The Annals of Statistics
2010, Vol. 38, No. 5, 2916–2957
DOI: 10.1214/10-AOS799

KERNEL DENSITY ESTIMATION VIA DIFFUSION

BY Z. I. BOTEV[1], J. F. GROTOWSKI AND D. P. KROESE[1]

University of Queensland

We present a new adaptive kernel density estimator based on linear diffusion processes. The proposed estimator builds on existing ideas for adaptive smoothing by incorporating information from a pilot density estimate. In addition, we propose a new plug-in bandwidth selection method that is free from the arbitrary normal reference rules used by existing methods. We present simulation examples in which the proposed approach outperforms existing methods in terms of accuracy and reliability.

1. Introduction. Nonparametric density estimation is an important tool in the statistical analysis of data. A nonparametric estimate can be used, for example, to assess the multimodality, skewness, or any other structure in the distribution of the data [47, 49]. It can also be used for the summarization of Bayesian posteriors, classification and discriminant analysis [50]. Nonparametric density estimation has even proved useful in Monte Carlo computational methods, such as the smoothed bootstrap method and the particle filter method [11]. Nonparametric density estimation is an alternative to the parametric approach, in which one specifies a model up to a small number of parameters and then estimates the parameters via the likelihood principle. The advantage of the nonparametric approach is that it offers a far greater flexibility in modeling a given dataset and, unlike the classical approach, is not affected by specification bias [37]. Currently, the most popular nonparametric approach to density estimation is *kernel density estimation* (see [47, 50, 53]).

Despite the vast body of literature on the subject, there are still many contentious issues regarding the implementation and practical performance of kernel density estimators. First, the most popular data-driven bandwidth selection technique, the *plug-in* method [26, 48], is adversely affected by the so-called *normal reference rule* [10, 25], which is essentially a construction of a preliminary normal model of the data upon which the performance of the bandwidth selection method depends. Although plug-in estimators perform well when the normality assumption holds approximately, at a conceptual level the use of the normal reference rule invalidates the original motivation for applying a nonparametric method in the first place.

Received December 2009.

[1]Supported by Australian Research Council Grant DP0985177.

AMS 2000 subject classifications. Primary 62G07, 62G20; secondary 35K05, 35K15, 60J60, 60J70.

Key words and phrases. Nonparametric density estimation, heat kernel, bandwidth selection, Langevin process, diffusion equation, boundary bias, normal reference rules, data sharpening, variable bandwidth.

Second, the popular Gaussian kernel density estimator [42] lacks *local adaptivity*, and this often results in a large sensitivity to outliers, the presence of spurious bumps, and in an overall unsatisfactory bias performance—a tendency to flatten the peaks and valleys of the density [51].

Third, most kernel estimators suffer from *boundary bias* when, for example, the data is nonnegative—a phenomenon due to the fact that most kernels do not take into account specific knowledge about the domain of the data [41, 44].

These problems have been alleviated to a certain degree by the introduction of more sophisticated kernels than the simple Gaussian kernel. Higher-order kernels have been used as a way to improve local adaptivity and reduce bias [28], but these have the disadvantages of not giving proper nonnegative density estimates, and of requiring a large sample size for good performance [42]. The lack of local adaptivity has been addressed by the introduction of *adaptive* kernel estimators [1, 15, 16, 27]. These include the *balloon* estimators, *nearest neighbor* estimators and *variable bandwidth* kernel estimators [39, 51], none of which yield bona fide densities, and thus remain somewhat unsatisfactory. Other proposals such as the *sample point adaptive* estimators are computationally burdensome (the fast Fourier transform cannot be applied [49]), and in some cases do not integrate to unity [44]. The *boundary kernel estimators* [24], which are specifically designed to deal with boundary bias, are either not adaptive away from the boundaries or do not result in bona fide densities [22]. Thus, the literature abounds with partial solutions that obscure a unified comprehensive framework for the resolution of these problems.

The aim of this paper is to introduce an adaptive kernel density estimation method based on the smoothing properties of linear diffusion processes. The key idea is to view the kernel from which the estimator is constructed as the transition density of a diffusion process. We utilize the most general linear diffusion process that has a given limiting and stationary probability density. This stationary density is selected to be either a pilot density estimate or a density that the statistician believes represents the information about the data prior to observing the available empirical data. The approach leads to a simple and intuitive kernel estimator with substantially reduced asymptotic bias and mean square error. The proposed estimator deals well with boundary bias and, unlike other proposals, is always a bona fide probability density function. We show that the proposed approach brings under a single framework some well-known bias reduction methods, such as the Abramson estimator [1] and other variable location or scale estimators [7, 18, 27, 46].

In addition, the paper introduces an improved plug-in bandwidth selection method that completely avoids the normal reference rules [25] that have adversely affected the performance of plug-in methods. The new plug-in method is thus genuinely "nonparametric," since it does not require a preliminary normal model for the data. Moreover, our plug-in approach does not involve numerical optimization and is not much slower than computing a normal reference rule [4].

The rest of the paper is organized as follows. First, we describe the Gaussian kernel density estimator and explain how it can be viewed as a special case of smoothing using a diffusion process. The Gaussian kernel density estimator is then used to motivate the most general linear diffusion that will have a set of essential smoothing properties. We analyze the asymptotic properties of the resulting estimator and explain how to compute the asymptotically optimal plug-in bandwidth. Finally, the practical benefits of the model are demonstrated through simulation examples on some well-known datasets [42]. Our findings demonstrate an improved bias performance and low computational cost, and a boundary bias improvement.

