) = \sum_{k\in\mathbb{Z}} \beta_{a,k}\varphi_{a,k}(x) =: P_{\mathcal{M}_a} f(x),$$

where $\beta_{a,k}'s$ are the true projection coefficients, since $\bar{\beta}_{a,k}(N)$ is unbiased. That is, $\bar{f}^a(x;N)$ is an unbiased estimator of the true orthogonal projection $P_{\mathcal{M}_a} f(x)$. The analysis of Unser (1996) establishes the following estimate,

$$\|P_{\mathcal{M}_a} f - f\|_2 \le C_p^{1/2}\cdot\|f^{(p+1)}\|_2\cdot h^{p+1},$$

where C_p is a constant independent of f. Hence,

$$\begin{aligned}\int \left(\mathbb{E}[\bar{f}^a(x;N)] - f(x)\right)^2 dx &= \int \left(P_{\mathcal{M}_a} f(x) - f(x)\right)^2 dx \\ &= \|P_{\mathcal{M}_a} f - f\|_2^2 \\ &\le C_p\cdot\|f^{(p+1)}\|_2^2\cdot h^{2(p+1)},\end{aligned}$$

and the result follows. This completes the proof. ■

Proof of Proposition 2.4. Combining (16) and (18), we have that the MISE can be bounded asymptotically by

$$\begin{aligned}&\mathbb{E}\left[\int \left(\bar{f}^a(x;N) - f(x)\right)^2 dx\right] = \int_{\mathbb{R}} \mathrm{Var}(\bar{f}^a(x;N))dx + \int \mathrm{Bias}^2(\bar{f}^a(x;N))dx \\ &\le \theta_p \sum_{k\in\mathbb{Z}} \bar{\sigma}_{a,k}^2 + \bar{C}_p\cdot\|f^{(p+1)}\|_2^2\cdot h^{2(p+1)}, \\ &\le \theta_p\left(\frac{1}{h}\frac{R(\widetilde{\varphi})}{N} + \frac{R(\widetilde{\varphi})C(f') - (m_0(\widetilde{\varphi}))^2 R(f)}{N}\right) + \bar{C}_p\cdot\|f^{(p+1)}\|_2^2\cdot h^{2(p+1)} := \zeta(h),\end{aligned} \tag{A.14}$$

where $|C(f')| \le \lambda\|f'\|_1$ for some constant λ.

Differentiating w.r.t h and setting it to zero yields

$$\frac{\partial\zeta(h)}{\partial h} = \theta_p\frac{-1}{h^2}\frac{R(\widetilde{\varphi})}{N} + 2(p+1)\bar{C}_p\cdot\|f^{(p+1)}\|_2^2\cdot h^{2p+1} = 0,$$

from which

$$\theta_p \frac{R(\widetilde{\varphi})}{N} = 2(p+1)\bar{C}_p \cdot \|f^{(p+1)}\|_2^2 \cdot h^{2p+3}.$$

This can be solved uniquely for h, yielding (20). It also holds for all $h > 0$ that

$$\frac{\partial^2 \zeta(h)}{\partial h^2} = 2\theta_p \frac{1}{h^3} \frac{R(\widetilde{\varphi})}{N} + 2(p+1)(2p+1)\bar{C}_p \cdot \|f^{(p+1)}\|_2^2 \cdot h^{2p} > 0,$$

from which h_p^* is indeed the unique minimum of $\zeta(h)$. The optimal asymptotic mean integrated squared error follows by plugging (20) into the last equation of (A.14), and noting that for small h, $1/N$ is dominated by $1/(hN)$.

Lastly, by applying Jensen's inequality to (A.14), we have that for small $h > 0$

$$\mathbb{E}\big[\|\big(\bar{f}^a(x;N) - f(x)\big)\|_2\big] \le \Big(\mathbb{E}\Big[\int \big(\bar{f}^a(x;N) - f(x)\big)^2 dx\Big]\Big)^{1/2} \le [\zeta(h)]^{1/2},$$

hence, as $h \to 0$, $Nh \to \infty$ and $N \to \infty$, then $\bar{f}^a(x;N) \xrightarrow{L^2} f(x)$. The conclusion that $\bar{f}^a(x;N)$ is a consistent estimator of $f(x)$ follows from this. This completes the proof. ■

