Statistics Surveys
Vol. 17 (2023) 75–118
ISSN: 1935-7516
https://doi.org/10.1214/23-SS142

Spline local basis methods for nonparametric density estimation*

J. Lars Kirkby

School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, GA 30318,
United States
e-mail: jkirkby3@gatech.edu

Álvaro Leitao

Universitat Oberta de Catalunya (UOC),
08018 Barcelona,
Spain

CITIC research center,
15071 A Coruña,
Spain
e-mail: aleitao@uoc.edu

Duy Nguyen

Department of Mathematics,
Marist College,
Poughkeepsie, NY 12601,
United States
e-mail: nducduy@gmail.com

Abstract: This work reviews the literature on spline local basis methods for non-parametric density estimation. Particular attention is paid to B-spline density estimators which have experienced recent advances in both theory and methodology. These estimators occupy a very interesting space in statistics, which lies aptly at the cross-section of numerous statistical frameworks. New insights, experiments, and analyses are presented to cast the various estimation concepts in a unified context, while parallels and contrasts are drawn to the more familiar contexts of kernel density estimation. Unlike kernel density estimation, the study of local basis estimation is not yet fully mature, and this work also aims to highlight the gaps in existing literature which merit further investigation.

Received August 2022.

1. Introduction

Nonparametric density estimation is one of the most fruitful and permeating research areas in mathematical statistics, and the applications are numerous and

*Á. Leitao wishes to acknowledge the support received from the CITIC research centre, funded by Xunta de Galicia and the European Union (European Regional Development Fund, Galicia 2014-2020 Program) by grant ED431G 2019/01.

constantly expanding. Since the seminal works of [35], [95], [84], [11], the literature on density estimation has steadfastly evolved in response to the growing demand for such techniques. Although kernel density estimators (KDE) prevail as the principal estimation approach, alternatives such as orthogonal sequence estimators have received attention, for example, in [101], [117], [48], [112], [50], and more recently [75], [129]. While orthogonal sequence estimators generally rely on a global basis expansion of the (unknown) density, local density estimators using B-splines and wavelets have been studied in [92, 4, 16, 23, 86, 85, 56, 1]. Related techniques, such as smoothing splines [43, 41, 42], penalized B-splines (P-splines) [27], and logsplines [70, 67] have also been investigated.

Even though KDE remain the most popular choice for typical problems, there are a number of practical applications for which local basis representations offer attractive advantages over KDE. For example, B-spline basis expansions are especially useful for problems which require intensive and repeated numerical computations, which benefit from the basis representation of the density and the closed-form computations that it facilitates. Examples are common in insurance and financial risk management, where inference involves the computation of various risk measures, such as value-at-risk (quantile estimation), expected shortfall (conditional integration), and scenario analysis (nonparametric simulation). Each of these computations is simplified by a B-spline basis. For example, [17] derives closed-form expressions for various portfolio risk measures, such as value at risk (VaR) and expected shortfall. Similarly, [18] provides closed-form efficient simulation from the nonparametric density. Other recent examples in risk management and insurance include [116, 76, 127], each of which exploits the computational tractability of B-splines for numerically intensive applications. Recently, [26] considers the novel use of B-splines as nonparametric Bayesian priors, which are combined effectively with a Markov chain Monte Carlo scheme to sample from the posterior distribution. B-splines are also important building blocks of density uncertainty quantification, see [82]. The broader roles of splines in statics is surveyed in [119].

This work provides a comprehensive and unified perspective on the current state of density estimation by local basis methods. There are a number of effective local basis methods, summarized in Table 1 along with some representative literature, to be discussed in turn. We focus primarily on the case of B-spline density estimation, in part due to the simplicity and versatility of these bases, but perhaps more importantly for the role of such estimators in practical applications which require a balance of accuracy, tractability, and computational performance. Moreover, B-spline density estimation has enjoyed some recent developments in both theory and methodology. For example, it was proved in [16] that the mean integrated squared error (MISE) of B-spline density estimators approaches the optimal convergence rate of N^{-1} as the basis order is increased, where N is the sample size. A similar result is proved in [124] in the context of logarithmic B-spline estimators, see Corollary 2 therein. From a methodological perspective, [66] demonstrates that Galerkin methods, which are somewhat common in applied mathematics [14, 81], can also be a powerful tool for statistical inference. Special attention is also paid to the hybrid application of kernels

TABLE 1
Spline methods and representative literature.

Method	References
Logsplines	[70, 71, 72, 67, 124]
Smoothing Splines	[43, 41, 42]
P-splines	[27, 74, 30]
B-splines	[37, 92, 93, 80, 87]
B-spline Duality	[16, 17, 66]
Shape Constrained	[83]

and B-splines, through a regularization (variance reduction) technique known as spectral filtering which operates on the empirical characteristic function of a sample.

This work is organized as follows: the remainder of Section 1 introduces the problem setting and reviews the development of delta sequence and basis density estimators, as well as estimation in a transformed space (logsplines). Section 2 narrows focus to the case of B-spline density estimation, which is the primary topic of this work. We introduce the general framework for analyzing B-spline density estimation, and contrast five alternative estimation procedures. The theory of duality is discussed, which leads to the more recent advances in B-spline density estimation, with comparisons to kernel density estimators. A new procedure for efficient estimation in the context streaming data is proposed, setting the stage for more research in this area. The important problem of bandwidth selection is addressed in Section 3, where we cover likelihood cross-validation, least squares (unbiased) cross-validation, rule of thumb approaches, and plug-in estimation for B-spline estimators. Numerical examples are presented to illustrate the density estimates, and compare the various bandwidth selection approaches. We discuss applications in Section 4, with a particular focus on risk management, where density and quantile estimation are critical to the functioning of financial institutions. Finally, we conclude the work in Section 5, with a view towards future development of B-spline estimation theory and its applications.

1.1. The problem setting

Consider a sample $\{X_n\}_{n=1}^N$ which is drawn (i.i.d.) from some unknown continuous probability density with respect to the Lebesgue measure, $X \sim f$, where the support of f is $\mathbb{E} \subset \mathbb{R}$. Our goal is to estimate the unknown function f using the sample $\{X_n\}_{n=1}^N$. The following benign assumption ensures that our approximations are well defined.

Assumption 1.1. *We assume that* $f : \mathbb{E} \to \mathbb{R}$ *is a density satisfying* $\|f\|_1 := \int_{\mathbb{E}} |f(x)|dx = 1$, *and* $\|f\|_\infty < \infty$.

Note that whenever Assumption 1.1 holds, $\|f\|_2^2 \leq \|f\|_1 \|f\|_\infty = \|f\|_\infty < \infty$, so $f \in L^2(\mathbb{E})$, where the norm $\|g\|_2 = \sqrt{\langle g, g \rangle}$ is induced by the inner product $\langle g, h \rangle = \int_E g(x)h(x)dx$. Unless otherwise specified, we will take $\mathbb{E} \subset \mathbb{R}$. While there are many approaches to the density estimation problem, they can be

unified under a family of methods known as delta sequence estimators. Let $\delta_\lambda(x,y)$ be a bounded function of $x,y \in \mathbb{R}$, parameterized by a smoothing parameter $\lambda > 0$. We refer to $\{\delta_\lambda(x,y)\}_\lambda$ as 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 $\lambda \to \infty$. The delta sequence performs a smoothing operation on g in a neighborhood of the point x, but is able to recover the original function value in the limit as $\lambda \to \infty$. In the context of a sample of data of size N, we form the corresponding *delta sequence estimator* for f, $\widetilde{f}_\lambda(x;N) = \frac{1}{N}\sum_{n=1}^{N} \delta_\lambda(x,X_n)$, which provides a smoothed (regularized) representation of the data, with the strength of smoothing controlled by λ. The distinction between the various density estimators can be viewed as a difference in the functional form of $\delta_\lambda(x,y)$, and how the smoothing parameter λ is chosen based on the data. As we will discuss further below, the determination of λ is often viewed a more important than the form of $\delta_\lambda(x,y)$. Its choice is also a key factor that drives the computational cost of the estimation procedure, as well as its asymptotic (and small sample) properties.

The theory of delta sequence estimators dates back to the early works of [121, 113, 108], and encompasses many of the most common density estimators, including wavelets, orthogonal series estimators, and B-spline density estimators. Most notable is the *kernel density estimator (KDE)*, where for some $K(\nu) : \mathbb{R} \to \mathbb{R}$ with $\int_{-\infty}^{\infty} K(\nu)d\nu = 1$, we define $\delta_h(x,X_n) := \frac{1}{h}K((x-X_n)/h)$, which yields

$$\widetilde{f}_h(x;N) = \frac{1}{hN}\sum_{n=1}^{N} K\left(\frac{x-X_n}{h}\right). \tag{1.1}$$

Several types of kernel functions are commonly used, including uniform, triangular, and Epanechnikov [32], see [114]. Perhaps, the most common example is the Gaussian kernel $K(\nu) = (2\pi)^{-1/2}e^{-\nu^2/2}$. For a KDE, the degree of smoothing is determined by h. For larger values of h, the KDE estimate becomes smoother as the neighborhood around any x includes more sample points with non-negligible weight. This idea of a neighborhood of points will be even more explicit when we consider local basis estimators. Among the many excellent works on KDE, we invite the reader to refer to the review works of [118], [58], [105], [6], or the classic book of [114].

1.2. Loss functions and risk

To determine the effectiveness of a density estimator, one typically defines a *loss* function $L(f_h, f)$ such as the L_2 loss,

$$L_2(f_h,f) := \left\|\widetilde{f}_h - f\right\|_2^2 = \int_{\mathbb{E}} (\widetilde{f}_h(x;N) - f(x))^2 dx, \tag{1.2}$$

which captures the distance between the true density and the estimate for *a particular sample*. Another common loss function is the Kullback-Leibler loss,

$$\int_{\mathbb{E}} f(x)\log\left(\frac{f(x)}{\widetilde{f}_h(x;N)}\right)dx, \tag{1.3}$$

also known as relative entropy, see [51]. To accentuate the tail error, the Hellinger loss if often used, defined by $\int_{\mathbb{E}} \left(\sqrt{\widetilde{f}_h(x;N)} - \sqrt{f(x)}\right)^2 dx$. To obtain an estimate that averages the error over all possible samples, we define the *risk* $R(f_h, f) := \mathbb{E}[L(f_h, f)]$. The risk corresponding to the L_2 loss in (1.2) is known as the mean integrated squared error (MISE), and is commonly used to assess the fit quality of an estimator.

It is important to note that $R(f_h, f)$ is a *pointwise* metric, in the sense that it captures the expected loss for a particular density function, f. However, the true population density is unknown a-priori, and so it often makes sense to think about the maximum risk posed over a family $\mathcal{P}$ of densities, say $\sup_{f\in\mathcal{P}} R(f_h, f)$. Typically we choose $\mathcal{P}$ to impose some degree of smoothness on f, such a Sobolev or Besov space. The Sobolev spaces W_p^s for $s \in \mathbb{N}$ and $p \geq 1$ are a common choice, where $f \in W_p^s$ if and only if $\sum_{\nu=0}^{s} \left\|f^{(\nu)}\right\|_p < \infty$. The problem of *minimax* estimation is to determine $\inf_{f_h} \sup_{f\in\mathcal{P}} R(f_h, f)$, or at least some bounds thereon, see for example [123]. Research in this area includes [124, 24, 2, 123, 61, 122, 39, 40].

1.3. Density basis estimators

The study of basis expansion density estimators began in [101] with the *orthogonal series estimator*, which is defined in terms of an orthonormal basis $\{\Psi_k\}_{k\in\mathbb{Z}}$ for $L^2(\mathbb{E})$, where[1] $\mathbb{E} \subset \mathbb{R}$. Recall that any $f \in L^2(\mathbb{E})$ can be represented exactly (in the $L^2(\mathbb{E})$ sense) by its orthogonal projection onto an orthonormal basis[2]

$$\widetilde{f}(x) = \sum_{k=-\infty}^{\infty} \Psi_k(x) \int_{\mathbb{E}} \overline{\Psi_k(y)} f(y) dy. \tag{1.4}$$

We will refer to $\alpha_k = \int_{\mathbb{E}} \overline{\Psi_k(y)} f(y) dy$ as the *(basis) coefficients* of the orthogonal projection. Harmonic bases are common candidates, and are studied in the works of [101], [117], and [75]. Later, we will relax the orthogonality constraint on the basis when we consider *biorthogonal* density estimators. These estimators obey a similar representation as (1.4), but require the concept of a dual basis to compute the coefficients, which we discuss in Section 2.1.

If we define $\delta_M(x, X_n)$ by $\sum_{k=-M}^{M} \Psi_k(x)\overline{\Psi_k(X_n)}$, then we can formulate the estimator

$$\widetilde{f}_M(x;N) = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=-M}^{M} \Psi_k(x)\overline{\Psi_k(X_n)} := \sum_{k=-M}^{M} \overline{\alpha}_k \Psi_k(x), \tag{1.5}$$

where $\overline{\alpha}_k := \frac{1}{N}\sum_{n=1}^{N} \overline{\Psi_k(X_n)}$ is an unbiased estimate of the basis coefficient. The Hermite function approach of [101] is a prominent example with $\mathbb{E} = \mathbb{R}$,

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

[2] Here $\overline{z}$ represents the complex conjugate of the number $z \in \mathbb{C}$.

where $\Psi_k(x) := (2^k k\pi^{1/2})^{-1/2} e^{-x^2/2} H_k(x)$ for $k \in \mathbb{N}$, and function $H_k(x)$ is defined as $H_k(x) := (-1)^k e^{x^2}(d^k/dx^k)(e^{-x^2})$. For more literature on orthogonal sequence estimators, see [9], [48], [112], [50], and more recently [19, 75]. For orthogonal series estimators, the smoothness of the estimate is determined by M. Truncating the series at a small value of M will tend to produce smoother estimates, although the "optimal" estimation of M is a delicate problem, see [75] for an interesting regularization approach.

Remark 1 (KDE vs Basis Estimators). The KDE is interesting in the sense that it always retains the entire sample, and all sample points are required to evaluate $\tilde{f}_h(x; N)$, for any x (similar to the K-nearest neighbor estimator in statistical learning). By contrast, if we represent the unknown density in terms of a basis, we are able to compress the sample data into a small set of basis coefficients, which encapsulate all necessary information to perform subsequent computations with the estimated density. This also highlights that while KDE literature is primarily focused on bandwidth selection, coefficient estimation for local bases is also of great importance, with potentially many estimation approaches available for a given basis. It is also interesting to note that, in the multivariate estimation case, the convergence rate for orthogonal series estimators is independent of the dimension, whereas the KDE experiences slower convergence as the dimension increases, see [101].

1.4. Logarithmic basis estimators

The first use of splines in density estimation dates back to the *histospline* approach of [5, 111], which applies a cubic spline interpolation of the empirical cumulative distribution function (CDF). More recent approaches have sought to represent the density (or its logarithm) in a spline basis, with coefficients that are estimated, rather than determined via interpolation of the CDF. A prominent example is the *logsplines*, introduced in [73] as a way to estimate $\log(f)$ via a B-spline expansion. In the case where $f > 0$ on $\mathbb{E}$, we have via logistic density transform [77] that $f(x) = e^{g(x)} / \int_{\mathbb{E}} e^{g(y)} dy \propto e^{g(x)}$. In [73], the authors approximate $g(x)$ with a cubic B-spline basis, denoted $g_d(x) = \theta_1 B_1(x) + \cdots \theta_d B_d(x)$. That is,

$$f_d(x) := \exp\left(\theta_1 B_1(x) + \cdots \theta_d B_d(x) - c(\boldsymbol{\theta})\right) := \exp(g_d(x) - c(\boldsymbol{\theta})), \tag{1.6}$$

where $c(\boldsymbol{\theta}) = \log(\int_{\mathbb{E}} e^{g_d(x)} dx)$, with weights to be determined via maximum likelihood. In [68, 69], the authors prove a rate of convergence of $N^{-2\alpha/(2\alpha+1)}$, where α is the smoothness of $\log(f(x))$ in a Besov space, using the Kullback-Leibler loss in (1.3). By contrast to orthogonal sequence estimators for which the estimated density can be negative in some regions, the logspline approach ensures that $f_d(x)$ is positive by estimating coefficients of $\log(f_d(x))$. The key issue for logspline estimation is the determination of B-spline knots: how many knots to choose, and where to place them. The selection of knots can be likened to the bandwidth parameter h for a KDE, or the size of an orthogonal series basis M, see for example [73].