2. Background. Given N independent realizations $\mathcal{X}_N \equiv \{X_1, \ldots, X_N\}$ from an unknown continuous probability density function (p.d.f.) f on $\mathscr{X}$, the *Gaussian kernel density estimator* is defined as

$$\hat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N} \phi(x, X_i; t), \qquad x \in \mathbb{R}, \tag{1}$$

where

$$\phi(x, X_i; t) = \frac{1}{\sqrt{2\pi t}} e^{-(x-X_i)^2/(2t)}$$

is a Gaussian p.d.f. (kernel) with location X_i and scale $\sqrt{t}$. The scale is usually referred to as the *bandwidth*. Much research has been focused on the optimal choice of t in (1), because the performance of $\hat{f}$ as an estimator of f depends crucially on its value [26, 48]. A well-studied criterion used to determine an optimal t is the *Mean Integrated Squared Error* (MISE),

$$\mathrm{MISE}\{\hat{f}\}(t) = \mathbb{E}_f \int [\hat{f}(x;t) - f(x)]^2 \, dx,$$

which is conveniently decomposed into integrated squared bias and integrated variance components:

$$\mathrm{MISE}\{\hat{f}\}(t) = \int \underbrace{\big(\mathbb{E}_f[\hat{f}(x;t)] - f(x)\big)}_{\text{pointwise bias of } f}^2 dx + \int \underbrace{\mathrm{Var}_f[\hat{f}(x;t)]}_{\text{pointwise variance of } f} dx.$$

Note that the expectation and variance operators apply to the random sample $\mathcal{X}_N$. The MISE depends on the bandwidth $\sqrt{t}$ and f in a quite complicated way. The analysis is simplified when one considers the asymptotic approximation to the MISE, denoted AMISE, under the consistency requirements that $t = t_N$ depends on the sample size N such that $t_N \downarrow 0$ and $N\sqrt{t_N} \to \infty$ as $N \to \infty$, and f is twice continuously differentiable [48]. The asymptotically optimal bandwidth is then the minimizer of the AMISE. The asymptotic properties of (1) under these assumptions are summarized in Appendix A.

A key observation about the Gaussian kernel density estimator (1) is that it is the unique solution to the diffusion partial differential equation (PDE)

$$\frac{\partial}{\partial t}\hat{f}(x;t) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\hat{f}(x;t), \qquad x \in \mathscr{X}, t > 0, \tag{2}$$

with $\mathscr{X} \equiv \mathbb{R}$ and initial condition $\hat{f}(x;0) = \Delta(x)$, where $\Delta(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - X_i)$ is the empirical density of the data $\mathcal{X}_N$ [here $\delta(x - X_i)$ is the Dirac measure at X_i]. Equation (2) is the well-known Fourier heat equation [36]. This link between the Gaussian kernel density estimator and the Fourier heat equation has been noted in Chaudhuri and Marron [6]. We will, however, go much further in exploiting this link. In the heat equation interpretation, the Gaussian kernel in (1) is the so-called Green's function [36] for the diffusion PDE (2). Thus, the Gaussian kernel density estimator $\hat{f}(x;t)$ can be obtained by evolving the solution of the parabolic PDE (2) up to time t.

To illustrate the advantage of the PDE formulation over the more traditional formulation (1), consider the case where the domain of the data is *known to be* $\mathscr{X} \equiv [0, 1]$. It is difficult to see how (1) can be easily modified to account for the finite support of the unknown density. Yet, within the PDE framework, all we have to do is solve the diffusion equation (2) over the finite domain $[0, 1]$ with initial condition $\Delta(x)$ and the Neumann boundary condition

$$\frac{\partial}{\partial x}\hat{f}(x;t)\Big|_{x=1} = \frac{\partial}{\partial x}\hat{f}(x;t)\Big|_{x=0} = 0.$$

The boundary condition ensures that $\frac{d}{dt}\int_{\mathscr{X}}\hat{f}(x;t)\,dx = 0$, from where it follows that $\int_{\mathscr{X}}\hat{f}(x;t)\,dx = \int_{\mathscr{X}}\hat{f}(x;0)\,dx = 1$ for all $t \geq 0$. The analytical solution of this PDE in this case is [3]

$$\hat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N}\kappa(x, X_i; t), \qquad x \in [0, 1], \tag{3}$$

where the kernel κ is given by

$$\kappa(x, X_i; t) = \sum_{k=-\infty}^{\infty}\phi(x, 2k + X_i; t) + \phi(x, 2k - X_i; t), \qquad x \in [0, 1]. \tag{4}$$

Thus, the kernel accounts for the boundaries in a manner similar to the boundary correction of the *reflection method* [49]. We now compare the properties of the kernel (4) with the properties of the Gaussian kernel ϕ in (1).