APPENDIX B: ECF Method Implementation Details

The ECF-based density estimator in (27) is determined upon evaluating the set of coefficients in (24) for a fixed basis of size N_φ. With bandwidth $h = x_k - x_{k-1}$ fixed, and $\Delta_\xi = 2\pi/(hN_\varphi)$ determined by the Nyquist frequency, coefficients are estimated based on a trapezoidal approximation of (24), with quadrature weights given by $v_j := 1 - (\delta_{j,1} + \delta_{j,N})/2$ and

$$\begin{aligned}\beta_{a,k} &\approx \frac{a^{-1/2}}{\pi} \Re\left\{\sum_{j=1}^{N_\varphi} \exp(-i x_k \xi_j) \cdot \phi_N(\xi_j) \cdot \widehat{\widetilde{\varphi}}\Big(\frac{\xi_j}{a}\Big) v_j \, \Delta_\xi\right\} \\ &= \frac{2a^{1/2}}{N_\varphi} \Re \sum_{j=1}^{N_\varphi} e^{-i\frac{2\pi}{N_\varphi}(j-1)(k-1)} F_j^N. \end{aligned} \tag{B.1}$$

We thus have $\beta_{a,k} \approx \widetilde{\beta}_{a,k}(N)$, where

$$\widetilde{\beta}_{a,k}(N) := \frac{2a^{1/2}}{N_\varphi} \Re\left\{\mathcal{D}_n\{F_j^N\}_{j=1}^{N_\varphi}\right\}, \quad F_j^N := \exp\big(-i x_1 \xi_j\big) \phi_N(\xi_j) \cdot \widehat{\widetilde{\varphi}}(\xi_j/a) v_j. \tag{B.2}$$

Here $\mathcal{D}$ denotes the discrete Fourier transform of a vector:

$$\mathcal{D}_n\{y_j\} := \sum_{j=1}^{N_\varphi} e^{-i\frac{2\pi}{N_\varphi}(j-1)(n-1)} y_j, \quad n = 1, \ldots, N_\varphi.$$

While N_φ is unrestricted, the FFT implementation is done most efficiently using a power of two. If needed, we can temporarily extend the basis beyond the desired N_φ, but keep only the coefficients up to N_φ. Unlike the dual approach, the discrete approximation here results in a biased estimator of $\beta_{a,k}$. However, the error caused by the integral approximation converges exponentially fast for a large class of processes, as discussed in Appendix B.2.

B.1. Choice of Parameters

We consider two types of domains for the density, $(-\infty,\infty)$ and $[l,\infty)$ (where the case of $(-\infty,u]$ follows analogously). Once the bandwidth h^* has been chosen based on the sample $\{X_n\}_{n=1}^N$, it remains to choose the truncated density support, $[l,u]$, and the number of basis elements, N_φ, as well as x_1, the leftmost basis grid point. Define $X_{(1)} := \min_{1\le n\le N} X_n$ and $X_{(N)} := \max_{1\le n\le N} X_n$. We start with the case of $(-\infty,\infty)$, and initially define $\tilde{\delta} = (1+\tau)\cdot(X_{(N)} - X_{(1)})$ for an amplification parameter $\tau \ge 0$ (a choice of $\tau = 1/10$ performs well in practice). To apply the FFT, we choose

$$P = \lceil \log_2(\tilde{\delta}/h^*)\rceil, \quad N_\varphi = 2^P.$$

Then $x_1 = (1+\tau/2)\cdot X_{(1)}$, and $\delta = (N_\varphi - 1)\cdot h^*$ is the truncated support width. The case of $[l,\infty)$ is handled similarly, with $\tilde{\delta} = (1+\tau)\cdot(X_{(N)} - l)$, and $x_1 = l$.

B.2. ECF Method Coefficient Bias

Our use of the ECF to estimate coefficients introduces a coefficient estimation bias when we employ numerical integration. In particular, the bias $\mathbb{E}[\widetilde{\beta}_{a,k}(N)] - \beta_{a,k} = \mathbb{E}[\mathcal{E}_k]$ where $\mathcal{E}_k$ can be decomposed as