Regularization was later introduced to temper the estimates, using smoothing splines. The smoothing spline estimator of [44] is defined as the minimizer of the penalized log-likelihood score

$$-\frac{1}{N}\sum_{n=1}^{N} g_d(X_n) + \log \int_{\mathbb{E}} e^{g_d(x)} dx + \frac{\lambda}{2} J(g_d),$$

where the final term is a regularization designed to smooth the resulting B-spline estimator, hence the name smoothing splines. For example, the typical roughness penalty $J(g_d) = \int_{\mathbb{E}} [g_d''(x)]^2 dx$ is commonly applied with cubic splines. Higher order roughness penalties have also been considered in [28], which the authors refer to as P-splines (as in penalized-splines). In [99], another approach is proposed using truncated power functions, with a ridge penalty and knots based on quantiles, but [29] finds a strong preference for the P-spline approach. Another interesting idea is introduced in [22] under the name H-splines, as it applies a hybrid approach of regression and smoothing splines. This approach is shown to provide some computational advantages over smoothing splines. Related B-spline smoothing algorithms can be found in [62]. The extension to bivariate logsplines was developed in [67].

In Section 2, we will take an alternative perspective on the estimation problem for splines, which is more in line with harmonic basis and wavelet estimation, in that it attempts to estimate f directly via basis expansion. This approach evolved in parallel to the smoothing spline technique, and in many ways has been overlooked. From a computational tractability perspective, dealing with an estimator for $f(x)$ directly is more expedient than a representation of the form $f(x) = e^{g(x)} / \int_{\mathbb{E}} e^{g(x)} dx$. Moreover, while the study of P-splines and estimation in the logarithm space is comparatively more mature, see for example [31] for a recent survey, B-spline density expansion, especially via more recent developments using basis duality, is at a much earlier stage of development.

2. B-spline basis expansions of the density

Density estimation via B-spline basis expansion of the density shares features in common with each of the estimation methods described above. Like orthogonal series estimation, it approaches the problem as a *direct* expansion of the density in a basis. Like KDE, it seeks a localized representation of the data, by choosing compactly supported basis elements. Like logsplines, it utilizes a B-spline basis representation. However, it differs in fundamental ways from each of these related approaches. These differences lead to computational benefits, discussed in the current section and further in Section 3, as well as advantages in various practical applications, illustrated in Section 4.

In contrast to traditional orthogonal sequence estimations, such as Hermite or Fourier bases which utilize globally defined basis elements over $\mathbb{E}$, B-spline basis approximation relies on local (compactly supported) basis elements in a direct attempt to capture local features of the estimated density. They form a

particular type of basis estimator known as a *partition of unity*, see for example [94]. Taking the linear B-splines as a convenient and useful example, the basis is spanned by scaled and shifted versions of the linear *generator*,

$$\varphi(y) = \begin{cases} 1+y, & y \in [-1,0], \\ 1-y, & y \in [0,1]. \end{cases} \tag{2.1}$$

Note that φ is real-valued and symmetric. For a fixed *resolution* $a > 0$ that defines the *bandwidth* $h := 1/a$, basis elements $\varphi_{a,k}(x) := a^{1/2}\varphi(a(x - x_k))$ are centered over the points

$$x_k = x_1 + (k-1)h, \quad k \in \mathbb{Z}, \tag{2.2}$$

where x_1 is a shift parameter determined below. The earliest B-spline density estimator, developed by [37, 92, 93, 80] proposed to estimate f using the expansion

$$f(x) \approx \sum_{k=1}^{N_\varphi} \overline{\lambda}_{a,k}(N)\varphi_{a,k}(x), \tag{2.3}$$

where $\overline{\lambda}_{a,k}(N)$ is estimated using

$$\overline{\lambda}_{a,k}(N) = \frac{1}{N} \sum_{1 \le n \le N} \varphi_{a,k}(X_n). \tag{2.4}$$

This approach is reasonable, by analogy with orthogonal sequence estimators. The problem is that $\{\varphi_{a,k}\}$ are not orthogonal, so this estimator does not correspond to our typical notion of a basis projection. For large enough samples, and as $N_\varphi \to \infty$, each $\varphi_{a,k}$ will approach a Dirac delta function, and the estimates are consistent in the limit. For finite samples, however, a much more accurate methodology was developed in [16, 66] based on basis duality. We will refer to (2.4) as the *primal* basis method for coefficient estimation, to distinguish it from the dual basis method described next.

2.1. B-spline basis duality

While the B-spline bases $\{\varphi_{a,k}\}_{k\in\mathbb{Z}}$ are not orthogonal, they belong to a special class known as the Riesz bases. According to the duality theory of Riesz bases (see [13, 55, 125], there exists a *dual generator* $\widetilde{\varphi}$ such that the biorthogonal projection of any $f \in L^2(\mathbb{R})$ onto $\mathcal{M}_a := \overline{\text{span}}\{\varphi_{a,k}\}_{k\in\mathbb{Z}} \subset L^2(\mathbb{R})$ satisfies

$$P_{\mathcal{M}_a} f(y) = \sum_{k\in\mathbb{Z}} \beta_{a,k}\varphi_{a,k}(y), \tag{2.5}$$

where we can express $\beta_{a,k} := \int f(x)\widetilde{\varphi}_{a,k}(x)dx = \mathbb{E}[\widetilde{\varphi}_{a,k}(X)]$ for a density $f(x)$, and $\sum_{k\in\mathbb{Z}} \beta_{a,k}^2 < \infty$. That is, $\{\beta_{a,k}\}_{k\in\mathbb{Z}} \in l^2(\mathbb{Z})$. Also note that, similar to $\varphi_{a,k}(x)$, we define $\widetilde{\varphi}_{a,k}(x) := a^{1/2}\widetilde{\varphi}(a(x - x_k))$. While the linear and higher order B-spline bases are not orthogonal, they are *biorthogonal* in the sense that $\langle \varphi_{a,j}, \widetilde{\varphi}_{a,k} \rangle = \mathbf{1}_{\{j=k\}}$, for any $j, k \in \mathbb{Z}$. Here we use $\mathbf{1}_S$ to denote the indicator function of the set S.

Assumption 2.1. *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, $\widetilde{\varphi} : \mathbb{R} \to \mathbb{R}$.*

Note from (2.5) that $P_{\mathcal{M}_a} f(y)$ is completely determined by its coefficients $\beta_{a,k}$, for any chosen generator φ. As a result, to estimate the density function f from a sample, one needs to estimate the coefficients of $P_{\mathcal{M}_a} f(y)$ from the given sample. This motivates us to define the coefficient estimator

$$\overline{\beta}_{a,k}(N) := \frac{1}{N} \sum_{1 \le n \le N} \widetilde{\varphi}_{a,k}(X_n), \quad \forall k \in \mathbb{Z}. \tag{2.6}$$

It is immediate to see that $\overline{\beta}_{a,k}(N)$ is an unbiased estimator of the coefficients $\beta_{a,k} = \mathbb{E}[\widetilde{\varphi}_{a,k}(X)]$, where for each fixed k, $\{\widetilde{\varphi}_{a,k}(X_n)\}_{n=1}^{N}$ is a sequence of i.i.d random variables. Throughout, we will use $\overline{\beta}_{a,k}(N)$ and $\overline{\beta}_{a,k}$ interchangeably.

The density estimator $\overline{f}^a(x; N)$ proposed in [16] is defined by

$$\overline{f}^a(x; N) = \sum_{k \in \mathbb{Z}} \overline{\beta}_{a,k}(N) \varphi_{a,k}(x). \tag{2.7}$$

It is easy to see that

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

so $\overline{f}^a(x; N)$ is an unbiased estimator of the true orthogonal projection $P_{\mathcal{M}_a} f(x)$ in (2.5), and $\int \overline{f}^a(x; N) dx = 1$, as required of a density. Moreover, it can be shown that (see [16]) $\overline{f}^a(x; N) \to P_{\mathcal{M}_a} f(x)$ a.s. as $N \to \infty$. Clearly, to estimate $\overline{f}^a(x; N)$, we need to estimate the coefficients $\overline{\beta}_{a,k}(N)$, which is the key differentiator between the various B-spline estimators, as described next.

2.1.1. Alternative estimation procedures

There are several feasible approaches for estimating $\overline{\beta}_{a,k}(N)$, which vary in terms of accuracy and computational cost, and offer some relative advantages depending on the application:

1. Direct Primal Evaluation: the early works of [37, 92, 93, 80] estimate the coefficients using the primal basis $\varphi_{a,k}$ directly, recall (2.4). As previously mentioned, this approach is inaccurate compared with methods that utilize the dual basis, although it is fast and trivial to implement numerically.
2. Direct Dual Evaluation: from (2.26) below, we can directly represent the dual $\widetilde{\varphi}_{a,k}(x)$ in terms of an accurate finite series with $2M$ terms, allowing us to compute all $\overline{\beta}_{a,k}(N)$ coefficients at a cost $\mathcal{O}(2M \cdot N_\varphi \cdot N)$. This approach is feasible, but it can be made more efficient using either of the following two procedures. It does however lead to an efficient procedure for streaming data applications, discussed in Section 2.6.

3. Fourier Inversion: the approach of [16] is to compute the coefficients in the frequency domain, using the empirical characteristic function, at a cost of $\mathcal{O}(N_\varphi \log_2 N_\varphi + N_\varphi \cdot N)$. The advantage of this approach is that is enables regularization using a spectral filtering technique discussed in Section 2.3.
4. Galerkin Solution: the complexity is further reduced in [66] to $\mathcal{O}(N + N_\varphi) = \mathcal{O}(N)$ by employing a Galerkin-based approach, discussed in Section 2.5. This method achieves optimal complexity, produces nearly identical estimates as [16], and is very simple to implement thanks to a closed-form solution.
5. Maximum Likelihood: an alternative to closed-form coefficient estimates is to use a MLE with the likelihood $\frac{1}{N}\sum_{n=1}^{N} \log \overline{f}^a(X_n; N)$, at a cost of $\mathcal{O}(N \cdot K)$, where K is the number of iterations until convergence. This idea is tightly linked to the likelihood-based cross-validation approach for bandwidth selection, as discussed in Section 3.1.
6. Bona Fide Projection: a very recent approach to this problem, proposed by [87], is to solve a constrained convex optimization for the optimal coefficients. As discussed in Section 2.7, this has the advantage over alternative methods of ensuring that the density is everywhere positive (a bona fide density), although it no longer admits a fast closed-form solution. Using this approach, the estimation procedure is similar to logsplines, and also shares the advantage of a positive density.

Remark 2 (Comparison with Logsplines). Compared with the logspline estimator in (1.6), for which a numerical maximum likelihood optimization is the only feasible estimation procedure, the wide variety of estimation candidates for density expansion is one of its advantages, offering multiple alternatives to meet the needs of an application. For example, fast estimation is available when necessary, using the highly efficient Galerkin estimator, along with its fast bandwidth selection procedure. The cubic splines that are commonly applied in logspline estimation are also available to density expansion procedures, and in either case maximum likelihood (with non-negativity constraints) is applicable. Moreover, the logspline approach is not well suited to streaming data applications, discussed in Section 2.6. Finally we note the relative convenience of the basis expansion estimator, as it provides a closed-form CDF (Section 2.8) and quantile estimator (Section 4.1), each of which contribute to its efficiency and ease of use in applications.

2.1.2. B-splines and their duals

Our primary focus is on the B-spline generators (scaling functions), $\varphi^{[p]}$ for $p \geq 0$ (the B-spline order), where $\varphi^{[p]}$ is defined in (2.8) below. These generators are distinguished by their smoothness order (having piece-wise continuous, non-zero derivatives of order p). Compared with global basis expansions (e.g. Fourier series), the impact of sharp features of the density (such as peaks and rapidly changing curvature) is more localized, and they are better equipped to capture the fine details of a probability density function.

Starting with the *Haar* scaling function $\varphi^{[0]}(y) := \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(y)$, we define the p-th order B-spline scaling functions recursively via convolution

$$\varphi^{[p]}(x) = \varphi^{[0]} \star \varphi^{[p-1]}(x) = \int_{-\infty}^{\infty} \varphi^{[p-1]}(y-x)\mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(y)dy. \tag{2.8}$$

Note that (2.8) reduces to (2.1) when $p = 1$. Importantly, similar to (2.5), one can define the B-spline projection

$$P_{\mathcal{M}_a} f(y) = \sum_{k\in\mathbb{Z}} \beta_{a,k}\varphi_{a,k}^{[p]}(y), \tag{2.9}$$

where $\beta_{a,k} = \mathbb{E}[\widetilde{\varphi}_{a,k}^{[p]}(X)]$. As a result, one can estimate the probability density function using the same procedure as in (2.7), for any B-spline order.

We also note that, from Proposition 3.3 of [109], for $x \in \mathbb{R}$,

$$|P_{\mathcal{M}_a} f(y) - f(y)| \leq \|P_{\mathcal{M}_a} f - f\|_\infty \leq C_p h^{p+1},$$

for some positive constant C_p. Hence, the basis order directly controls the bias of the estimator, in a similar manner as the order of a kernel. While the bias decreases for higher order bases (and small h), the variance can become harder to control. Under certain assumptions on the choice of bandwidth, discussed further in Section 3.2, we can quantify the bias-variance tradeoff as a function of the B-spline order. A detailed analysis of the theoretical properties of B-spline estimators is provided in [16]. In practice, the choice of bandwidth h is often much more important than the basis order, and the linear basis is well-suited for many applications. A notable exception is when smoothness of the density is required, as needed to estimate derivatives of the density, see [92].

While there are several alternative definitions for the B-splines, for example the normalized splines considered in [92], we find this characterization preferable as it unites estimation techniques that operate in the frequency and physical domains. In particular, the convolution representation in (2.8) leads immediately to a characterization of the dual $\widetilde{\varphi}$ in the frequency domain.

Define the Fourier transform $\mathcal{F}[g](\xi) := \int e^{ix\xi} g(x)dx$, $\xi \in \mathbb{R}$. From [13], we can obtain the Fourier transform of the dual generator

$$\widehat{\widetilde{\varphi}}(\xi) := \frac{\widehat{\varphi}(\xi)}{\mathbf{\Phi}(\xi)}, \qquad \mathbf{\Phi}(\xi) := \sum_{k\in\mathbb{Z}} |\widehat{\varphi}(\xi + 2\pi k)|^2, \quad \xi \in \mathbb{R}, \tag{2.10}$$

where $\widehat{\varphi}(\xi) = \mathcal{F}\varphi(\xi)$, and $\widehat{\widetilde{\varphi}}(\xi) = \mathcal{F}\widetilde{\varphi}(\xi)$. Given that φ is compactly supported, $\mathbf{\Phi}(\xi)$ is a trigonometric polynomial with a finite cosine series expansion. From [63, 64], 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

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

Simplified expressions are given in [64] for orders $p = 0, \ldots, 3$, and the coefficients c_m in (2.27) can be easily evaluated. Note that, equipped with a closed-form for $\widehat{\widetilde{\varphi}}^{[p]}(\xi)$, we are able to estimate the B-spline coefficients using the empirical characteristic function of a sample, as discussed in Section 2.3.