First, the series representation (4) is useful for deriving the small bandwidth properties of the estimator in (3). The asymptotic behavior of $\kappa(x, X_i; t)$ as $t \to 0$ in the interior of the domain $[0, 1]$ is no different from that of the Gaussian kernel, namely,

$$\sum_{k=-\infty}^{\infty}\phi(x, 2k + X_i; t) + \phi(x, 2k - X_i; t) \sim \phi(x, X_i; t), \qquad t \downarrow 0,$$

for any fixed x in the interior of the domain $[0, 1]$. Here $q(t) \sim z(t), t \downarrow t_0$ stands for $\lim_{t\downarrow t_0} \frac{q(t)}{z(t)} = 1$. Thus, for small t, the estimator (3) behaves like the Gaussian kernel density estimator (1) in the interior of $[0, 1]$. Near the boundaries at $x = 0, 1$, however, the estimator (3) is consistent, while the Gaussian kernel density estimator is inconsistent. In particular, a general result in Appendix D includes as a special case the following boundary property of the estimator (3):

$$\mathbb{E}_f \hat{f}(x_N; t_N) = f(x_N) + O\big(\sqrt{t_N}\big), \qquad N \to \infty,$$

where $x_N = \alpha t_N$ for some $\alpha \in [0, 1]$, and $t_N \downarrow 0$ as $N \to \infty$. This shows that (3) is consistent at the boundary $x = 0$. Similarly, (3) can be shown to be consistent at the boundary $x = 1$. In contrast, the Gaussian kernel density estimator (1) is inconsistent [53] in the sense that

$$\mathbb{E}_f \hat{f}(0; t_N) = \tfrac{1}{2} f(0) + O\big(\sqrt{t_N}\big), \qquad N \to \infty.$$

The large bandwidth behavior ($t \to \infty$) of (3) is obtained from the following equivalent expression for (4) (see [3]):

$$\kappa(x, X_i; t) = \sum_{k=-\infty}^{\infty} e^{-k^2\pi^2 t/2} \cos(k\pi x) \cos(k\pi X_i). \tag{5}$$

From (5), we immediately see that

$$\kappa(x, X_i; t) \sim 1 + 2e^{-\pi^2 t/2} \cos(\pi x) \cos(\pi X_i), \qquad t \to \infty, x \in [0, 1]. \tag{6}$$

In other words, as the bandwidth becomes larger and larger, the kernel (4) approaches the uniform density on $[0, 1]$.

REMARK 1. An important property of the estimator (3) is that the number of local maxima or modes is a nonincreasing function of t. This follows from the *maximum principle* for parabolic PDE; see, for example, [36].

For example, a necessary condition for a local maximum at, say, $(x_0, t_0), t_0 > 0, x_0 \in (0, 1)$ is $\frac{\partial^2}{\partial x^2} \hat{f}(x_0; t_0) \le 0$. From (2), this implies $\frac{\partial}{\partial t} \hat{f}(x_0; t_0) \le 0$, from which it follows that there exists an $\varepsilon > 0$ such that $\hat{f}(x_0; t_0) \ge \hat{f}(x_0; t_0 + \varepsilon)$. As a consequence of this, as t becomes larger and larger, the number of local maxima of (3) is a nonincreasing function. This property is shared by the Gaussian kernel density estimator (1) and has been exploited in various ways by Silverman [49].

EXAMPLE 1. Figure 1 gives an illustration of the performance of estimators (3) and (1), where the true p.d.f. is the beta density $4(1 - x)^3, x \in [0, 1]$, and the estimators are build from a sample of size $N = 1000$ with a common bandwidth $\sqrt{t} = 0.05248$. Note that the Gaussian kernel density estimator is close to half the value of the true p.d.f. at the boundary $x = 0$. Overall, the diffusion estimator (3) is much closer to the true p.d.f. The proposed estimator (3) appears to be the first

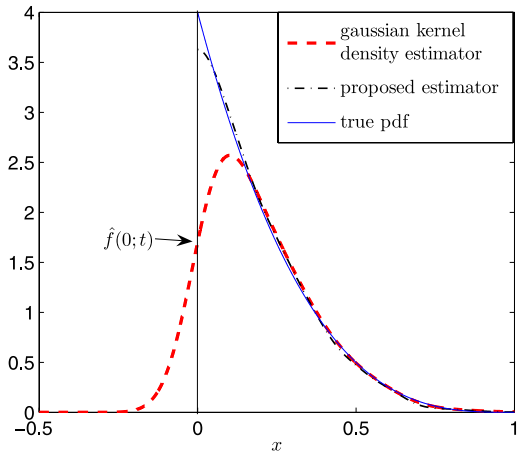

FIG. 1. *Boundary bias in the neighborhood of* $x = 0$.

kernel estimator that does not use boundary transformation and yet is consistent at all boundaries and remains a genuine p.d.f. (is nonnegative and integrates to one). Existing boundary correction methods [19, 31, 32] either account for the bias at a single end-point, or the resulting estimators are not genuine p.d.f.'s.

REMARK 2. In applications such as the smoothed bootstrap [11], there is a need for efficient random variable generation from the kernel density estimate. Generation of random variables from the kernel (4) is easily accomplished using the following procedure. Generate $Z \sim \mathsf{N}(0, t)$ and let $Y = X_i + Z$. Compute $W = Y \bmod 2$, and let $X = |W|$. Then it is easy to show (e.g., using characteristic functions) that X has the density given by (4).

Given the nice boundary bias properties of the estimator that arises as the solution of the diffusion PDE (2), it is of interest to investigate if equation (2) can be somehow modified or generalized to arrive at an even better kernel estimator. This motivates us to consider in the next section the most general linear time-homogeneous diffusion PDE as a starting point for the construction of a better kernel density estimator.