$$\begin{aligned}
\mathcal{E}_k &= \frac{a^{-1/2}}{\pi}\Re\left\{\Delta_\xi \sum_{j=1}^{N_\varphi} v_j h_{a,k}(\xi_j)\cdot\phi_N(\xi_j) - \int_0^\infty h_{a,k}(\xi)\cdot\phi(\xi)d\xi\right\} \\
&= \frac{a^{-1/2}}{\pi}\Re\left\{\Delta_\xi {\sum_{j=1}^{\infty}}' h_{a,k}(\xi_j)\cdot\phi(\xi_j) - \int_0^\infty h_{a,k}(\xi)\cdot\phi(\xi)d\xi\right\} \\
&\quad - \frac{a^{-1/2}}{\pi}\Re\left\{\Delta_\xi {\sum_{j=N_\varphi}^{\infty}}' h_{a,k}(\xi_j)\cdot\phi(\xi_j)\right\} + \frac{a^{-1/2}}{\pi}\Re\left\{\Delta_\xi {\sum_{j=1}^{N_\varphi}}' h_{a,k}(\xi_j)\left(\phi_N(\xi_j) - \phi(\xi_j)\right)\right\} \\
&:= \mathcal{E}_k^1 + \mathcal{E}_k^2 + \mathcal{E}_k^3(N),
\end{aligned}$$

where the prime indicates the first term in the summation is halved, and $h_{a,k}(\xi) := \exp(-ix_k\xi)\cdot\widehat{\widetilde{\varphi}}\left(\frac{\xi}{a}\right)$. The third error source depends on the sample, but satisfies $\mathbb{E}[\mathcal{E}_k^3(N)] = 0$, from which

$$\mathbb{E}[\widetilde{\beta}_{a,k}(N)] - \beta_{a,k} = \mathcal{E}_k^1 + \mathcal{E}_k^2.$$

The first two error sources, $\mathcal{E}_k^1, \mathcal{E}_k^2$ are deterministic, and decay exponentially for a large class of processes, as demonstrated in Kirkby (2015). In practice, we find that the difference is negligible between the ECF estimator and the estimator based on the dual in Section 2.

B.3. Coefficient Estimation by Duality

This section briefly describes a low-cost alternative for coefficient estimation, which is promising for higher dimensional applications in future research. In general, $\widetilde{\varphi}$ does not have a closed-form expression, although it must belong to the closed span of $\{\varphi(x-m)\}_{m\in\mathcal{Z}}$,

$$\widetilde{\varphi}(x) = \sum_{m\in\mathbb{Z}} \alpha_m \varphi(x-m), \tag{B.3}$$

for a set of coefficients $\{\alpha_m\}$ with a finite $l^2(\mathbb{Z})$ norm, that is $\sum_{m\in\mathbb{Z}} \alpha_m^2 < \infty$. Coefficients for the linear basis are known in closed-form as follows.

LEMMA A.4 (Kirkby and Deng, 2019). *The coefficients α_m of the piecewise linear dual generator $\widetilde{\varphi}^{[1]}$ converge to zero exponentially in m. In particular,*

$$\widetilde{\varphi}^{[1]}(x) = \sum_{m\in\mathbb{Z}} \left(\frac{3}{\sqrt{3}} (\sqrt{3}-2)^{|m|} \right) \varphi^{[1]}(x-m). \tag{B.4}$$

To evaluate $\widetilde{\varphi}_{a,k}(X_n)$, we need to truncate the infinite series representation. Since the coefficients α_m decay exponentially, accurate approximations are obtained with a dozen or fewer coefficients. (For higher order B-splines, the coefficients of the dual scaling function can be obtained numerically.)

Rather than truncating the dual, Kirkby and Deng (2019) develop an alternative approach based on an *alternative bi-orthogonal sequence* (ABS). In particular, we can search for an alternative function that is still bi-orthogonal to the generator, but resides in a different space. (Recall that in order to be the true dual, a function must live in the same space as the generator.) For the linear basis, the ABS_2 generator of Kirkby and Deng (2019) satisfies

$$\breve{\varphi}(x) = \sum_{|m|\le 3} c_{|m|} \varphi^{[1]}(2x-m), \tag{B.5}$$

where $(c_0, c_1, c_2, c_3) = (2, 5/12, -1/2, 1/12)$. This generator lives in the span of functions at one higher order resolution than the dual, and produces equivalent approximations for polynomials of degrees three or less. Because of its narrow support, approximations using $\breve{\varphi}(x)$ are computationally inexpensive, requiring only a handful of evaluations per basis element (i.e., at a linear cost, $\mathcal{O}(N_\varphi)$). For higher dimensional tensor bases, the cost savings of this approach could be substantial. We refer the reader to Kirkby and Deng (2019) for more details on this procedure.