2.2. The order of B-splines and kernels

While B-splines and kernels offer two very distinct estimation approaches, they do share some common theoretical links. For example, we can express $\overline{f}^a(x; N)$ from (2.7) via the following reproducing kernel representation[3]

$$\overline{f}^a(x; N) = \frac{1}{N} \sum_{1 \le n \le N} \left(\sum_{k \in \mathbb{Z}} \varphi_{a,k}(x) \widetilde{\varphi}_{a,k}(X_n) \right) = \frac{1}{hN} \sum_{1 \le n \le N} K\left(\frac{x}{h}, \frac{X_n}{h} \right), \tag{2.11}$$

which is reminiscent of (1.1), where

$$K(x, y) := \sum_{k \in \mathbb{Z}} \varphi(x-k)\widetilde{\varphi}(y-k). \tag{2.12}$$

Moreover, there is an interesting parallel to be made between the order of B-splines and kernels, which ultimately governs the convergence rate for sufficiently smooth functions.

Definition 2.1. *A kernel $K : \mathbb{R} \to \mathbb{R}$ is said to be of* order *p if it satisfies: $\int_{\mathbb{R}} K(u)du = 1$, $\int_{\mathbb{R}} u^j K(u)du = 0$ for $j = 1, \ldots, (p-1)$, and $\int_{\mathbb{R}} u^p K(u)du \neq 0$.*

For a p-th order kernel, and $f \in C_b^{(p+1)}$, from the Taylor's expansion, it follows that

$$\begin{aligned} \mathbb{E}[\widetilde{f}_h(x; N)] &= \int_{\mathbb{R}} K(u) f(x - hu) du \\ &= f(x) + \sum_{l=1}^{p} \frac{(-h)^l f^{(l)}(x)}{l!} \int_{\mathbb{R}} u^l K(u) du + o(h^p), \end{aligned} \tag{2.13}$$

so that a p-th order kernel has bias of order $\mathcal{O}(h^p)$. This results in an asymptotic mean integrated squared error (AMISE) of $O\left(R(f^{(p+1)})N^{-\frac{2p}{2p+1}}\right)$, given an optimally selected bandwidth, where $R(g) := \int_{\mathbb{R}} g^2(x)dx$ is the *roughness* of g. This relationship between the kernel order and its convergence rate has

[3] The idea of equivalent kernel representations for B-spline estimators is explored in [107].

been long understood, see [84, 96, 115]. The most common kernel estimators are second order ($p = 2$), such as the Epanechnikov and Gaussian kernel, which achieve AMISE of[4] $\mathcal{O}(R(f'')N^{-4/5})$, for $f \in C_b^2(\mathbb{R})$. All kernels of order $p > 2$ must take negative as well as positive values, so $N^{-4/5}$ is the fastest attainable convergence for a positive kernel.

By comparison, Proposition 2.4 of [16] proves 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. This demonstrates that for $f \in C^\infty(\mathbb{R})$, one can arbitrarily approach the optimal rate of $\mathcal{O}(N^{-1})$ by increasing the basis order, which is equivalent to increasing the basis smoothness. For KDE, the optimal rate is achieved (for example) by the family of "flat-top" kernels [88, 89], which are infinitely smooth. The logspline basis discussed in Section 1.4 also achieves the optimal rate for infinitely smooth functions as measured by the Kullback-Leibler risk, see [69].

By direct analogy to kernels, the only way for B-splines to beat the $N^{-4/5}$ convergence rate is for the dual scaling function to take negative values, hence permitting negativity of the estimates.[5] By utilizing the primal basis when computing B-spline coefficients, as in [92], one guarantees positivity of the estimator, but caps the convergence rate to $N^{-4/5}$, even for higher order B-splines (see Table 3 of [92]). What is further intriguing is that the linear B-splines achieve (theoretically) a convergence rate of $N^{-4/5}$ when coefficients are computed using *either* the primal or dual, where the linear dual is negative over part of its domain. In practice, the linear dual estimator converges at a faster rate than predicted, while the primal approaches converges much more slowly than predicted.

It is also interesting to note the relationship between the standard histogram estimator, where AMISE decays at a rate of $\mathcal{O}(R(f')N^{-2/3})$, and the Haar basis, which achieves the same convergence order. The Haar basis is the only *orthogonal* B-spline basis, and it is also the only B-spline with an everywhere non-negative dual. While the convergence rate is relatively poor for the Haar basis, it is still sometimes useful in applications due to its tractability [76].

2.3. The empirical fourier estimator

As previously discussed, the B-spline density estimator in (2.7) can be obtained in several ways, each offering trade-offs in terms of complexity, accuracy, and computational cost. The various approaches also illuminate the nature of the estimation problem in unique ways, and offer clues for designing new and improved procedures. We first review a dual basis method developed in [16] which utilizes the *empirical characteristic function* (ECF) of the sample. While this approach does not achieve the optimal complexity (i.e. $\mathcal{O}(N)$) of the estimator in [66], it facilitates a convenient and efficient regularization approach via

[4] While the rate of convergence is identical for all kernels of the same order, the constant governing convergence varies by kernel.

[5] But unlike kernels, it is trivial to ensure positivity of a B-spline basis (if needed) by simply flooring the coefficients to zero and re-normalizing.

spectral filtering, as described below, which provides some variance reduction in small to medium sample settings. Other applications of the ECF in statistics are provided in [126]. The ECF approach is well-suited for B-splines in a particular, because their characteristic functions are known in closed-form, along with those of the dual scaling function, recall (2.10).

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

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

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

$$\phi_N(\xi) := \frac{1}{N}\sum_{n=1}^{N} \exp(iX_n\xi) = \frac{1}{N}\sum_{n=1}^{N} \{\cos(X_n\xi) + i\sin(X_n\xi)\}. \tag{2.14}$$

Remark 3. We note that $\phi_N(\xi)$ is well defined and finite for $\xi \in \mathcal{I} \subset \mathbb{C}$ where $\mathcal{I}$ is a strip of the form: $\mathcal{I} := \{\xi \in \mathbb{C} | \Im(\xi) \in (A, B), \Re(\xi) \in \mathbb{R}\}$, for $A, B \in \mathbb{R}$ with $A < B$. By the Strong Law of Large Numbers, the ECF is a consistent estimator of the true characteristic function $\phi(\xi)$ at each point $\xi \in \mathcal{I}$. Moreover from [34], we have for any $0 < U \in \mathbb{R}$,

$$P\Big(\lim_{N\to\infty} \sup_{|\xi|\le U} |\phi_N(\xi) - \phi_X(\xi)| = 0\Big) = 1. \tag{2.15}$$

Theorem 2 of [15] showed that the above result holds true if U is replaced by $U_N \to \infty$, but this cannot be improved further in general.

With the help of the ECF, we can estimate the coefficients $\beta_{a,k}$ using the Fourier transform $\widehat{\widetilde{\varphi}} = \widehat{\widetilde{\varphi}}^{[p]}$ of the dual from (2.10). Using 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 can be shown to satisfy

$$\begin{aligned}\beta_{a,k} &= \frac{a^{-1/2}}{2\pi}\mathbb{E}\left[\int_{-\infty}^{\infty} \exp\left(i\xi(X - x_k)\right) \cdot \widehat{\widetilde{\varphi}}(-\xi/a)d\xi\right] \\ &= \frac{a^{-1/2}}{\pi}\Re\left\{\int_0^{\infty} \exp(-ix_k\xi) \cdot \phi_X(\xi) \cdot \widehat{\widetilde{\varphi}}(\xi/a)d\xi\right\}, \quad k \in \mathbb{Z}.\end{aligned} \tag{2.16}$$

The estimates for $\overline{\beta}_{a,k}(N)$ in (2.6) are obtained upon replacing $\phi_X(\xi)$ with the sample ECF, $\phi_N(\xi)$. That is,

$$\overline{\beta}_{a,k}(N) = \frac{a^{-1/2}}{\pi}\Re\left\{\int_0^{\infty} \exp(-ix_k\xi) \cdot \phi_N(\xi) \cdot \widehat{\widetilde{\varphi}}(\xi/a)d\xi\right\}, \quad k \in \mathbb{Z}. \tag{2.17}$$

Since $\phi_N(\xi)$ is an unbiased estimator of $\phi_X(\xi)$, i.e. $\mathbb{E}[\phi_N(\xi) - \phi_X(\xi)] = 0$, so is $\overline{\beta}_{a,k}(N)$ when estimated in this way. In fact, it can be shown (see [17]) that $\overline{\beta}_{a,k}(N) \to \beta_{a,k}$ a.s. as $N \to \infty$.

Remark 4 (Implementation). Practical implementation of the B-spline projection require a truncated domain, where we restrict the basis to N_φ basis elements centered over the points $x_k = x_1 + (k-1)h,\ k = 1, ..., N_\varphi$, where x_1 is the leftmost grid point. The truncated density support by $[l, u] = [x_1, x_{N_\varphi}]$ is chosen simply to cover the observed sample. As described in [16], the coefficients are then computed using the fast Fourier transform (FFT) at a cost of $\mathcal{O}(N_\varphi \log_2(N_\varphi))$. The dominant cost in this procedure is actually the evaluation of $\phi_N(\xi)$ in (2.14) for N_φ values of ξ, with a complexity of $\mathcal{O}(N \cdot N_\varphi)$.

2.4. Regularization by spectral filters

The ECF-based method described in the previous section offers a convenient and effective regularization technique for the estimation of coefficients using a spectral filter, at essentially no additional cost. This procedure is effective in small to medium sample settings at reducing the variance in coefficient estimation, which helps improve the quality of estimates. As further discussed below, this procedure can be thought of intuitively as a light kernel-smoothing applied to the sample prior to estimating the B-spline density projection.

Definition 2.2. *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* $0 \le l \le q-1$.

We form the spectrally filtered characteristic function, supported on the interval $[-2\pi a, 2\pi a]$, which is defined by $\widehat{\varphi}_N(\xi) := \Gamma_a(\xi)\phi(\xi)$, where $\Gamma_a(\xi) := \Gamma(\xi/(2\pi a))$. The regularized density is estimated by simply replacing the coefficients in (2.17) by those of the spectrally filtered ECF:

$$\overline{\beta}_{a,k}(N) \approx \frac{a^{-1/2}}{\pi} \Re \left[\int_0^{2\pi a} \exp(-i x_n \xi) \cdot \Gamma_a(\xi) \phi_N(\xi) \cdot \widehat{\widehat{\varphi}}^{[p]}\left(\frac{\xi}{a}\right) d\xi \right]. \tag{2.18}$$

Hence, the estimation procedure is the same as before, with a simple multiplicative adjustment made to the ECF. It is also interesting to note that the dual itself acts similarly to a filter on the ECF, via the multiplication $\phi_N(\xi) \cdot \widehat{\widehat{\varphi}}^{[p]}\left(\frac{\xi}{a}\right)$, while $\Gamma_a(\xi)$ provides additional smoothing.

Example 2.1. For practical purposes, we find that the exponential filter is quite effective (see for example [98, 16]). It is defined by $\Gamma(\xi) = \exp(-\tau \xi^q)$, where $\tau := \log \epsilon_m$ and ϵ_m is the machine precision epsilon. Figure 1 illustrates the effectiveness of spectral filtering to dampen the spurious oscillations of the ECF, due to sampling error. Outside of the frequency window $[-2\pi a, 2\pi a]$, the filtered ECF is zero, while the ECF to oscillate about zero. By contrast, the filtered ChF decays to zero (faster than any polynomial), and the corresponding density estimate is smooth. An interesting area for future research is to investigate the design of spectral filters for this estimation problem, as well as the theoretical properties of the filtered estimator.

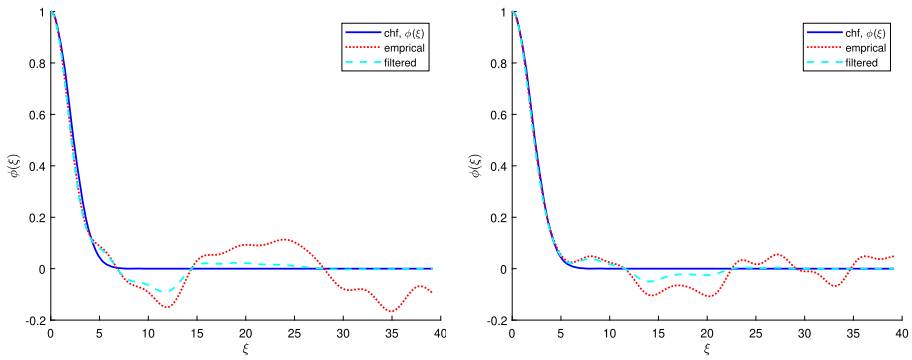


FIG 1. *Empirical characteristic function decay and spectral filtering with normal*$(0, 0.5)$ *data using an exponential filter of order* $q = 6$. *Left:* $N = 50$. *Right:* $N = 100$.

Remark 5 (Preservation of Moments). The spectral filter possesses an interesting property in the way it preserves moments of the sample. Let $\widehat{X}$ denote the filtered sample, defined implicitly by the filtered ECF $\widehat{\varphi}_N(\xi)$. From the binomial theorem,

$$\widehat{\varphi}_N^{(j)}(0) = \sum_{k=0}^{j} \binom{j}{k} \phi_N^{(j-k)}(0) \cdot \Gamma_a^{(k)}(0) = \phi_N^{(j)}(0) + \sum_{k=q}^{j} \binom{j}{k} \phi_N^{(j-k)}(0) \cdot \frac{\Gamma^{(k)}(0)}{(2\pi a)^k},$$

where $\Gamma_a^{(k)}(0) = \Gamma^{(k)}(0) \cdot (2\pi a)^{-k}$. In particular, using the fact that $\mathbb{E}[\widehat{X}^j] = i^{-j}\widehat{\varphi}_N^{(j)}(0)$, we have preservation of $\mathbb{E}[\widehat{X}^j] = \mathbb{E}[X^j]$ (in the sample) for $j = 1, \ldots, q-1$, and the added bias behaves as $(2\pi a)^{-q}$ for $j \geq q$. Moreover, as $a \to \infty$, we observe preservation of all moments in the limit. In practice, as $N \to \infty$, $a \to \infty$, which means that a feature of this regularization procedure is that it "deactivates" automatically as our sample size increases, regardless of the filter order.

2.4.1. Why does spectral filtering work?

Intuitively, spectral filtering replaces the empirical distribution with a (lightly) kernel-smoothed version prior to performing the projection. In terms of the theoretical density, at any continuity point we have

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ix\xi} \phi_X(\xi) d\xi \\ &\approx \frac{1}{2\pi} \int_{\mathbb{R}} e^{-ix\xi} \phi_X(\xi) \Gamma_a(\xi) d\xi \\ &= \int_{\mathbb{R}} \left[\mathcal{F}^{-1} \circ \phi_X(y)\right] \left[\mathcal{F}^{-1} \circ \{\Gamma_a(\cdot) e^{-ix\cdot}\}(y)\right] dy, \end{aligned}$$

by Parseval's identity. Hence, we can define the implicit kernel function

$$K(x) = \mathcal{F}^{-1} \circ \Gamma(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \Gamma(\xi) e^{-i\xi x} d\xi, \tag{2.19}$$

and let $\lambda := 1/2\pi a$, then we have

$$f(x) \approx \frac{1}{\lambda} \int_{\mathbb{R}} f(y) K\left(\frac{x-y}{\lambda}\right) dy := (f * K_\lambda)(x). \tag{2.20}$$

In other words, we can think of Γ in terms of its corresponding action in the physical domain through the convolution operator. Starting with a spectral filter Γ, we arrive at smoothed approximation to $f(x)$ given by $(f * K_\lambda)(x)$, where $K_\lambda : \mathbb{R} \to \mathbb{R}_+$ is the *mollifier* (kernel) defined by $K_\lambda(y) = K(y/\lambda)/\lambda$. The truncation parameter which determines the frequency cutoff corresponds to the bandwidth in the physical domain. Upon inverting the filtered ECF, we obtain

$$f(x) \approx \mathcal{F}^{-1} \circ \widehat{\varphi}_N(x) = \frac{1}{\lambda N} \sum_{n=1}^{N} K\left(\frac{x - X_n}{\lambda}\right),$$

which is the kernel density estimator defined in terms of $K = \mathcal{F}^{-1} \circ \Gamma$. In fact, K is actually a q-th order kernel (recall Definition 2.1), given the moment conditions on Γ. Higher order kernels (say $q \geq 6$) correspond to a low bias from (2.13) (and moment preservation, recall Remark 5), which is why we refer to this as a "light" filtering prior to coefficient estimation. Higher order spectral filters are effective at removing high frequency noise without adding too much bias to sample. Determining the optimal filter order is still an open problem.