3. The diffusion estimator. Our extension of the simple diffusion model (2) is based on the smoothing properties of the linear diffusion PDE

$$\frac{\partial}{\partial t} g(x;t) = Lg(x;t), \qquad x \in \mathscr{X}, t > 0, \tag{7}$$

where the linear differential operator L is of the form $\frac{1}{2}\frac{d}{dx}(a(x)\frac{d}{dx}(\frac{\cdot}{p(x)}))$, and a and p can be any arbitrary positive functions on $\mathscr{X}$ with bounded second derivatives, and the initial condition is $g(x, 0) = \Delta(x)$. If the set $\mathscr{X}$ is bounded, we add

the boundary condition $\frac{\partial}{\partial x}(\frac{g(x;t)}{p(x)}) = 0$ on $\partial\mathscr{X}$, which ensures that the solution of (7) integrates to unity. The PDE (7) describes the p.d.f. of X_t for the Itô diffusion process $(X_t, t > 0)$ given by [12]

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dB_t, \tag{8}$$

where the drift coefficient $\mu(x) = \frac{a'(x)}{2p(x)}$, the diffusion coefficient $\sigma(x) = \sqrt{\frac{a(x)}{p(x)}}$, the initial state X_0 has distribution $\Delta(x)$, and $(B_t, t > 0)$ is standard Brownian motion. Obviously, if $a = 1$ and $p = 1$, we revert to the simpler model (2). What makes the solution $g(x;t)$ to (7) a plausible kernel density estimator is that $g(x;t)$ is a p.d.f. with the following properties. First, $g(\cdot;0)$ is identical to the initial condition of (7), that is, to the empirical density $\Delta(x)$. This property is possessed by both the Gaussian kernel density estimator (1) and the diffusion estimator (3). Second, if $p(x)$ is a p.d.f. on $\mathscr{X}$, then

$$\lim_{t\to\infty} g(x;t) = p(x), \qquad x \in \mathscr{X}.$$

This property is similar to the property that the kernel (6) and the estimator (3) converge to the uniform density on $\mathscr{X} \equiv [0,1]$ as $t \to \infty$. In the context of the diffusion process governed by (8), p is the limiting and stationary density of the diffusion. Third, similar to the estimator (3) and the Gaussian kernel density estimator (1), we can write the solution of (7) as

$$g(x;t) = \frac{1}{N}\sum_{i=1}^{N} \kappa(x, X_i; t), \tag{9}$$

where for each fixed $y \in \mathscr{X}$ the diffusion kernel κ satisfies the PDE

$$\begin{cases} \dfrac{\partial}{\partial t}\kappa(x,y;t) = L\kappa(x,y;t), & x \in \mathscr{X}, t > 0, \\ \kappa(x,y;0) = \delta(x-y), & x \in \mathscr{X}. \end{cases} \tag{10}$$

In addition, for each fixed $x \in \mathscr{X}$ the kernel κ satisfies the PDE

$$\begin{cases} \dfrac{\partial}{\partial t}\kappa(x,y;t) = L^*\kappa(x,y;t), & y \in \mathscr{X}, t > 0, \\ \kappa(x,y;0) = \delta(x-y), & y \in \mathscr{X}, \end{cases} \tag{11}$$

where L^* is of the form $\frac{1}{2p(y)}\frac{\partial}{\partial y}(a(y)\frac{\partial}{\partial y}(\cdot))$; that is, L^* is the adjoint operator of L. Note that L^* is the *infinitesimal generator* of the Itô diffusion process in (8). If the set $\mathscr{X}$ has boundaries, we add the Neumann boundary condition

$$\frac{\partial}{\partial x}\left(\frac{\kappa(x,y;t)}{p(x)}\right)\bigg|_{x\in\partial\mathscr{X}} = 0 \qquad \forall t > 0 \tag{12}$$

and $\frac{\partial}{\partial y}\kappa(x,y;t)|_{y\in\partial\mathscr{X}} = 0$ to (10) and (11), respectively. These boundary conditions ensure that $g(x;t)$ integrates to unity for all $t \geq 0$. The reason that the kernel

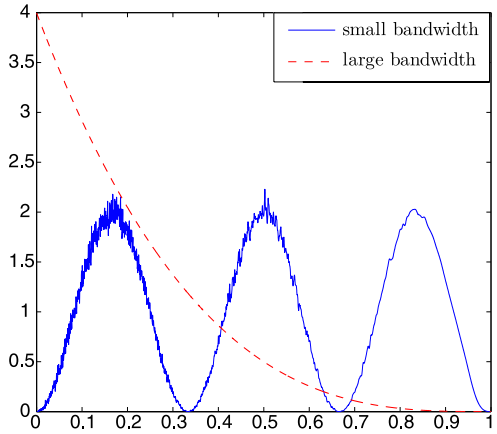

FIG. 2. *Small and large bandwidth behavior of the diffusion density in Example 2.*

κ satisfies both PDEs (10) and (11) is that (10) is the Kolmogorov forward equation [12] corresponding to the diffusion process (8), and (11) is a direct consequence of the Kolmogorov backward equation. We will use the forward and backward equations to derive the asymptotic properties of the diffusion estimator (9). Before we proceed with the asymptotic analysis, we illustrate how the model (7) possesses adaptive smoothing properties similar to the ones possessed by the adaptive kernel density estimators [1, 15, 16, 27].