B.4. Plug-in Rule Details

Note that the Gaussian kernel satisfies the Chapman-Kolmogorov equation, and we can estimate $\|f^{(p)}\|_2^2$ as follows

$$\widehat{\|f^{(p)}\|_2^2} := \frac{(-1)^p}{N^2} \sum_{k=1}^{N} \sum_{m=1}^{N} \phi^{(2p)}(X_k, X_m; 2t_p), \tag{B.6}$$

where $\phi^{(2p)}(x, y; t)$ is the derivative with respect to the first variable x of $\phi(x, y; t) = \frac{1}{\sqrt{2\pi t}} e^{-(x-y)^2/(2t)}$. Next let $\widehat{{}_*t_p}$ be the optimal estimate for $\widehat{\|f^{(p)}\|_2^2}$. Note that the computation of $\widehat{\|f^{(p)}\|_2^2}$ requires the estimate of $\widehat{{}_*t_p}$ itself. From Botev et al. (2010), we have

$$\widehat{{}_*t_p} = \left(\frac{1 + 1/2^{p+1/2}}{3} \frac{1 \times 3 \times 5 \times \cdots \times (2p-1)}{N\sqrt{\pi/2}\, \widehat{\|f^{(p+1)}\|_2^2}} \right)^{2/(3+2p)} =: \gamma_p(\widehat{{}_*t_{p+1}}).$$

Let $\gamma^{[k]}(t) = \gamma_1(\ldots\gamma_{k-1}(\gamma_k(t))\ldots)$ with $k \ge 1$. In particular, Botev et al. (2010) suggest the following algorithm to estimate $\widehat{\|f^{(2)}\|_2^2}$ from data, which we recall here for the reader's convenience:

1. Initialize with $z_0 = \epsilon$, where ϵ is the machine precision, and $n = 0$.
2. Set $z_{n+1} = ((6\sqrt{2}-3)/7)^{2/5}\gamma^{[5]}(z_n)$.
3. If $|z_{n+1} - z_n| < \epsilon$, stop and set $\widehat{{}_*t_2} = \gamma^{[4]}(z_{n+1})$; otherwise, set $n := n+1$ and repeat from step 2.
4. Recover $\widehat{\|f^{(2)}\|_2^2}$ from (B.6).

In the numerical examples, for a general density $f(x)$, we will use the norm estimated from the above algorithm and plug it into (20) for the linear basis. (In practice, we find that a value of $\theta_1 = 1/4$ performs ideally for the linear basis (in place of $\theta_1 = 4/3$), which amounts to under-smoothing. We combine this with a spectral filter in Section 4 to reduce variance.)

B.5. Least Squares Cross-Validation

The approach of Rudemo (1982), Bowman (1984) known as least squares cross-validation (LSCV) is to choose h to minimize

$$\mathrm{LSCV}(h) := \int (\bar{f}^a(x;N))^2 dx - \frac{2}{N}\sum_{i=1}^{N} \log \bar{f}^a_{-i}(X_i), \tag{B.7}$$

and is also known as "unbiased cross-validation". (By Lemma A.3, LSCV(h) is finite.) For a pth order estimator in (7), we have

$$\begin{aligned}\int (\bar{f}^a(x;N))^2 dx &= \sum_{k\in\mathbb{Z}} \sum_{0\le|k-j|\le p} \bar{\beta}_{a,k}\bar{\beta}_{a,j} \int \varphi_{a,k}(x)\varphi_{a,j}(x)dx \\ &= \|\varphi\|_2^2 \sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}^2 + \sum_{1\le|m|\le p} \gamma_m \sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}\bar{\beta}_{a,k-m},\end{aligned}$$

where $\bar{\beta}_{a,k} = \bar{\beta}_{a,k}(N)$, $\gamma_m := \int \varphi(x)\varphi(x-m)dx$. For example, the linear B-spline basis yields

$$\int (\bar{f}^a(x;N))^2 dx = \tfrac{2}{3}\sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}^2 + \tfrac{1}{6}\sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}\bar{\beta}_{a,k-1} = \tfrac{1}{6}\sum_{k\in\mathbb{Z}} \bar{\beta}_{a,k}(4\bar{\beta}_{a,k} + \bar{\beta}_{a,k-1}).$$