2.5. The Galerkin estimator

A significant computational improvement over the ECF estimator is developed in [66], which utilizes a novel statistical Galerkin method to estimate the projection coefficients. Like the method of [16], this approach seeks to estimate the L^2 optimal projection of the density, but it does so over a compact interval $[l, u]$, with basis elements spanning the finite space $V_a := \text{span}\{\varphi_{a,k}\}_{k=1}^{N_\varphi}$. The bandwidth is given by $h = (u-l)/(N_\varphi - 1)$, which is related to the basis resolution by $a = (N_\varphi - 1)/(u-l)$.

If the density f were known, we could approximate it by the $L^2([l,u])$ projection $P_{V_a} : L^2([l,u]) \to V_a$,

$$f(x) \approx P_{V_a} f(x) = \sum_{1 \leq k \leq N_\varphi} \alpha_{a,k} \varphi_{a,k}(x).$$

As a projection, P_{V_a} must satisfy $P_{V_a} f - f \perp V_a$, or equivalently $P_{V_a} f - f \perp \varphi_{a,m}$ for $m = 1, ..., N_\varphi$. This leads to the set of *normal equations* $\langle P_{V_a} f - f, \varphi_{a,m} \rangle = 0, m = 1, ..., N_\varphi$, written equivalently as

$$\sum_{k=1}^{N_\varphi} \alpha_{a,k} \langle \varphi_{a,k}, \varphi_{a,m} \rangle = \langle f, \varphi_{a,m} \rangle, \quad m = 1, ..., N_\varphi. \tag{2.21}$$

Since f is a density function, we can replace the unknown integrals $\langle f, \varphi_{a,m} \rangle$ with their unbiased estimates to obtain

$$\langle f, \varphi_{a,m} \rangle = \mathbb{E}[\varphi_{a,m}(X)] \approx \frac{1}{N} \sum_{i=1}^{N} \varphi_{a,m}(X_i) =: \overline{\theta}_{a,m}(N)$$

which results in a solvable system of equations,

$$\sum_{k=1}^{N_\varphi} \overline{\alpha}_{a,k}(N) \langle \varphi_{a,k}, \varphi_{a,m} \rangle = \overline{\theta}_{a,m}(N), \quad m = 1, ..., N_\varphi, \tag{2.22}$$

with solution $\overline{\alpha}_{a,k}(N)$ depending on the sample. The system in (2.22) can be represented as

$$\boldsymbol{A}\overline{\boldsymbol{\alpha}} = \overline{\boldsymbol{\theta}}, \tag{2.23}$$

where $A_{m,k} := \langle \varphi_{a,k}, \varphi_{a,m} \rangle$, $k, m = 1, \ldots, N_\varphi$. The *Galerkin density estimator* is defined by

$$\overline{P}_{V_a} f(x) = \overline{\boldsymbol{\alpha}}^\top \boldsymbol{\varphi}_a(x) = \sum_{1 \le k \le N_\varphi} \overline{\alpha}_{a,k} \varphi_{a,k}(x). \tag{2.24}$$

From Lemma 3.1 and 3.2 of [66], the solution to (2.23) is well defined for any p-th order B-spline basis, and the solution is stable for any basis resolutions $a > 0$ due to $\left\|\boldsymbol{A}^{-1}\right\|_\infty \le C_{A,\infty}$.

For a p-th order B-spline basis, the matrix $\boldsymbol{A}$ in (2.23) is a banded symmetric matrix with bandwidth $\lfloor \frac{p+1}{2} \rfloor$, and the system $\boldsymbol{A}\overline{\boldsymbol{\alpha}} = \overline{\boldsymbol{\theta}}$ can be solved efficiently for the projection coefficients, $\overline{\boldsymbol{\alpha}}$. In particular, the cost is $\mathcal{O}(N_\varphi)$ operations, compared with $\mathcal{O}(N_\varphi^3)$ for a dense system. Moreover, the cost of calculating $\overline{\boldsymbol{\theta}}$ is $\mathcal{O}(N)$, resulting in an overall computational complexity of $\mathcal{O}(N_\varphi + N) = \mathcal{O}(N)$, which is optimal.

Of fundamental importance in the context of density estimation is the asymptotic normality of the estimator. The next result establishes this fact for densities with sufficient regularity, with an analogous result holding for the ECF estimator (see [16], Proposition 2.1).

Proposition 2.1 (Asymptotic Normality, [66], Proposition 3.3)**.** *Let $\varphi = \varphi^{[p]}$ be a pth order B-spline generator. Assume that $f \in C_b^{p+1}(\mathbb{R})$, and that $f(x) \sim C|x|^{-\gamma}$ for some $C > 0, \gamma > 0$ and as $|x| \to \infty$. Moreover, suppose that $N \to \infty, h \to 0$, and $Nh \to \infty$. Then*

$$\frac{\overline{P}_{V_a} f(x) - f(x) - \mu_p(h)}{\sqrt{Var\left(\overline{P}_{V_a} f(x)\right)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{2.25}$$

where the bias $\mu_p(h)$ satisfies $\mu_p(h) \le 2\lambda_p \left\|f^{(p+1)}\right\|_\infty h^{p+1} + \min\{|l|, |u|\}^{-\gamma}$. The variance is uniformly bounded for $x \in \mathbb{R}$,

$$\sup_{x \in \mathbb{R}} \{ Var(\overline{P}_{V_a} f(x)) \} \le \frac{\kappa}{Nh} \|f\|_\infty,$$

where $0 < \kappa < \infty$ is bounded uniformly in N, h.

2.5.1. Galerkin vs biorthogonal projection

The Galerkin density estimator can be thought of as a finite dimensional version of the biorthogonal projection in (2.5). Here we review the "closeness" between biorthogonal projection on $L^2(\mathbb{R})$ and Galerkin's projection on $L^2([l,u])$. Recall that $\{\varphi_{a,k}(x)\}_k$ and $\{\widetilde{\varphi}_{a,k}(x)\}_k$ are biorthogonal. It should then hold that

$$\begin{aligned}\beta_{a,j} = \mathbb{E}[\widetilde{\varphi}_{a,j}(X)] = \langle f, \widetilde{\varphi}_{a,j}\rangle &\approx \sum_{1\le k\le N_\varphi} \alpha_{a,k}\langle \varphi_{a,k}, \widetilde{\varphi}_{a,j}\rangle \\ &= \sum_{1\le k\le N_\varphi} \alpha_{a,k}\mathbf{1}_{\{j=k\}} = \alpha_{a,j},\end{aligned}$$

so $\beta_{a,j} \approx \alpha_{a,j}$. As a result, we have $(P_{\mathcal{M}_a} f)|_{[l,u]} \approx P_{V_a} f$. This is made mathematically rigorous in Theorem 3.4 of [66], which shows that $\|P_{\mathcal{M}_a} f - P_{V_a} f\|_2 \le 2C_p \left\|f^{(p+1)}\right\|_2 h^{p+1} + \tau(l,u)$, where $\tau(l,u) \to 0$ as $[l,u]$ covers $(-\infty,\infty)$. In particular, the $L^2([l,u])$ and $L^2(\mathbb{R})$ approximations become arbitrarily close. In any finite sample, as long as $[l,u]$ covers the observed sample, the two estimators are essentially identical in practice.

2.5.2. Regularization

Here we briefly remark on how regularization can be applied with the Galerkin approach, similar to using a spectral filter with the ECF method. Recalling from Section 2.4 that projecting the spectrally-filtered ECF is equivalent to projecting a (lightly) kernel-smoothed estimator (in the physical domain), instead of solving $\boldsymbol{A}\overline{\boldsymbol{\alpha}} = \overline{\boldsymbol{\theta}}$, we solve $\boldsymbol{A}\overline{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\theta}}$, where $\widetilde{\boldsymbol{\theta}}$ are kernel-smoothed mean estimates. That is, rather than $\langle f, \varphi_{a,m}\rangle \approx \frac{1}{N}\sum_{i=1}^N \varphi_{a,m}(X_i)$, we can calculate

$$\langle f, \varphi_{a,m}\rangle \approx \int_{I_{a,m}} \varphi_{a,m}(x)\widetilde{f}_\lambda(x;N)dx$$

where $f_\lambda(x;N)$ is defined in (1.1), but with the smoothing/regularization bandwidth parameter λ.[6] Further development of the Galerkin approach to include regularization is a promising research direction, as it could capitalize on the optimal complexity of this method, but also incorporate the main comparative advantage provided by the ECF approach of Section 2.3.

2.5.3. Estimation examples

We now illustrate the density estimation methodology, applying for concreteness the Galerkin linear basis estimator of the previous section, together with the Least-Squares Cross Validation (LSCV) bandwidth selection method, to be

[6] The integral can be computed efficiently by a trapezoidal (3 point) approximation on each subdomain, for a total of 5 points per basis element, to maintain the overall optimal complexity.

discussed in Section 3.1. Several conceptually different distributions are selected, whose definition and parameter configuration is presented in Table 2. In Figure 2, the estimations of the selected densities for a range of sample sizes are depicted. As N increases from 10^3 to 10^5, we observe convergence of estimator to the true density. Later experiments confirm the convergence in terms of the MISE, and show that it is faster than theoretically expected.

TABLE 2
Test case densities for numerical experiments.

Test Case	Density	Test Case	Density
Normal	$\mathcal{N}(0,1)$	Claw	$\frac{1}{2}\mathcal{N}(0,1) + \sum_{k=0}^{4} \frac{1}{10}\mathcal{N}\left(\frac{k}{2} - 1, \left(\frac{1}{10}\right)^2\right)$
Gamma	$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}, \quad x > 0,$ $\theta = \frac{1}{2}, k = 9$	Skewed Bimodal	$\frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}\left(\frac{3}{2}, \left(\frac{1}{3}\right)^2\right)$
Separated Bimodal	$\frac{1}{2}\mathcal{N}\left(-2, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}\mathcal{N}\left(2, \left(\frac{1}{2}\right)^2\right)$	Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp(-(x/\lambda)^k), \quad x \geq 0$ $\lambda = 1, k = 5$

2.6. Streaming data and direct dual estimation

The Galerkin and ECF approaches are similar in that they both estimate all coefficients simultaneously. This is perfectly reasonable for common estimation problems, but it is not well-suited for applications with streaming data when estimation speed is critical. This section proposes a new alternative approach which is ideally suited for streaming data applications, allowing the density estimator to update efficiently as new information arrives.

We first note that we can utilize $\widehat{\widetilde{\varphi}}$ defined in (2.10) to obtain an exponentially convergent series expansion of $\widetilde{\varphi}$. This also leads to a fast cross-validation method for determining the estimator bandwidth, discussed in Section 3.1. From Theorem 3.1 of [66], we have the following representation of the dual, given its Fourier transform $\widehat{\widetilde{\varphi}}$:

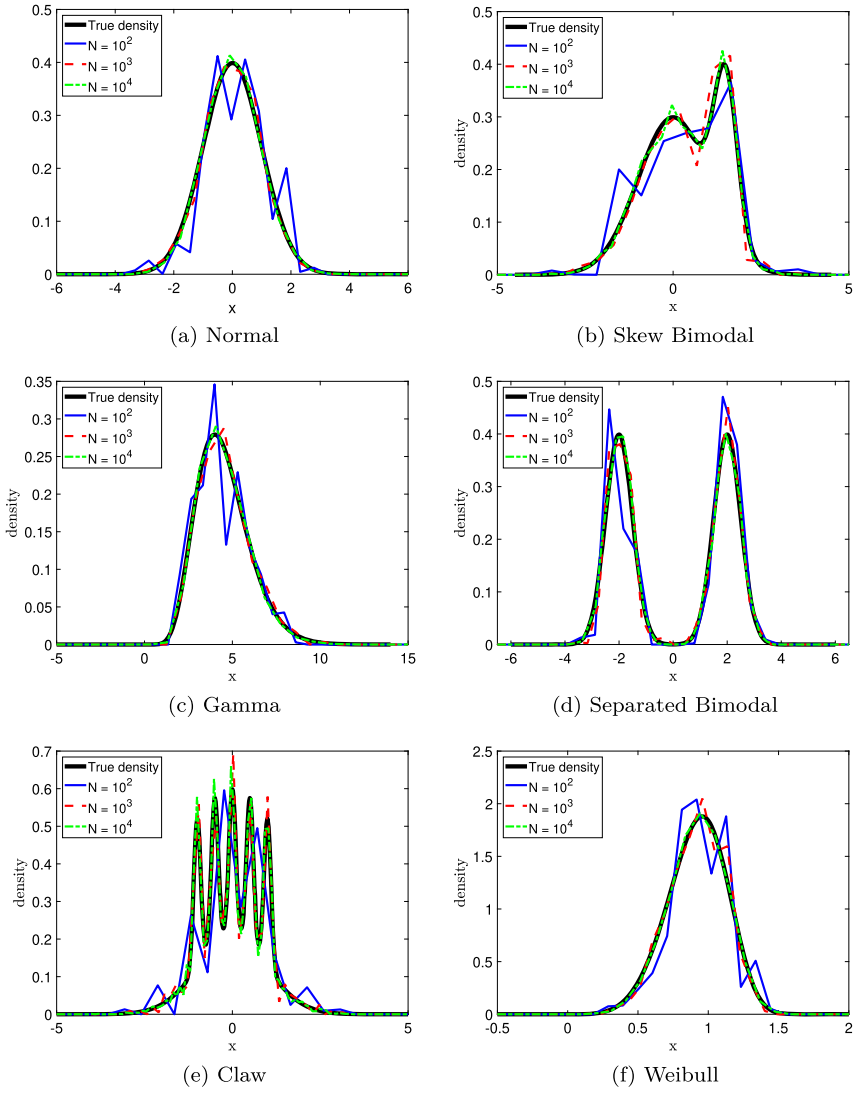
$$\widetilde{\varphi}_{a,k}(x) = \sum_{m\in\mathbb{Z}} c_m \varphi_{a,k-m}(x) \in L^2(\mathbb{R}), \qquad k \in \mathbb{Z}, \tag{2.26}$$

where $\{c_m\} \in l^2(\mathbb{Z})$ are given by

$$c_m = \frac{1}{\pi}\int_0^\infty \left(\widehat{\widetilde{\varphi}}(\xi)\right)^2 \cos(m\xi)d\xi = \frac{1}{\pi}\int_0^\infty \cos(m\xi)\frac{\widehat{\varphi}^2(\xi)}{\boldsymbol{\Phi}^2(\xi)}d\xi. \tag{2.27}$$

As mentioned above, we can use (2.26) directly to estimate the basis coefficients, after we truncate the infinite series,

$$\overline{\beta}_{a,k}(N) = \frac{1}{N}\sum_{1\leq n\leq N} \widetilde{\varphi}_{a,k}(X_n) = \sum_{m\in\mathbb{Z}} c_m \left(\frac{1}{N}\sum_{1\leq n\leq N} \varphi_{a,k-m}(X_n)\right). \tag{2.28}$$



(a) Normal (b) Skew Bimodal

(c) Gamma (d) Separated Bimodal

(e) Claw (f) Weibull

FIG 2. *Linear B-spline density estimates in terms of the number of samples, N.*

Rather than truncating the dual in (2.28), in the context of function approximation [65] develop an approach based on an *alternative bi-orthogonal sequence* (ABS), which leads to a more efficient implementation yet preserves the convergence rate. In particular, they consider a sequence that is bi-orthogonal to $\{\varphi_{a,k}\}$, but resides in a different space[7]. They define an ABS$_\gamma$ generator which

[7]Recall that in order to be the true dual, a function must live in the same space as the generator.

is supported on $[-\gamma, \gamma]$ and for which all moments $p \leq 2\gamma - 1$ coincide with the true dual (in addition to all odd moments), see [65] Proposition 5.4. For example, the ABS_2 generator for the linear basis satisfies

$$\breve{\varphi}^{[1]}(x) = \sum_{|m|\leq 3} c_{|m|}\varphi^{[1]}(2x - m), \tag{2.29}$$

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 $\breve{\varphi}^{[1]}(x)$ are computationally inexpensive, requiring only a handful of evaluations per basis element. For higher dimensional tensor bases, the cost savings of this approach could be substantial. Figure 3 illustrates the ABS_1 and ABS_2 generators, in addition to the exact dual.