EXAMPLE 2. Suppose that the initial condition of PDE (7) is $\Delta(x)$ with $N = 500{,}000$ and $X_1, \ldots, X_N$ are independent draws from $f(x) = 1 - \cos(6\pi x), x \in [0, 1]$. Suppose further that $p(x) = 4(1-x)^3$ and $a(x) = 1$ on $[0, 1]$. The aim of this example is not to estimate f, but to illustrate the various shapes that the estimator can take, given data from f. Figure 2 shows the solution of the PDE (7) for two values of the bandwidth: $\sqrt{t} = 4 \times 10^{-4}$ (small) and $\sqrt{t} = 0.89$ (large). Since $p(x)$ is the limiting and stationary density of the diffusion process governed by (7), the large bandwidth density is indistinguishable from $p(x)$. The small bandwidth density estimate is much closer to $f(x)$ than to $p(x)$. The crucial feature of the small bandwidth density estimate is that $p(x)$ allows for varying degrees of smoothing across the domain of the data, in particular allowing for greater smoothing to be applied in areas of sparse data, and relatively less in the high density regions. It can be seen from Figure 2 that the small time density estimate is noisier in regions where $p(x)$ is large (closer to $x = 0$), and smoother in regions where $p(x)$ is small (closer to $x = 1$). The adaptive smoothing is a consequence of the fact that the diffusion kernel (10) has a state-dependent diffusion coefficient $\sigma(x) = \sqrt{a(x)/p(x)}$, which helps diffuse the initial density $\Delta(x)$ at a different rate throughout the state space.

REMARK 3. Even though there is no analytical expression for the diffusion kernel satisfying (10), we can write κ in terms of a generalized Fourier series in the case that $\mathscr{X}$ is bounded:

$$\kappa(x, y; t) = p(x) \sum_{k=0}^{\infty} e^{\lambda_k t} \varphi_k(x) \varphi_k(y), \qquad x, y \in [0, 1], \tag{13}$$

where $\{\varphi_k\}$ and $\{\lambda_k\}$ are the eigenfunctions and eigenvalues of the Sturm–Liouville problem on $[0, 1]$:

$$\begin{aligned} L^* \varphi_k &= \lambda_k \varphi_k, \qquad k = 0, 1, 2, \ldots, \\ \varphi_k'(0) &= \varphi_k'(1) = 0, \qquad k = 0, 1, 2, \ldots. \end{aligned} \tag{14}$$

It is well known (see, e.g., [36]) that $\{\varphi_k\}$ forms a complete orthonormal basis with respect to the weight p for $L^2(0, 1)$. From the expression (13), we can see that the kernel satisfies the *detailed balance* equation for a continuous-time Markov process [12]

$$p(y)\kappa(x, y; t) = p(x)\kappa(y, x; t) \qquad \forall t > 0, x, y \in \mathscr{X}. \tag{15}$$

The detailed balance equation ensures that the limiting and stationary density of the diffusion estimator (9) is $p(x)$. In addition, the kernel satisfies the Chapman–Kolmogorov equation

$$\int_{\mathscr{X}} \kappa(x_1, x_0; t_1)\kappa(x_2, x_1; t_2)\, dx_1 = \kappa(x_2, x_0; t_1 + t_2). \tag{16}$$

Note that there is no loss of generality in assuming that the domain is $[0, 1]$, because any bounded domain can be mapped onto $[0, 1]$ by a linear transformation.

REMARK 4. When $p(x)$ is a p.d.f., an important distance measure between the diffusion estimator (9) and $p(x)$ is the divergence measure of Csiszár [9]. The Csiszár distance measure between two continuous probability densities g and p is defined as

$$\mathcal{D}(g \to p) = \int_{\mathbb{R}} p(x) \psi\left(\frac{g(x)}{p(x)}\right) dx,$$

where $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a twice continuously differentiable function; $\psi(1) = 0$; and $\psi''(x) > 0$ for all $x \in \mathbb{R}_+$. The diffusion estimator (9) possesses the monotonicity property

$$\frac{d}{dt}\mathcal{D}(g \to p) = -\frac{1}{2} \int_{\mathscr{X}} \left(\frac{g(x; t)}{p(x)}\right)^2 \psi''\left(\frac{g(x; t)}{p(x)}\right) dx < 0, \qquad g \neq p, t > 0.$$

In other words, the distance between the estimator (9) and the stationary density p is a monotonically decreasing function of the bandwidth $\sqrt{t}$. This is why the solution of (7) in Figure 2 approaches p as the bandwidth becomes larger and larger.

Note that Csiszár's family of measures subsumes all of the information-theoretic distance measures used in practice [21, 30]. For example, if $\psi(x) = \frac{x^\alpha - x}{\alpha(\alpha-1)}, \alpha \neq 0, 1$, for some parameter α, then the family of distances indexed by α includes the *Hellinger distance* for $\alpha = 1/2$, *Pearson's* χ^2 *discrepancy measure* for $\alpha = 2$, *Neymann's* χ^2 *measure* for $\alpha = -1$, the *Kullback–Leibler distance* in the limit as $\alpha \to 1$ and *Burg's distance* as $\alpha \to 0$.

4. Bias and variance analysis. We now examine the asymptotic bias, variance and MISE of the diffusion estimator (9). In order to derive the asymptotic properties of the proposed estimator, we need the small bandwidth behavior of the diffusion kernel satisfying (10). This is provided by the following lemma.

LEMMA 1. *Assume that the functions $a(x)$ and $p(x)$ are such that*

$$c_1 = \sqrt{\int_{-\infty}^{\infty} \left(\frac{Lq(z)}{q(z)}\right)^2 dz} < \infty, \qquad q(z) := \frac{p(z)}{a^{1/4}(z)p^{1/4}(z)}, \tag{17}$$
$$\lim_{z\to\infty} \int_{z_0}^{z} \sqrt{p(s)/a(s)}\, ds = \infty.$$

Then, the leading small bandwidth asymptotic behavior of the kernel satisfying (10) *and* (11) *on* $\mathscr{X} \equiv \mathbb{R}$ *is*

$$\kappa(x, y; t) \sim \frac{p(x)}{\sqrt{2\pi t}[p(x)a(x)a(y)p(y)]^{1/4}} \times \exp\left\{-\frac{1}{2t}\left[\int_y^x \sqrt{\frac{p(s)}{a(s)}}\, ds\right]^2\right\}, \qquad t \downarrow 0.$$

We denote the asymptotic approximation on the right-hand side by $\widetilde{\kappa}(x, y; t)$. *Thus,* $\kappa(x, y; t) \sim \widetilde{\kappa}(x, y; t)$ *as* $t \downarrow 0$.

The somewhat lengthy and technical proof is given in Appendix B. A few remarks about the technical conditions on a and p now follow. Conditions (17) are trivially satisfied if a, p and its derivatives up to order 2 are all bounded from above, and $p(x) \geq p_0 > 0$ and $a(x) \geq a_0 > 0$. In other words, if we clip $p(x)$ away from zero and use $a(x) = p^\alpha(x)$ for $\alpha \in [0, 1]$, then the conditions (17) are satisfied. Such clipping procedures have been applied in the traditional kernel density estimation setting, see [1, 7, 16, 18, 27]. Note that the conditions are more easily satisfied when p is heavy-tailed. For example, if $a(x) = p(x)$, then p could be any regularly varying p.d.f. of the form $p \propto (1 + |x|)^{-\alpha}, \alpha > 1$. Lemma 1 is required for deriving the asymptotic properties of the estimator, all collected in the following theorem.

THEOREM 1. *Let $t = t_N$ be such that $\lim_{N\to\infty} t_N = 0$, $\lim_{N\to\infty} N\sqrt{t_N} = \infty$. Assume that f is twice continuously differentiable and that the domain $\mathscr{X} \equiv \mathbb{R}$. Then*:

1. *The pointwise bias has the asymptotic behavior*

$$\mathbb{E}_f[g(x;t)] - f(x) = tLf(x) + O(t^2), \qquad N \to \infty. \tag{18}$$

2. *The integrated squared bias has the asymptotic behavior*

$$\|\mathbb{E}_f[g(\cdot;t)] - f\|^2 \sim t^2\|Lf\|^2 = \tfrac{1}{4}t^2\|(a(f/p)')'\|^2, \qquad N \to \infty. \tag{19}$$

3. *The pointwise variance has the asymptotic behavior*

$$\mathrm{Var}_f[g(x;t)] \sim \frac{f(x)}{2N\sqrt{\pi t}\sigma(x)}, \qquad N \to \infty, \tag{20}$$

where $\sigma^2(x) = a(x)/p(x)$.

4. *The integrated variance has the asymptotic behavior*

$$\int \mathrm{Var}_f[g(x;t)]\,dx \sim \frac{\mathbb{E}_f[\sigma^{-1}(X)]}{2N\sqrt{\pi t}}, \qquad N \to \infty. \tag{21}$$

5. *Combining the leading order bias and variance terms gives the asymptotic approximation to the MISE*

$$\mathrm{AMISE}\{g\}(t) = \frac{1}{4}t^2\|(a(f/p)')'\|^2 + \frac{\mathbb{E}_f[\sigma^{-1}(X)]}{2N\sqrt{\pi t}}. \tag{22}$$

6. *Hence, the square of the asymptotically optimal bandwidth is*

$$t^* = \left(\frac{\mathbb{E}_f[\sigma^{-1}(X)]}{2N\sqrt{\pi}\|Lf\|^2}\right)^{2/5}, \tag{23}$$

which gives the minimum

$$\min_t \mathrm{AMISE}\{g\}(t) = N^{-4/5}\frac{5[\mathbb{E}_f\sigma^{-1}(X)]^{4/5}\|Lf\|^{2/5}}{2^{14/5}\pi^{2/5}}. \tag{24}$$

The proof is given in Appendix C.

We make the following observations. First, if $p \not\equiv f$, the rate of convergence of (24) is $O(N^{-4/5})$, the same as the rate of the Gaussian kernel density estimator in (39). The multiplicative constant of $N^{-4/5}$ in (24), however, can be made very small by choosing p to be a *pilot density estimate* of f. Preliminary or pilot density estimates are used in most adaptive kernel methods [53]. Second, if $p \equiv f$, then the leading bias term (18) is 0. In fact, if f is infinitely smooth, the pointwise bias is exactly zero, as can be seen from

$$\mathbb{E}_f[g(x;t)] = \sum_{k=0}^{\infty}\frac{t^k}{k!}L^k f(x), \qquad f \in C^\infty,$$

where $L^{n+1} = LL^n$ and L^0 is the identity operator. In addition, if $a = p \propto 1$, then the bias term (18) is equivalent to the bias term (35) of the Gaussian kernel density estimator. Third, (20) suggests that in regions where the pilot density $p(x)$ is large [which is equivalent to small diffusion coefficient $\sigma(x)$] and $f(x)$ is large, the pointwise variance will be large. Conversely, in regions with few observations [i.e., where the diffusion coefficient $\sigma(x)$ is high and $f(x)$ is small] the pointwise variance is low. In other words, the ideal variance behavior results when the diffusivity $\sigma(x)$ behaves inversely proportional to $f(x)$.