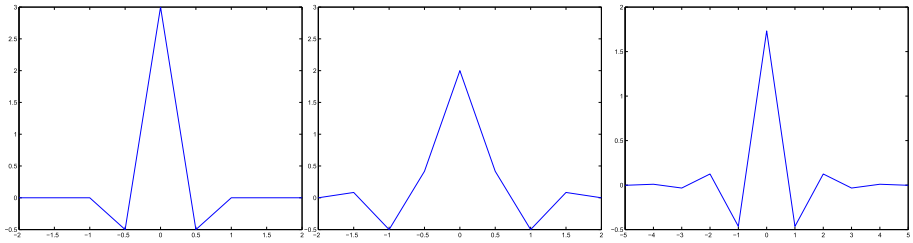


FIG 3. *Left: ABS_1 generator supported on $[-1, 1]$. Middle: ABS_2 generator on $[-2, 2]$. Right: true dual $\widetilde{\varphi}^{[1]}$, supported on $(-\infty, \infty)$.*

2.6.1. Streaming density estimation

Given a compactly supported ABS, such as in (2.29), we can update the projection coefficients using an on-the-fly update

$$\begin{aligned}\breve{\beta}_{a,k}(N) &:= \frac{1}{N}\sum_{1\leq n\leq N}\breve{\varphi}_{a,k}(X_n) \\ &= \frac{N-1}{N}\cdot\breve{\beta}_{a,k}(N-1) + \frac{1}{N}\cdot\breve{\varphi}_{a,k}(X_n).\end{aligned} \tag{2.30}$$

As N becomes large, the contribution of each new $\breve{\varphi}_{a,k}(X_n)$ diminishes, to the point where the estimates will no longer reflect new information. This is clearly a problem if the distribution of $\{X_n\}$ is changing over time, as our estimator should adapt with it. As pointed out in [10], for nonstationary data, we can consider an update rule of the form

$$\breve{\beta}_{a,k}(N) = \theta\cdot\breve{\beta}_{a,k}(N-1) + (1-\theta)\cdot\breve{\varphi}_{a,k}(X_n). \tag{2.31}$$

As $\theta \in (0, 1)$ is fixed, we have $1 - \theta > 0$ and the contribution of $\breve{\varphi}_{a,k}(X_n)$ never ceases to update the coefficient of $\breve{\beta}_{a,k}(N)$. Moreover, each update is performed

efficiently at a cost of $\mathcal{O}(1)$, since X_n falls in the support of only a handful of $\breve{\varphi}_{a,k}$, which makes this approach ideal for streaming data applications.

We also note that the proposed method can be combined with the sliding window technique developed in [36], which proposes the sliding window estimator for $w \in \mathbb{N}$,

$$\breve{\beta}_{a,k}(N) = \frac{N-1}{N} \cdot \breve{\beta}_{a,k}(N-1) + \frac{1}{w}\left(\breve{\varphi}_{a,k}(X_n) - \breve{\varphi}_{a,k}(X_{n-w})\right), \tag{2.32}$$

and is more suitable for data which arrives naturally in blocks, such as the case of environmental monitoring.

2.7. Bona fide estimators

Similar to higher order kernels, higher order B-splines are not guaranteed to produce a non-negative density estimate. For B-splines, a simple procedure of flooring the basis coefficients, and redistributing probability mass to ensure integration to one works well in many applications, see [17]. An alternative procedure is proposed in [87], which solves a convex optimization problem for the coefficients, subject to the constraints that the density integrates to one and is non-negative. Let p_δ denote the empirical measure,

$$p_\delta(x) = \frac{1}{N}\sum_{n=1}^{N} \delta_{X_n}(x), \tag{2.33}$$

where $\delta_{X_n}(x)$ is the Dirac Delta function centered over data point X_n. We can then solve the following problem,

$$\begin{aligned} \tilde{f}_+^a := \operatorname{argmin}_{\tilde{f}\in\mathcal{M}_a} &\left\{ \left\| \langle p_\delta, \varphi_k^a\rangle - \langle \varphi_k^a, \tilde{f}\rangle \right\|_{l_2}^2 \right\} \\ \text{s.t.} \quad \tilde{f}(x) \geq 0, x \in \mathbb{R}, &\quad \int_{\mathbb{R}} \tilde{f}(x)dx = 1. \end{aligned} \tag{2.34}$$

This estimator, coined the "Bona Fide Projection", is also promising for applications, and (2.34) is solvable using standard quadratic programming techniques, as demonstrated in Section 3 of [87].

2.8. Distribution function estimation

Before moving on to the topic of bandwidth selection, we briefly discuss the related problem of estimating the cumulative distribution function (CDF). Naturally, each of the density estimation methods described above yields a CDF estimator. Given $\overline{f}^a(y;N) := \sum_{k=1}^{N_\varphi} \overline{\beta}_{a,k}(N)\varphi_{a,k}(x)$, restricted to the set of N_φ coefficients that overlap the sample, define $\overline{F}^a(x;N) := \int_{-\infty}^{x} \overline{f}^a(y;N)dy$. From [17], Proposition 2.2, $\overline{F}^a(x;N)$ converges uniformly to $F(x)$ under mild assumptions,

$$\sup_{x\in\mathbb{R}} |\overline{F}^a(x;N) - F(x)| \to 0 \quad a.s \quad \text{as} \quad N \to \infty, a \to \infty.$$

Moreover, closed-form expressions are easily derivable for the B-spline bases. For example, the linear basis has a simple closed-form expression for $x \in \mathbb{R}$,

$$\overline{F}^a(x;N) = a^{-1/2}\left(\sum_{j\le k^*-1}\overline{\beta}_{a,j} + \overline{\beta}_{a,k^*}\left(\frac{1}{2}+\zeta-\frac{\zeta^2}{2}\right) + \frac{\zeta^2}{2}\overline{\beta}_{a,k^*+1}\right), \quad (2.35)$$

where $k^* := \lfloor (x-x_1)/a \rfloor + 1$, and $\zeta := a(x - x_{k^*})$. Formulas for higher order B-splines can also be derived. While many properties of the B-spline *density* estimators are documented in [92, 16, 66], theoretical properties of the CDF estimators are not yet established, which offers an interesting area for further research. In Section 4.2, we provide an application of the B-spline CDF estimator to nonparametric simulation.

3. B-spline bandwidth selection

Of equal (or greater) importance to the estimation of B-spline coefficients is the bandwidth selection procedure. There have been significant research efforts in determining the optimal choice of bandwidth in the literature. Early approaches can be found in [52], [106], [59], and references therein. A fairly comprehensive comparison of techniques for KDE is conducted in [12]. In general, the appropriate selection rule is application dependent, and no single bandwidth approach is universally preferred, [79]. This section reviews two complementary approaches to bandwidth selection for B-splines. The first approach, which is appropriate for all sample sizes, is based on a closed-form cross-validation formula for the B-spline bases. The second approach, which is ideal for larger sample sizes, chooses the bandwidth to optimize the asymptotic mean integrated squared error (asymptotic MISE).

3.1. Bandwidth selection by cross validation

A general approach to bandwidth selection is based on cross-validation, as discussed in [49, 8], for example. Specifically, likelihood-based cross-validation (see [25] for more details) chooses h to minimize the average log-likelihood

$$\mathrm{LCV}(h) = -\frac{1}{N}\sum_{i=1}^{N}\log \overline{f}^a_{-i}(X_i; N-1),$$

where $\overline{f}^a_{-i}$ is the estimator formed from the size $N-1$ sample with X_i removed. That is

$$\overline{f}^a_{-i}(X_i;N-1) = \sum_{k\in\mathbb{Z}}\overline{\beta}^{-i}_{a,k}(N-1)\varphi_{a,k}(X_i), \quad (3.1)$$

where $\overline{\beta}^{-i}_{a,k}(N-1) = \frac{1}{N-1}\sum_{1\le n\le N, n\ne i}\widetilde{\varphi}_{a,k}(X_n)$. This approach minimizes the Kullback-Leibler distance between the true and estimated density. Let $\varphi^{[p]}$ be

a p-th order B-spline generator, with dual generator $\widetilde{\varphi}^{[p]}$. From Lemma 4.1 of [66],

$$\mathrm{LCV}(h) = -\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{q \in \mathcal{K}(p)} \left(N\overline{\beta}_{a,k(i)+q} - \sum_{m \in \mathcal{K}(p)-q} c_m \varphi^{[p]}_{a,k(i)+q+m}(X_i) \right) \cdot \varphi^{[p]}_{a,k(i)+q}(X_i), \tag{3.2}$$

where $\gamma_m := \int \varphi^{[p]}(x)\varphi^{[p]}(x-m)dx$, c_m are the coefficients of the dual generator in (2.26), and $k(i) := \lfloor (X_i - l)/h + 1 \rfloor$ is the grid point left of X_i. The set $\mathcal{K}(p) \subset \{-\lceil \frac{p+1}{2} \rceil + 1, \ldots, \lceil \frac{p+1}{2} \rceil\}$ is the is the set of potentially affected coefficients.

The related approach, proposed in [97, 7] and known as least squares cross-validation (LSCV), chooses h to minimize

$$\mathrm{LSCV}(h) := \int (\overline{f}^a(x;N))^2 dx - \frac{2}{N} \sum_{i=1}^{N} \overline{f}^a_{-i}(X_i; N-1). \tag{3.3}$$

This approach is commonly referred to as "unbiased cross-validation", see [102]. It follows that

$$\mathrm{LSCV}(h) = \sum_{k \in \mathbb{Z}} \overline{\beta}_{a,k} \left(\left\| \varphi^{[p]} \right\|_2^2 \cdot \overline{\beta}_{a,k} + \sum_{1 \le m \le p} \gamma_m \left(\overline{\beta}_{a,k+m} + \overline{\beta}_{a,k-m}(N) \right) \right) + 2 \cdot \mathrm{LCV}(h). \tag{3.4}$$

Simplified closed-form expressions for $\mathrm{LSCV}(h)$ are provided for both the Haar and Linear basis (Theorem 4.1 and Theorem 4.2) in [66]. The formulas for $\mathrm{LCV}(h)$ and $\mathrm{LSCV}(h)$ are estimated with optimal efficiency in a single pass through the data. Hence, the cost to evaluate either is just $\mathcal{O}(N_\varphi + N) = \mathcal{O}(N)$, which compares favorably to $\mathcal{O}(N^2)$ for a kernel density estimator [105] or an orthogonal sequence estimator. While the LCV method has some nice theoretical properties, it is commonly observed to under-smooth in practice, and tends to under-perform the LSCV method, as demonstrated empirically below.

3.2. Plugin bandwidth with asymptotic MISE

For moderate to large sample sizes, an asymptotically optimal bandwidth can be determined by minimizing the asymptotic MISE of the estimator. Recall that

$$\begin{aligned} \mathrm{MISE} &:= \mathbb{E}\left[\int \left(\overline{f}^a(x;N) - f(x) \right)^2 dx \right] \\ &= \int \mathbb{E}\left[\left(\overline{f}^a(x;N) - \mathbb{E}[\overline{f}^a(x;N)] \right)^2 \right] dx + \int \left(\mathbb{E}[\overline{f}^a(x;N)] - f(x) \right)^2 dx, \end{aligned} \tag{3.5}$$

which is the sum of the integrated variance and the integrated squared bias. We have the following proposition regarding the MISE and asymptotically optimal bandwidth.

Proposition 3.1 ([16]). *For a p-th order B-spline basis, suppose that the roughness of $f^{(p+1)}$ is finite, i.e. $R(f^{(p+1)}) = \left\|f^{(p+1)}\right\|_2^2 < \infty$, and $f \in C_b^4(\mathbb{R})$[8]. For each of $j = 1, 2, 3$, it is assumed further that $f^{(j)}$ is absolutely continuous, and $f^{(j)} \in L^1(\mathbb{R})$. The following hold:*
(i) The asymptotic MISE satisfies the following bound for small $h > 0$:

$$MISE \leq \theta_p \frac{R(\widetilde{\varphi})}{hN} + \frac{\gamma(f)}{N} + \overline{C}_p \cdot \left\|f^{(p+1)}\right\|_2^2 \cdot h^{2(p+1)}, \tag{3.6}$$

with $\gamma(f) < \infty$, and where $\overline{C}_p$ and $\theta_p \leq 2$ are provided in Table 3. Moreover, as $h \to 0$, $Nh \to \infty$ and $N \to \infty$, then $\overline{f}^a(x; N) \xrightarrow{L^2} f(x)$, from which $\overline{f}^a(x; N)$ is a consistent estimator of $f(x)$.
(ii) The asymptotically optimal bandwidth with respect to (3.6) satisfies

$$h_p^* = \left(\frac{\theta_p}{2(p+1)\overline{C}_p} \cdot \frac{R(\widetilde{\varphi})}{\left\|f^{(p+1)}\right\|_2^2} \cdot \frac{1}{N} \right)^{\frac{1}{2p+3}}, \tag{3.7}$$

where $\overline{C}_p$ and $\theta_p \leq 2$ are provided in Table 3.

TABLE 3
B-spline constants and asymptotically optimal bandwidth for normal data. The column $AMISE_p^*$ provides the AMISE given the optimal bandwidth, h_p^*.

p	$AMISE_p^*$	h_p^*	$\overline{C}_p^{1/2}$	$R(\tilde{\varphi})$	θ_p
0	$N^{-2/3}$	$\sigma\left(\frac{2\sqrt{\pi}}{\overline{C}_0}\right)^{1/3} N^{-1/3}$	0.288675	1	1
1	$N^{-4/5}$	$\sigma\left(\frac{2}{3}\frac{\theta_1\sqrt{\pi}}{\overline{C}_1}R(\tilde{\varphi})\right)^{1/5} N^{-1/5}$	3.72678×10^{-2}	1.73205	4/3
2	$N^{-6/7}$	$\sigma\left(\frac{4}{45}\frac{\theta_2\sqrt{\pi}}{\overline{C}_2}R(\tilde{\varphi})\right)^{1/7} N^{-1/7}$	5.75055×10^{-3}	2.84217	2
3	$N^{-8/9}$	$\sigma\left(\frac{4}{105}\frac{\theta_3\sqrt{\pi}}{\overline{C}_3}R(\tilde{\varphi})\right)^{1/9} N^{-1/9}$	9.09241×10^{-4}	4.96473	2

(iii) The corresponding optimal asymptotic mean integrated squared error is given by

$$AMISE_p^* = (\alpha_p + 1)R(\widetilde{\varphi}) \cdot \left(\frac{\overline{C}_p \left\|f^{(p+1)}\right\|_2^2}{\alpha_p R(\widetilde{\varphi})} \right)^{\frac{1}{2p+3}} \cdot N^{-\frac{2p+2}{2p+3}}, \tag{3.8}$$

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

Remark 6 (Alternative Convergence Measures). Like most existing literature on B-spline density expansion, the previous discussions focus on MISE and

[8] $C_b^n(\mathbb{R})$ is the set of nth order continuously differentiable bounded functions with bounded derivatives

AMISE convergence for a given density f belonging to some space of (smooth) functions. However, it would be interesting to consider the *minimax* convergence discussed in Section 1.2 as it provides an ideal trade-off between approximation error, estimation error, and model complexity relative to the sample size; see for example, Theorem 1 of [124].

We note that the $AMISE_p^*$ in (3.8) depends on f only through $\left\|f^{(p+1)}\right\|_2^2$. In particular, if we consider the family of densities $\mathcal{P}$ that satisfy the assumptions of Proposition 3.1, and further obey $\left\|f^{(p+1)}\right\|_2^2 \leq \kappa$, then

$$\sup_{f\in\mathcal{P}} AMISE_p^* \leq \kappa_p \cdot N^{-\frac{2p+2}{2p+3}},$$

where $\kappa_p := (\alpha_p+1)R(\widetilde{\varphi}) \cdot \left(\frac{\overline{C}_p\kappa}{\alpha_p R(\widetilde{\varphi})}\right)^{\frac{1}{2p+3}}$ depends only on the order of the basis. We leave a detailed investigation of the minimax convergence properties of the density expansion estimators, including smaller sample properties, as interesting an interesting topic for future research.

3.2.1. Rule of thumb and plug-in estimation

Proposition 3.1 provides an asymptotically optimal bandwidth, which becomes operational only once we have an estimate for $\left\|f^{(p+1)}\right\|_2^2$. Of course, this is itself a non-trivial estimation problem, for which several approaches have been proposed as follows:

1. Rule of Thumb: Assume that f is normally distributed, that is $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/(2\sigma^2)}$. We can then compute $\left\|f^{(p+1)}\right\|_2^2$ explicitly, plugging in $\widehat{\sigma}$ for σ:
$$\left\|f^{(p+1)}\right\|_2^2 = (-1)^{3(p+1)}\frac{H_{2(p+1)}(0)}{\sqrt{\pi}\cdot(2\sigma)^{2p+3}},$$
where $H_n(x) := (-1)^n e^{x^2}d^n/dx^n\left(e^{-x^2}\right)$ is the nth order Hermite polynomial. Pluging this into (3.7) yields the rule of thumb bandwidth. The rule of thumb works well for near-normal data, but is otherwise not robust to outliers or multiple modes.
2. Adaptive Heuristic Rule of Thumb: One can also consider the *robust rule of thumb* where the standard estimate $\widehat{\sigma}$ is replaced by $\sigma^* = \min(\widehat{\sigma}, \widehat{R}/\zeta)$. Here $\widehat{R}$ is the inter-quartile range and $\zeta := 2.5 + (1 - e^{-\kappa})$, where κ is the maximum of the excess kurtosis and zero. As demonstrated in [16] (see Section 3.1.2), this choice of σ^* not only protects against outliers but also performs well for multi-modal data. The authors then provide an adaptive heuristic which adjusts the bandwidth according to a binning procedure. Like the standard rule of thumb, it is a poor choice for multi-modal data.
3. Plug-In Estimation: a more careful estimate of $\left\|f^{(p+1)}\right\|_2^2$ can improve the B-spline estimation performance. The so-called *plug-in methods* use an it-

erative approach to estimate $\left\|f^{(p+1)}\right\|_2^2$ (see [60], [106]). A more recent (and highly effective) approach is developed in [6]. Unlike the rules of thumb, this approach performs well even for most multi-modal distributions, and provides a competitive general-purpose alternative to the LCV and LSCV methods of Section 3.1.

Once an estimate for $\left\|f^{(p+1)}\right\|_2^2$ is determined, we plug it into (3.7) to obtain h_p^*.

3.3. Bandwidth selection comparison

In the next sequence of experiments, we analyze the behavior of four bandwidth selection procedures for several density estimation examples. As cautioned in [79], 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. Here we compare the following four methods:

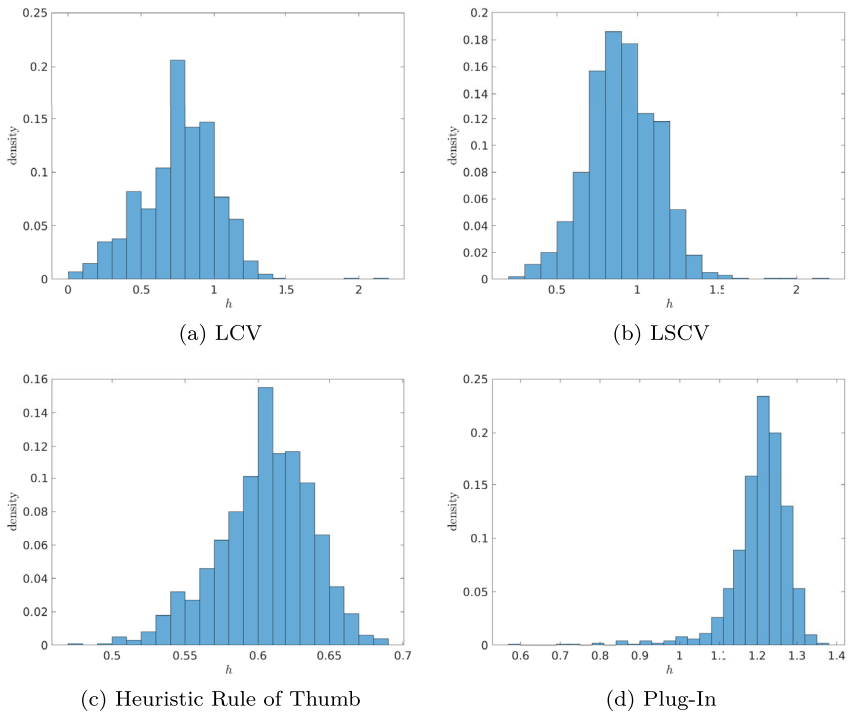
1. Likelihood Cross Validation (LCV) – defined in (3.2)
2. Least-Squares Cross Validation (LSCV) – defined in (3.4)
3. Adaptive Heuristic Rule of Thumb (RoT) – discussed in Section 3.2.1
4. Plug-In Estimator (Plug-In) – discussed in Section 3.2.1, using the method of [6] to estimate $\left\|f^{(p+1)}\right\|_2^2$

For concreteness, we estimate the bandwidth for the linear basis using the Galerkin estimator of Section 2.5. For an empirical comparison of various estimation procedures, see [66]. Experiments comparing the accuracy of each approach are considered later in Figure 7, while the current experiments aim to highlight the disparities among each type of selection approach, in terms of their tendency to over vs under-smooth.

In Figure 4 we present the histograms (out of 1000 realizations) of the optimal bandwidth values provided by the four alternatives described above for the standard normal distribution. First note how tightly distributed the RoT bandwidths are (given that the assumed distribution matches the true distribution), especially compared with the LCV approach, which can be overly sensitive to the observed sample. The Plug-In estimator tends to over-smooth in this case.

We also consider the skewed bimodal and gamma distributions defined in Table 2, with bandwidth histograms displayed in Figure 5 and 6. While gamma presents a relatively stable behaviour, for the skewed bimodal distribution the RoT is clearly inappropriate (given its unimodal assumption). Both the RoT and Plug-In approaches over-smooth in this case, resulting in a higher MISE. By contrast, the LCV method under-smooths (as it is well-known to do, see for example [79]), and LSCV performs ideally well.

Next, we compare the bandwidth selection methods presented above in terms of the MISE convergence. The observed outcomes are presented in Figure 7. Besides the distributions employed in the previous experiment, we consider ad-



(a) LCV (b) LSCV

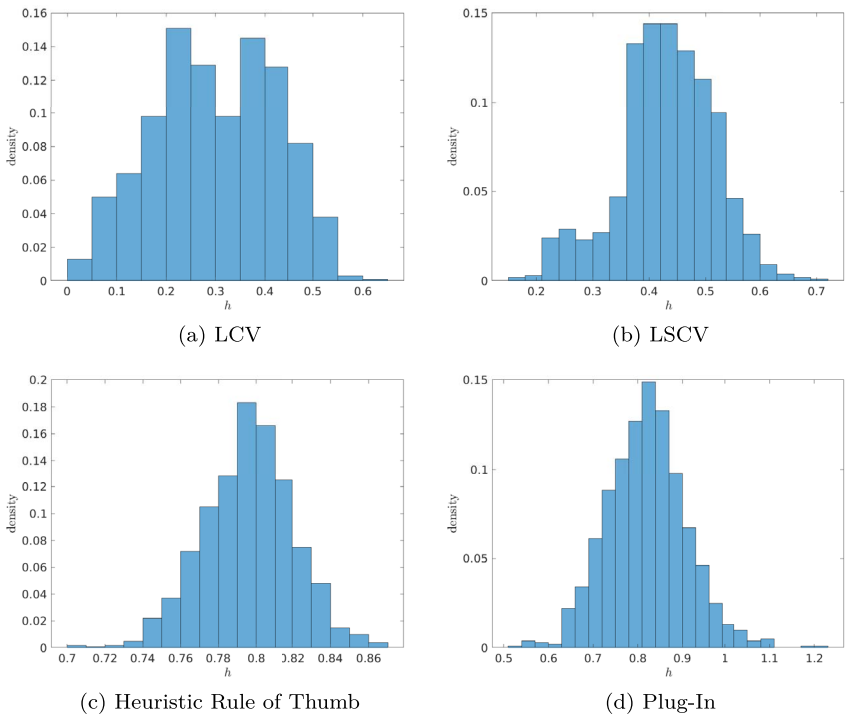
(c) Heuristic Rule of Thumb (d) Plug-In

FIG 4. *Histogram (out of 1000 realizations) of bandwidths for the Normal distribution with* $N = 10^3$.

ditional ones with their own peculiarities, namely the Separated Bimodal, Claw, and Weibull distributions, defined in Table 2. In general, strong performance is provided by LCV, LSCV and Plug-In methodologies, achieving a convergence rate usually superior to $\frac{4}{5}$, i.e. superior to the theoretical convergence predicted by Proposition 3.1. This fact is often observed in practical applications, so when estimating derivatives of the density is not a concern, the linear basis is an excellent choice and is highly tractable.

3.3.1. Recommendation

From these and numerous other experiments (including [66, 16, 17]), we find that the LSCV method has the best all-around performance. It is robust to idiosyncrasies in the data, such as heavy tails or multiple modes, and it is quite efficient as demonstrated in [66]. Compared with LCV, it is less prone to undersmoothing, and has a similar computational cost. The Plug-In methodology works well in larger samples (say $N \geq 5,000$), but is not as robust as the LSCV method for idiosyncratic data, especially at smaller sample sizes ($N < 1,000$). As expected, the heuristic RoT performs poorly in multi-modal cases, like for the Claw or Separated Bimodal distributions. However, when the data is known to



(a) LCV (b) LSCV

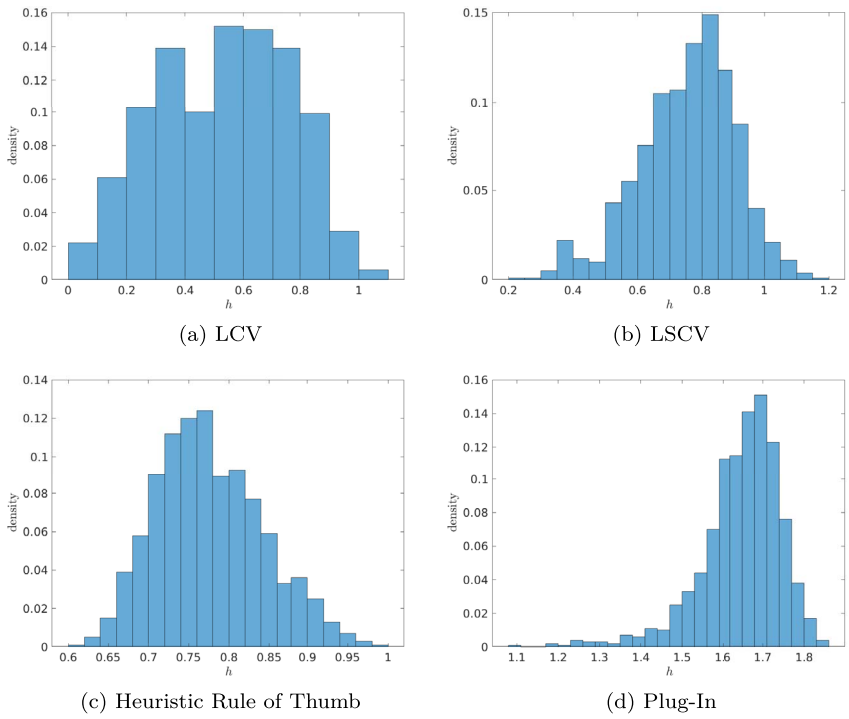
(c) Heuristic Rule of Thumb (d) Plug-In

FIG 5. *Histogram (out of 1000 realizations) of bandwidths for the Skewed Bimodal distribution with* $N = 10^3$.

be unimodal, and reasonably close to normal (including some heavy-tailed distributions), the heuristic RoT performs quite well, as can be seen for the Normal, Gamma, and Weibull examples. When performance is a key consideration, the heuristic RoT has obvious computational advantages over both cross-validation and plug-in estimators, but should be applied judiciously.

4. Applications

There are numerous existing and potential applications of B-spline estimation techniques, and this section illustrates a few interesting examples. For instance, there is a long history of nonparametric statistical techniques in financial econometrics (see [33]), with ever-increasing demand for statistical risk management since the 2008 financial crisis, and the financial fallout of the COVID-19 pandemic. Section 4.1 illustrates the problem of quantile estimation, which is especially important in statistical risk management through measures such Value-at-Risk (VaR) and expected shortfall. We then discuss the application B-spline density estimation to nonparametric simulation in Section 4.2. Other interesting recent applications of B-spline estimators include [110, 91, 104, 100, 47, 46, 26, 54].



(a) LCV (b) LSCV

(c) Heuristic Rule of Thumb (d) Plug-In

FIG 6. *Histogram of bandwidths for the gamma distribution (1000 samples).*

4.1. Quantile estimation

One way to exploit the tractability of the B-splines is to obtain formulas for quantiles and conditional moments directly from the estimated density, rather than via a quantile regression approach. Recall that $F^{-1}(\alpha) := \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$. For each $N \in \mathbb{N}$, define the B-spline quantile function $\overline{F}_a^{-1}(\alpha; N) := \inf\{x \in \mathbb{R} : \overline{F}^a(x; N) \geq \alpha\}$. Under reasonable assumptions, Corollary 4 of [17] shows that $\overline{F}_a^{-1}(\alpha; N) \xrightarrow{a.s.} F^{-1}(\alpha)$ as $N \to \infty$ and $a = a(N) \to \infty$. Closed-form expressions can be derived for various B-splines, such as (4.1) below for the linear basis.

Remark 7. An alternative B-spline approach to quantile estimation is via *quantile regression*. For example, [53] prove that with sufficiently many knots, B-splines can achieve the optimal convergence rate of $N^{-r/(2r+1)}$ for a conditional quantile if the quantile function is smooth up to order r. A related approach is provided more recently in [20].