4.1. *Special cases of the diffusion estimator.* We shall now show that the diffusion kernel estimator (9) is a generalization of some well-known modifications of the Gaussian kernel density estimator (1). Examples of modifications and improvements subsumed as special cases of (9) are as follows.

1. If $a(x) = p(x) \propto 1$ in (9) and $\mathscr{X} \equiv \mathbb{R}$, then the kernel κ reduces to the Gaussian kernel and we obtain (1).
2. If $a(x) = 1$ and $p(x) = f_p(x)$, where f_p is a clipped pilot density estimate of f (see [1, 18, 27]), then from Lemma 1, we have

$$\kappa(x, y; t) \sim \widetilde{\kappa}(x, y; t) = \frac{f_p(x)}{\sqrt{2\pi t}(f_p(x) f_p(y))^{1/4}} \exp\left\{-\frac{1}{2t}\left[\int_y^x \sqrt{f_p(s)}\, ds\right]^2\right\}.$$

   Thus, in the neighborhood of y such that $|x - y| = O(t^\beta)$, $\beta > 1/3$, we have

$$\kappa(x, y; t) \sim \frac{1}{\sqrt{2\pi t / f_p(x)}} \exp\left\{-\frac{(x-y)^2}{2t/f_p(x)}\right\}, \qquad t \downarrow 0.$$

   In other words, in the neighborhood of y, κ is asymptotically equivalent to a Gaussian kernel with mean y and bandwidth $\sqrt{t/f_p(y)}$, which is precisely the Abramson's variable bandwidth [1] modification as applied to the Gaussian kernel. Abramson's square root law states that the asymptotically optimal variable bandwidth is proportional to $f_p^{-1/2}(y)$.
3. If we choose $a(x) = p(x) = f_p(x)$, then in an $O(t^\beta)$, $\beta > 0$ neighborhood of y, the kernel $\kappa(x, y; t)$ behaves asymptotically as a Gaussian kernel with location $y + \frac{t}{2}\frac{f_p'(y)}{f_p(y)}$ and bandwidth $\sqrt{t}$:

$$\kappa(x, y; t) \sim \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{1}{2t}\left(x - y - \frac{t}{2}\frac{f_p'(y)}{f_p(y)}\right)^2\right\}, \qquad t \downarrow 0.$$

   This is precisely the data sharpening modification described in [46], where the locations of the data points are shifted prior to the application of the kernel density estimate. Thus, in our paradigm, data sharpening is equivalent to using the diffusion (7) with drift $\mu(x) = \frac{f_p'(x)}{2f_p(x)}$ and diffusion coefficient $\sigma(x) = 1$.

4. Finally, if we set $p(x) = f_p(x)$ and $a(x) = p^\alpha(x)$, $\alpha \in [0,1]$, then we obtain a method that is a combination of both the data sharpening and the variable bandwidth of Abramson. The kernel κ behaves asymptotically [in an $O(t^\beta)$, $\beta > 1/3$ neighborhood of y] like a Gaussian kernel with location $y + t\mu(y) = y + \frac{\alpha t}{2} f_p^{\alpha-2}(y) f_p'(y)$ and bandwidth $\sqrt{t\sigma^2(y)} = \sqrt{t f_p^{\alpha-1}(y)}$. Similar variable location and scale kernel density estimators are considered in [27].

The proposed method thus unifies many of the already existing ideas for variable scale and location kernel density estimators. Note that these estimators all have one common feature: they compute a pilot density estimate (which is an infinite-dimensional parameter) prior to the main estimation step.

Our choice for $a(x)$ will be motivated by regularity properties of the diffusion process underlying the smoothing kernel. In short, we prefer to choose $a(x) = 1$ so as to make the diffusion process in (8) nonexplosive with a well-defined limiting distribution. A necessary and sufficient condition for explosions is Feller's test [13].

THEOREM 2 (Feller's test). *Let $\mu(x) > 0$ and $\sigma(x) > 0$ be bounded and continuous. Then the diffusion process (8) explodes if and only if there exists $z \in \mathbb{R}$ such that either one of the following two conditions holds*:

1.
$$\int_{-\infty}^{z} \int_{x}^{z} \exp\left(\int_{x}^{y} \frac{2\mu(s)}{\sigma^2(s)}\, ds \right) \sigma^{-2}(y)\, dy\, dx < \infty,$$

2.
$$\int_{z}^{\infty} \int_{z}^{x} \exp\left(\int_{x}^{y} \frac{2\mu(s)}{\sigma^2(s)}\, ds \right) \sigma^{-2}(y)\, dy\, dx < \infty.$$

A corollary of Feller's test is that when $\mu(x) = 0$ both of Feller's conditions fail, and diffusions of the form $dX_t = \sigma_t\, dW_t$ are nonexplosive.

Since in our case we have $\sigma^2(x) = a(x)/p(x)$ and $a(x) = \exp(\int_{x_0}^{x} 2\mu(y)/\sigma^2(y)\, dy)$, Feller's condition becomes the following.