We now describe an important (and common) practical application in risk management which requires quantile estimation on a large scale. The problem is to estimate the risk posed by a financial position (a portfolio of stocks, options, futures, etc.), for example to a bank or the clearing house which safeguards

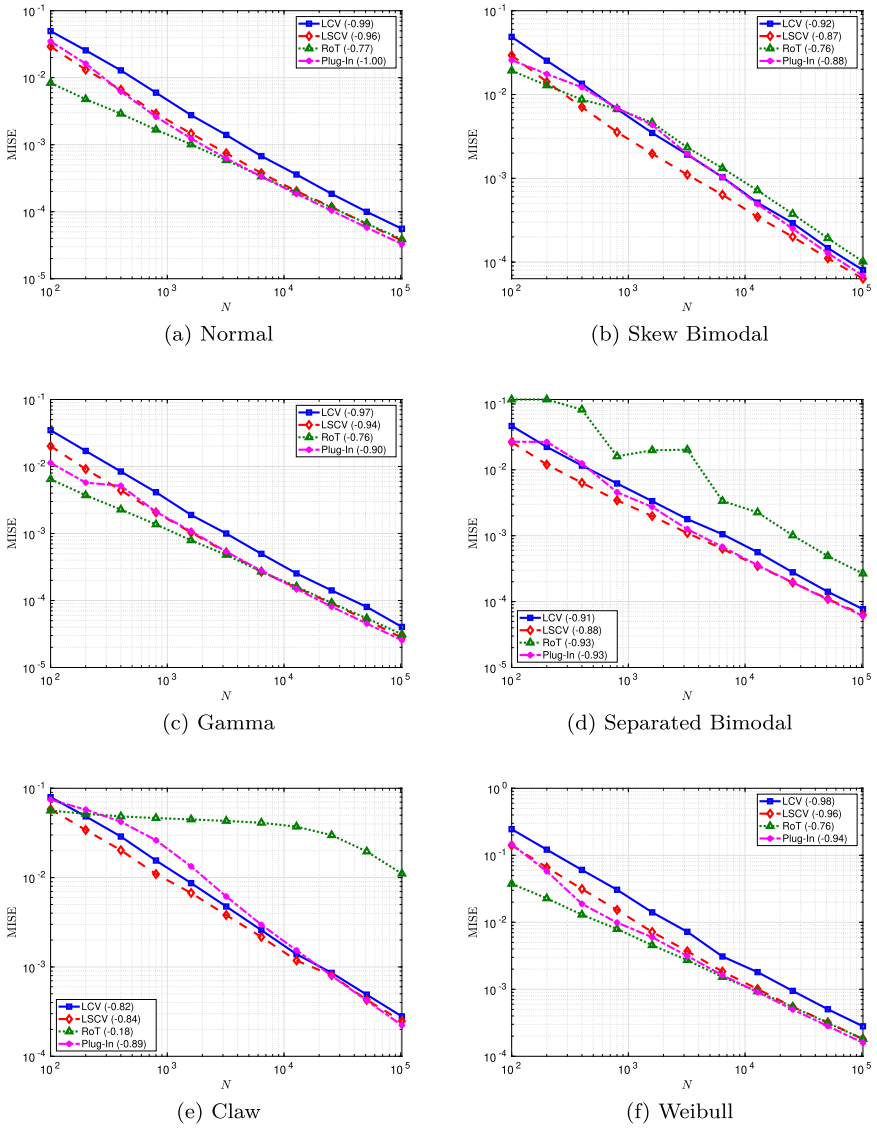


FIG 7. *Comparison of bandwidth selection methods: LCV, LSCV, heuristic Rule of Thumb (denoted by RoT) and Plug-In. MISE(N) (log-scale) computed by means of 1000 trials.*

market participants in default events. Let $\{S_{t_n}\}_{n=0}^{N}$ be a historical sample of position values[9], and $\{\sigma_{t_n}\}_{n=0}^{N}$ a series of estimates for the latent volatility state,

[9]For example, for a long position in a single asset, S_t is just the time series of values for that asset.

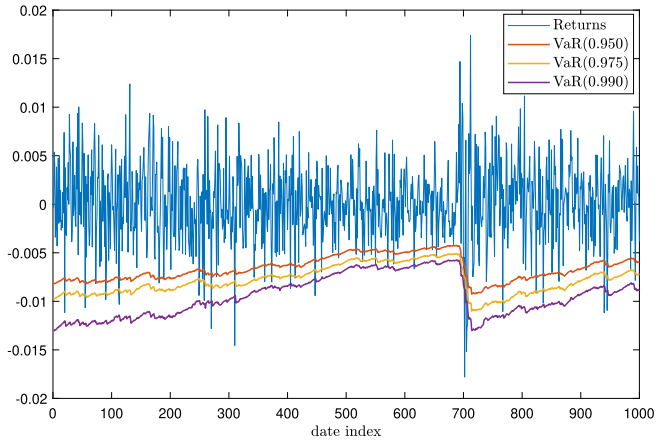


FIG 8. *Sequential quantile estimation (VaR backtest) of a long position in USD/EUR FX futures. FX returns are displayed, along with three VaR(α) estimates using a sliding window* $\{X_n\}_{n=1}^{750}$.

estimated from the returns of the $\{S_{t_n}\}$.[10] The "de-volatilized" return series is given by $Y_{t_n} = \sigma_{t_{n-1}}^{-1} \cdot \log(S_{t_n}/S_{t_{n-1}})$, and σ_{t_N} serves as an estimate for the current volatility level, see [45] for more discussions on this common approach.

Hence, we obtain the conditional sample of "re-volatilized" returns, $X_n := \sigma_{t_N} \cdot Y_{t_n}$, $n = 1, \ldots, N$, which serves as a *reasonably stationary* series of returns and captures the current volatility state. From $\{X_n\}_{n=1}^N$, we can estimate the Value-at-Risk (VaR), defined as $\text{VaR}(\alpha) := \overline{F}_a^{-1}(1-\alpha; N)$, which is interpreted as the maximum losses expected to occur with probability α. VaR is used by risk managers to internally manage the risk of an institution's financial positions, as well as clearing houses to determine the amount of money (margin) that must be set aside to cover losses and mitigate the risk of default. Additional risk-measures are based on conditional tail events, such as

$$\frac{1}{1-\alpha}\int_\alpha^1 \overline{F}_a^{-1}(x; N)dx \xrightarrow{a.s.} \frac{1}{1-\alpha}\int_\alpha^1 F^{-1}(x)dx \quad \text{as } N \to \infty,$$

which captures the "expected-shortfall".

Model risk evaluation, that is the ongoing statistical assessment of model quality, produces a sequence of VaR(α) estimates, known as a backtest, computed from a sliding window of data to assess the statistical integrity of a risk model. Financial institutions calculate millions of these VaR/expected shortfall estimates on a daily basis, using both realized (historical/nonparametrically simulated) and synthetically generated time series data. Figure 8 illustrates the procedure for USD/EUR futures, using data between 5/19/2014 and 5/21/2021. On each of the 1000 dates in the backtest window, we compute the B-spline es-

[10] Traditionally, GARCH and EWMA are the standard models in practice. In this example, we use and EWMA estimate.

timator VaR from (4.1), using the most recent 750 point series of $\{X_n\}$. Three typical levels of α are displayed, and we can see how the VaR estimates evolve and react over time, in response to changing market conditions. Since speed is an important requirement of this type of large scale testing, we utilize the Adaptive Heuristic Rule-of-Thumb bandwidth, which performs very well on the unimodal data that are typical of financial returns time series (see experiments/discussion in Section 3.3.)

4.2. Nonparametric simulation

B-spline estimators are also useful in the context of nonparametric simulation. From a given sample, the obvious approach to re-sampling is to simply bootstrap the observed sample (that is, sample with replacement from the empirical histogram). However, this has the disadvantage of a slow convergence rate as well as an inability to draw points outside of the observed discrete sample. The B-splines offer efficient simulation from a continuous representation of the sample as outlined below, performing a sort of interpolation between observed data points to produce a continuously-drawn sample.

Recall the definition of the linear CDF $\overline{F}^a(x;N)$ from (2.35), where $\overline{\beta}_{a,k}$ are estimated by any of the alternative approaches. We can perform efficient simulation from $\overline{F}^a(x;N)$ by first tabulating $\overline{F}_k := \overline{F}^a(x_k;N)$ at the grid points x_k. Then, for any $y \in (0,1)$, let $k \in \{0,\dots,N_\varphi\}$ be the unique integer satisfying $\overline{F}_k \le y < \overline{F}_{k+1}$, and set $d_k := \overline{\beta}_{a,k+1} - \overline{\beta}_{a,k}$, where $\overline{\beta}_{a,0} = \overline{\beta}_{a,N_\varphi+1} = 0$. The inverse CDF is derived in [17], and satisfies

$$\overline{F}_a^{-1}(y;N) = \begin{cases} x_k + \dfrac{1}{a \cdot d_k}\left(-\overline{\beta}_{a,k} + \sqrt{\overline{\beta}_{a,k}^2 + 2a^{1/2} \cdot d_k (y - \overline{F}_k)}\right), & d_k \neq 0; \\ x_k + \dfrac{y - \overline{F}_k}{a \cdot (\overline{F}_{k+1} - \overline{F}_k)}, & d_k = 0. \end{cases} \tag{4.1}$$

Figure 9 illustrates the procedure, by estimating the nonparametric CDF from a sample of $\mathcal{N}(0, \sigma^2 \Delta t)$ data (left), which corresponds to the log-changes of a geometric Brownian motion, sampled at a weekly frequency. The sample size of $N = 1000$ is effectively reduced to a few dozen basis coefficients (indicated by '+'). After estimation, we simulate $N_{sim} = 10^5$ new data points from the nonparametric density, and obtain a smooth empirical density for the simulated data (right), which closely matches the true density.

The simulated sample reflects the nonparametric CDF, and this can be thought of as a way to increase the sample size by *interpolating* the data, while preserving the empirical distribution.[11] From Figure 9 we can see that the resulting CDF is smooth and closely matches the distribution used to generate the

[11] This is important for sensitivity analysis and calculations of risk measures such as the Value-at-risk (VaR), which can be sensitive to the coarseness of an empirical distribution. The continuous density representation benefits from the improved convergence of the B-spline estimators over the standard histogram, and provides a smoothing function for the empirical distribution.

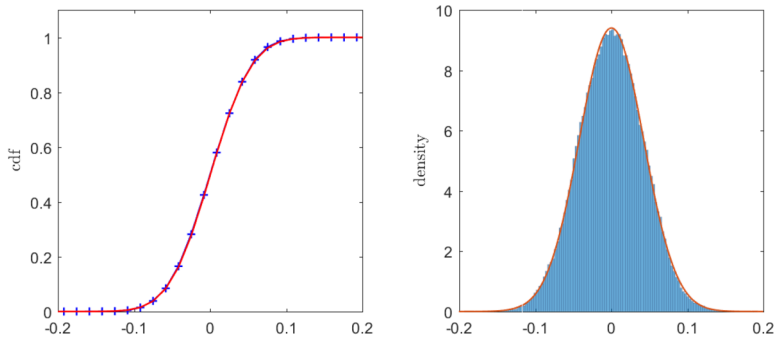


FIG 9. *Nonparametric CDF: estimation and simulation. Left: nonparametric CDF of* $\mathcal{N}(0, \sigma^2\Delta t)$*,* $\sigma = 0.3, \Delta t = 1/52$*; estimated from sample of size* $N = 1000$*, with true CDF overlay. Right: sample of size* $N_{sim} = 10^5$ *simulated from nonparametric CDF.*

sample. This methodology can also be applied to simulate parametric stochastic models from known characteristic functions, as studied for example in [38, 3].

Remark 8. An alternative approach to simulation is to recognize that the density $\overline{f}^a(x; N) = \sum_{k=1}^{N_\varphi} a^{-1/2}\overline{\beta}_{a,k}(N) \cdot a^{1/2}\varphi_{a,k}(x)$ is a mixture distribution. Each $a^{1/2}\varphi_{a,k}(x)$ is triangularly distributed, and can be simulated as the sum of two uniforms. By sampling from the discrete distribution with weights $a^{-1/2}\overline{\beta}_{a,k}(N)$, we determine in which bin the triangular variate falls. This idea generalizes to higher order B-splines, and was first proposed in [92].

5. Conclusion

The work provides an in-depth account of the current state of nonparametric density estimation by local basis methods, with particular attention paid to B-spline basis estimation. Most often, B-splines are applied in the transformed log-density space, as is the case for logsplines, smoothing splines, and P-Splines. However, direct estimation by B-spline basis expansion of the density is a powerful technique for which recent progress have been made. This work surveys the current state-of-the art and recent developments in B-spline density estimation, with a particular focus on the direct expansion of the density. We detail the progress of existing literature, and offer new insights, experiments, and analyses to cast the various estimation concepts in a unified context. Parallels and contrasts are drawn to the more familiar context of kernel density estimation. Applications to quantile estimation with time series data, as well as nonparametric simulation are provided to illustrate the versatility of B-spline density estimation.

5.1. Future research directions

Before concluding this work, we highlight just a few interesting topics which merit further study:

1. Analysis of higher order splines: in theory, the MISE of a p-th order B-spline projection behaves like $N^{-\frac{2p+2}{2p+3}}$ for the optimally selected bandwidth, with higher order splines offering accuracy that approaches the optimal rate for sufficiently smooth densities (recall Section 2.2). While the Haar and Linear basis are well-studied, higher order B-splines have not been extensively analyzed in the literature. For applications which require derivative estimation, higher order B-splines possess the necessary differentiability (see for example [92]).
2. Extension to multivariate density estimation: multivariate density estimation is another important problem in nonparametric statistics, see for example [103]. While some recent progress has been made for B-splines, such as [120, 90, 128] for bivariate tensor product B-splines, and [21] for an adaptive tensor product approach using B-splines as nonparametric Bayesian priors, the literature is still quite limited, and the approach via multivariate duality has not yet been considered.
3. On-line coefficient estimation: recent "big-data" applications require the estimation of coefficients on-the-fly, in a computationally and memory efficient manner, see for example [36]. Section 2.6 proposes an adaptive coefficient estimator that is well-suited for streaming data applications. Given the compact (and narrow) support of this "alternative dual" estimator, it is also promising for multivariate extensions. We leave these extensions for future research.
4. Theory of spectral filtering: at this point, the use of spectral filtering in combination with B-spline projection remains an empirical tool (recall Section 2.4), and the filter order is treated as hyper-parameter to be tuned to a particular problem area. Analysis of the theoretical properties of filtered projection, such as the choice of optimal filter order as a function of the sample size (and properties of the data), remains wanting.
5. Nonparametric regression: a natural extension is to the related strain of research on nonparametric regression, where it remains to be seen if the concept of B-spline duality is equally powerful, see for example [78, 57].
6. Additional theory: the theory of B-spline density estimation is far from complete at this stage. As discussed in Remark 6, most of the literature focuses on pointwise converge results based on MISE, and measures such as minimax convergence, which are well developed in the context of logslines [124], are not yet established for B-spline density expansions.

Compared with kernel density estimation, the literature on local basis estimation is not nearly as mature, and there remain many possible avenues for future research, including those outlined above.

References

[1] Antoniadis, A. (2007). Wavelet methods in statistics: some recent developments and their applications. *Statistics Surveys* **1** 16–55. MR2520413

[2] Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE transactions on information theory* **44** 2743–2760. MR1658898

[3] Bernard, C., Cui, Z. and Mcleish, D. (2012). Nearly exact option price simulation using characteristic functions. *International Journal of Theoretical and Applied Finance* **15** 1–29. MR2999572

[4] Blu, T. and Unser, M. (2004). Quantitative L/sup 2/approximation error of a probability density estimate given by its samples. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* **3** iii–952. IEEE.

[5] Boneva, L. I., Kendall, D. and Stefanov, I. (1971). Spline transformations: Three new diagnostic aids for the statistical data-analyst. *Journal of the Royal Statistical Society. Series B (Methodological)* **33** 1–71. MR0288888

[6] Botev, Z. I., Grotowski, J. F. and Kroese, D. P. (2010). Kernel density estimation via diffusion. *Annals of Statistics* **38** 2916–2957. MR2722460

[7] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360. MR0767163

[8] Bowman, A. W., Hall, P. and Titterington, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* **71** 341–351. MR0767162

[9] Brunk, H. D. (1978). Univariate density estimation by orthogonal series. *Biometrika* **65(3)** 521–528. MR0521820

[10] Caudle, K. A. and Wegman, E. (2009). Nonparametric density estimation of streaming data using orthogonal series. *Computational Statistics and Data Analysis* **53** 3980–3986. MR2744299

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

[12] Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica* 129–145. MR1379053

[13] Christensen, O. (2003). *An Introduction to Frames and Riesz Bases*. Birkhauser Boston. MR1946982

[14] Ciarlet, P. G. (2002). *The finite element method for elliptic problems* **40**. Siam. MR1930132

[15] Csorgo, S. and Totik, V. (1983). On how long interval is the empirical characteristic function uniformly consistent. *Acta Scientiarum Mathematicarum* **45** 141–149. MR0708779

[16] Cui, Z., Kirkby, J. L. and Nguyen, D. (2020). Nonparametric density estimation by B-spline duality. *Econometric Theory* 1–42. MR4078115

[17] Cui, Z., Kirkby, J. L. and Nguyen, D. (2021). A data-driven framework for consistent financial valuation and risk measurement. *European Journal of Operational Research* **289(1)** 381–398. MR4159671

[18] Cui, Z., Kirkby, J. L. and Nguyen, D. (2021). Efficient Simulation of Generalized SABR and Stochastic Local Volatility Models based on Markov chain approximations. *European Journal of Operational Research*

290(3) 1046–1062. MR4195068

[19] Dai, X., Muller, H.-G. and Yao, F. (2017). Optimal Bayes Classifiers for functional data and density ratios. *Biometrika* **104** 545–560. MR3694582

[20] Das, P. and Ghosal, S. (2017). Bayesian quantile regression using random B-spline series prior. *Computational Statistics & Data Analysis* **109** 121–143. MR3603645

[21] De Jonge, R., Van Zanten, J. et al. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electronic Journal of Statistics* **6** 1984–2001. MR3020254

[22] Dias, R. (1998). Density estimation via hybrid splines. *Journal of Statistical Computation and Simulation* **60** 277–293. MR1704852

[23] Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Annals of Statistics* **24(2)** 508–539. MR1394974

[24] Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The annals of Statistics* **26** 879–921. MR1635414

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

[26] Edwards, M. C., Meyer, R. and Christensen, N. (2019). Bayesian nonparametric spectral density estimation using B-spline priors. *Statistics and Computing* **29** 67–78. MR3905541

[27] Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11(2)** 89–121. MR1435485

[28] Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science* **11** 89–121. MR1435485

[29] Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* **2** 637–653.

[30] Eilers, P. H. and Marx, B. D. (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge University Press.

[31] Eilers, P. H., Marx, B. D. and Durbán, M. (2015). Twenty years of P-splines. *SORT: statistics and operations research transactions* **39** 0149–186. MR3467488

[32] Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* **14** 153–158.

[33] Fan, J. (2005). A selective overview of nonparametric methods in financial econometrics. *Statistical Science* 317–337. MR2210224

[34] Feuerverger, A. and Mureika, R. (1977). The empirical characteristic function and its applications. *Annals of Statistics* **5** 88–97. MR0428584

[35] Fix, E. and Hodges, J. L. (1951). Nonparametric discrimination: consistency properties. Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, February.

[36] Garcia-Trevino, E. S. and Barria, J. A. (2012). Online wavelet-based density estimation for non-stationary streaming data. *Computa-*

tional Statistics and Data Analysis **56** 327–344. MR2842342

[37] Gehringer, K. R. and Redner, R. A. (1992). Nonparametric probability density estimation using normalized b–splines. *Communications in Statistics-Simulation and Computation* **21** 849–878. MR1185175

[38] Glasserman, P. and Liu, Z. (2010). Sensitivity estimates from characteristic functions. *Operations Research* **58** 1611–1623. MR2752708

[39] Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *The Annals of Statistics* **39** 1608–1632. MR2850214

[40] Goldenshluger, A. and Lepski, O. (2014). On adaptive minimax density estimation on R^d. *Probability Theory and Related Fields* **159** 479–543. MR3230001

[41] Gu, C. (1993). Smoothing spline density estimation: a dimensionless automatic algorithm. *Journal of the American Statistical Association* **88(422)** 495–504. MR1224374

[42] Gu, C. (1995). Smoothing spline density estimation: conditional distribution. *Statistica Sinica* 709–726. MR1347615

[43] Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: theory. *Annals of Statistics* **21(1)** 217–234. MR1212174

[44] Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics* 217–234. MR1212174

[45] Gurrola-Perez, P. and Murphy, D. (2015). Filtered historical simulation Value-at-Risk models and their competitors.

[46] Hadrich, A., Zribi, M. and Masmoudi, A. (2012). A proposed normalized B-spline density estimator and it application in unsupervised statistical image segmentation. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 480–483. IEEE.

[47] Hadrich, A., Zribi, M. and Masmoudi, A. (2015). Using B-splines functions and EM algorithm for Hidden Markov Model-based Unsupervised Image Segmentation. In *Applied Mathematics in Tunisia* 203–214. Springer. MR3440555

[48] Hall, P. (1981). On trigonometric series estimates of densities. *Annals of Statistics* **9** 683–685. MR0615446

[49] Hall, P. (1982). Cross-validation in density estimation. *Biometrika* **69** 383–390. MR0671976

[50] Hall, P. (1987). Cross-validation and the smoothing of orthogonal series density estimators. *Journal of Multivariate Analysis* **21** 189–206. MR0884096

[51] Hall, P. (1987). On Kullback-Leibler loss and density estimation. *The Annals of Statistics* 1491–1519. MR0913570

[52] Hardle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics* **13** 1465–1481. MR0811503

[53] He, X. and Shi, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journaltitle of Nonparametric Statistics* **3** 299–308. MR1291551

[54] He, Y., Fan, H., Lei, X. and Wan, J. (2021). A runoff probability density prediction method based on B-spline quantile regression and kernel density estimation. *Applied Mathematical Modelling* **93** 852–867. MR4202521

[55] Heil, C. (2011). *A Basis Theory Primer, expanded edition*. Birkhauser. MR2744776

[56] Huang, S.-Y. (1999). Density estimation by wavelet-based reproducing kernels. *Statistica Sinica* **9** 137–151. MR1678885

[57] Imoto, S. and Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics* **55** 671–687. MR2028612

[58] Izenman, A. J. (1991). Recent Developments in Nonparametric Density Estimation. *Journal of the American Statistical Association* **86(413)** 205–223. MR1137112

[59] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91** 401–407. MR1394097

[60] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* **11** 337–381. MR1415761

[61] Juditsky, A. and Lambert-Lacroix, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10** 187–220. MR2046772

[62] Kauermann, G., Schellhase, C. and Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical B-splines. *Scandinavian Journal of Statistics* **40** 685–705. MR3145112

[63] Kirkby, J. L. (2015). Efficient Option Pricing by Frame Duality with the Fast Fourier Transform. *SIAM Journal on Financial Mathematics* **6(1)** 713–747. MR3384832

[64] Kirkby, J. L. (2017). Robust Option Pricing with Characteristic Functions and the B-Spline Order of density projection. *Journal of Computational Finance* **21(2)** 101–127.

[65] Kirkby, J. L. and Deng, S. J. (2019). Static Hedging and Pricing of Exotic Options with Payoff frames. *Mathematical Finance* **29(2)** 612–658. MR3925432

[66] Kirkby, J. L., Leitao, Á. and Nguyen, D. (2021). Nonparametric density estimation and bandwidth selection with B-spline bases: A novel Galerkin method. *Computational Statistics & Data Analysis* **159** 107202. MR4233347

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

[68] Koo, J.-Y. and Kim, W.-C. (1996). Wavelet density estimation by approximation of log-densities. *Statistics & probability letters* **26** 271–278. MR1394903

[69] Koo, J.-Y., Kooperberg, C. and Park, J. (1999). Logspline density estimation under censoring and truncation. *Scandinavian journal of statistics* **26** 87–105. MR1685304

[70] Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis* **12** 327–347. MR1144152

[71] Kooperberg, C. and Stone, C. J. (1992). Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1** 301–328. MR0870445

[72] Kooperberg, C. and Stone, C. J. (2004). Comparison of Parametric and bootstrap approaches to obtaining confidence intervals for Logspline density estimation. *Journal of Computational and Graphical Statistics* **1** 106–122. MR2044873

[73] Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis* **12** 327–347. MR1144152

[74] Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of computational and graphical statistics* **13** 183–212. MR2044877

[75] Leitao, A., Oosterlee, C. W., Ortiz-Gracia, L. and Bohte, S. M. (2018). On the data-driven COS method. *Applied Mathematics and Computation* **317** 68–84. MR3709219

[76] Leitao, A. and Ortiz-Gracia, L. (2020). Model-free computation of risk contributions in credit portfolios. *Applied Mathematics and Computation* **382** 125351. MR4100766

[77] Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)* **40** 113–132. MR0517434

[78] Li, G., Shi, P. and Li, G. (1995). Global convergence rates of B-spline M-estimators in nonparametric regression. *Statistica Sinica* 303–318. MR1329300

[79] Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Annals of Statistics* **27(2)** 415–438. MR1714723

[80] Masri, R. and Redner, R. A. (2005). Convergence rates for uniform B-spline density estimators on bounded and semi-infinite domains. *Nonparametric Statistics* **17(5)** 555–582. MR2141362

[81] Matthies, H. G. and Keese, A. (2005). Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Computer methods in applied mechanics and engineering* **194** 1295–1331. MR2121216

[82] McDonald, S. and Campbell, D. (2021). A review of uncertainty quantification for density estimation. *Statistics Surveys* **15** 1–71. MR4255286

[83] Papp, D. and Alizadeh, F. (2014). Shape-Constrained Estimation Using Nonnegative Splines. *Journal of Computational and Graphical Statistics* **23** 211–231. MR3173768

[84] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics,* **33** 1065–1076. MR0143282

[85] Penev, S. and Dechevsky, L. (1997). On non-negative wavelet-based density estimators. *Journal of Nonparametric Statistics* **7** 365–394. MR1460206

[86] Peter, A. M. and Rangarajan, A. (2008). Maximum likelihood

wavelet density estimation with applications to image and shape matching. *IEEE Transactions on Image Processing* **17(4)** 458–468. MR2512451

[87] Pla, P. D. A. and Unser, M. (2022). Bona Fide Riesz Projections for Density Estimation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5613–5616. IEEE.

[88] Politis, D. N. (2001). On nonparametric function estimation with infinite-order flat-top kernels. *Probability and Statistical Models with applications* 469–483.

[89] Politis, D. N. and Romano, J. P. (1999). Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis* **68** 1–25. MR1668848

[90] Price, M. J., Yu, C. L., Hennessy, D. A. and Du, X. (2019). Are actuarial crop insurance rates fair?: an analysis using a penalized bivariate B-spline method. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **68** 1207–1232. MR4022810

[91] Qingguo, T. and Longsheng, C. (2010). B-spline estimation for spatial data. *Journal of Nonparametric Statistics* **22** 197–217. MR2598962

[92] Redner, R. A. (1999). Convergence rates for uniform B-spline density estimators part I: one dimension. *SIAM Journal on Scientific Computing* **20(6)** 1929–1953. MR1694647

[93] Redner, R. A. (2000). Convergence rates for uniform B-spline density estimators part II: Multiple dimensions. *Journal of nonparametric statistics* **12** 753–777. MR1802575

[94] Redner, R. A. and Gehringer, K. (1994). Function estimation using partitions of unity. *Communications in Statistics-Theory and Methods* **23** 2059–2078. MR1281903

[95] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27** 832–837. MR0079873

[96] Rosenblatt, M. (1971). Curve estimates. *The Annals of Mathematical Statistics* **42** 1815–1842. MR0301851

[97] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9** 65–78. MR0668683

[98] Ruijter, M. J., Versteegh, M. and Oosterlee, C. W. (2015). On the application of spectral filters in a Fourier option pricing technique. *Journal of Computational Finance* **19(1)** 76–106.

[99] Ruppert, D., Wand, M. and Carroll, R. (2003). Semiparametric Regression.,(Cambridge University Press: Cambridge, UK.). MR1998720

[100] Savaci, F. A. and Güngör, M. (2012). Estimating probability density functions and entropies of chua's circuit using b-spline functions. *International Journal of Bifurcation and Chaos* **22** 1250107.

[101] Schwartz, S. C. (1967). Estimation of a probability density by an orthogonal series. *Annals of Mathematical Statistics,* **38** 1261–1265. MR0221638

[102] Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-

validation in density estimation. *Journal of the American Statistical Association* **82(400)** 1131–1146. MR0922178

[103] Scott, D. W. and Sain, S. R. (2005). Multidimensional density estimation. *Handbook of statistics* **24** 229–261.

[104] Sharef, E., Strawderman, R. L., Ruppert, D., Cowen, M. and Halasyamani, L. (2010). Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electronic journal of statistics* **4** 606–642. MR2660535

[105] Sheather, S. J. (2004). Density Estimation. *Statistical Science* **19** 588–597. MR2185580

[106] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 683–690. MR1125725

[107] Staudenmayer, J., Ruppert, D. and Buonaccorsi, J. P. (2008). Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association* **103** 726–736. MR2524005

[108] Terrell, G. R. and Scott, D. W. (1992). Variable Kernel Density Estimation. *Annals of Statistics* **20(3)** 1236–1265. MR1186249

[109] Unser, M. and Daubechies, I. (July 1997). On the approximation power of convolution-based least squares versus interpolation. *IEEE Transactions on Signal Processing* **45(7)** 1697–1711.

[110] Venelli, A. (2010). Efficient entropy estimation for mutual information analysis using B-splines. In *IFIP International Workshop on Information Security Theory and Practices* 17–30. Springer.

[111] Wahba, G. (1975). Interpolating spline methods for density estimation I. Equi-spaced knots. *The Annals of Statistics* 30–48. MR0370906

[112] Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Annals of Statistics* **9** 146–156. MR0600541

[113] Walter, G. and Blum, J. (1979). Probability density estimation using delta sequences. *Annals of Statistics* **7(2)** 328–340. MR0520243

[114] Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. CRC press. MR1319818

[115] Wand, M. P. and Schucany, W. R. (1990). Gaussian-based kernels. *Canadian Journal of Statistics* **18** 197–204. MR1079592

[116] Wang, W. and Zhang, Z. (2019). Computing the Gerber–Shiu function by frame duality projection. *Scandinavian Actuarial Journal* **4** 291–307. MR3929238

[117] Watson, G. S. (1969). Density estimation by orthogonal series. *The Annals of Mathematical Statistics* **38** 1262–1265. MR0242332

[118] Wegman, E. J. (1972). Nonparametric Probability Density Estimation: A summary of available methods. *Technometrics* **14(3)** 533–546.

[119] Wegman, E. J. and Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association* **78** 351–365. MR0711110

[120] Wood, S. N. (2017). P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. *Statistics and Computing* **27** 985–989. MR3627558

[121] Woodroofe, M. (1970). On choosing a delta sequence. *Annals of Mathematical Statistics,* **41** 1665–1671. MR0270515

[122] Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Transactions on Information Theory* **45** 2271–2284. MR1725115

[123] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 1564–1599. MR1742500

[124] Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory* **44** 95–116. MR1486651

[125] Young, R. (1980). *An Introduction to Nonharmonic Fourier Series*, (revised) ed. Academic Press, New York. MR0591684

[126] Yu, J. (2004). Empirical characteristic function estimation and its applications. *Econometric reviews* **23** 93–123. MR2075094

[127] Zhang, Z., Yong, Y. and Yu, W. (2020). Valuing equity-linked death benefits in general exponential Lévy models. *Journal of Computational and Applied Mathematics* **365** 112377. MR3990773

[128] Zhao, J. and Li, S. (2020). Efficient pricing of European options on two underlying assets by frame duality. *Journal of Mathematical Analysis and Applications* **486** 123873. MR4053054

[129] Zhao, Y., Zhang, M., Ni, Q. and Wang, X. (2023). Adaptive Nonparametric Density Estimation with B-Spline Bases. *Mathematics* **11**.