PROPOSITION 1 (Feller's test). *Given $a(x)$ and $p(x)$ in (7), the diffusion process (8) explodes if and only if there exists $z \in \mathbb{R}$ such that either one of the following two conditions holds*:

1.
$$\int_{-\infty}^{z} \int_{x}^{z} \frac{p(y)}{a(x)}\, dy\, dx < \infty,$$

2.
$$\int_{z}^{\infty} \int_{z}^{x} \frac{p(y)}{a(x)}\, dy\, dx < \infty.$$

The easiest way to ensure nonexplosiveness of the underlying diffusion process and the existence of a limiting distribution is to set $a(x) = 1$, which corresponds to $\mu(x) = 0$. Note that a necessary condition for the existence of a limiting p.d.f. is the existence of z such that $\int_z^\infty 1/a(x)\,dx = \infty$. In this case, both of Feller's conditions fail. The nonexplosiveness property ensures that generation of random variables from the diffusion estimator does not pose any technical problems.

5. Bandwidth selection algorithm. Before we explain how to estimate the bandwidth $\sqrt{t^*}$ in (23) of the diffusion estimator (9), we explain how to estimate the bandwidth $\sqrt{{}_*t}$ in (38) (see Appendix A) of the Gaussian kernel density estimator (1). Here, we present a new plug-in bandwidth selection procedure based on the ideas in [23, 26, 40, 48] to achieve unparalleled practical performance. The highlighting feature of the proposed method is that it does not use normal reference rules and is thus completely data-driven.

It is clear from (38) in Appendix A that to compute the optimal ${}_*t$ for the Gaussian kernel density estimator (1) one needs to estimate the functional $\|f''\|^2$. Thus, we consider the problem of estimating $\|f^{(j)}\|^2$ for an arbitrary integer $j \geq 1$. The identity $\|f^{(j)}\|^2 = (-1)^j \mathbb{E}_f[f^{(2j)}(X)]$ suggests two possible plug-in estimators. The first one is

$$\begin{aligned}(-1)^j \widehat{\mathbb{E}_f f^{(2j)}} &:= \frac{(-1)^j}{N} \sum_{k=1}^N \hat{f}^{(2j)}(X_k; t_j) \\ &= \frac{(-1)^j}{N^2} \sum_{k=1}^N \sum_{m=1}^N \phi^{(2j)}(X_k, X_m; t_j),\end{aligned} \tag{25}$$

where $\hat{f}$ is the Gaussian kernel density estimator (1). The second estimator is

$$\begin{aligned}\widehat{\|f^{(j)}\|^2} &:= \|\hat{f}^{(j)}(\cdot; t)\|^2 \\ &= \frac{1}{N^2} \sum_{k=1}^N \sum_{m=1}^N \int_{\mathbb{R}} \phi^{(j)}(x, X_k; t_j)\phi^{(j)}(x, X_m; t_j)\,dx \\ &= \frac{(-1)^j}{N^2} \sum_{k=1}^N \sum_{m=1}^N \phi^{(2j)}(X_k, X_m; 2t_j),\end{aligned} \tag{26}$$

where the last line is a simplification following easily from the fact that the Gaussian kernel ϕ satisfies the Chapman–Kolmogorov equation (16). For a given bandwidth, both estimators $(-1)^j \widehat{\mathbb{E}_f f^{(2j)}}$ and $\widehat{\|f^{(j)}\|^2}$ aim to estimate the same quantity, namely $\|f^{(j)}\|^2$. We select t_j so that both estimators (25) and (26) are asymptotically equivalent in the mean square error sense. In other words, we choose $t_j = {}_*t_j$ so that both $(-1)^j \widehat{\mathbb{E}_f f^{(2j)}}$ and $\widehat{\|f^{(j)}\|^2}$ have equal asymptotic mean square error. This gives the following proposition.

PROPOSITION 2. *The estimators* $(-1)^j\widehat{\mathbb{E}_f f^{(2j)}}$ *and* $\widehat{\|f^{(j)}\|^2}$ *have the same asymptotic mean square error when*

$$
{}_*t_j = \left(\frac{1+1/2^{j+1/2}}{3}\frac{1\times 3\times 5\times\cdots\times(2j-1)}{N\sqrt{\pi/2}\|f^{(j+1)}\|^2}\right)^{2/(3+2j)}. \tag{27}
$$

PROOF. The arguments are similar to the ones used in [53]. Under the assumptions that t_j depends on N such that $\lim_{N\to\infty} t_j = 0$ and $\lim_{N\to\infty} N t_j^{j+1/2} = \infty$, we can take the expectation of the estimator (25) and obtain the expansion ($t_j = t$):

$$
\begin{aligned}
&\mathbb{E}_f\big[\widehat{\mathbb{E}_f f^{(2j)}}\big]\\
&\quad= \frac{1}{N}\phi^{(2j)}(0,0;t) + \frac{N-1}{N}\iint f(x)f(y)\phi^{(2j)}(x,y;t)\,dx\,dy\\
&\quad= -\frac{1\times 3\times\cdots\times(2j-1)}{t^{j+1/2}\sqrt{2\pi}N}\\
&\qquad+\int f(x)\left(f^{(2j)}(x)+\frac{t}{2}f^{2(j+1)}(x)+o(t)\right)dx + O(N^{-1})\\
&\quad= -\frac{1\times 3\times 5\times\cdots\times(2j-1)}{t^{j+1/2}\sqrt{2\pi}N} + \frac{t}{2}\|f^{(j+1)}\|^2\\
&\qquad+(-1)^j\|f^{(j)}\|^2 + O(N^{-1}), \qquad N\to\infty.
\end{aligned}
$$

Hence, the squared bias has asymptotic behavior ($N\to\infty$)

$$
\big((-1)^j\mathbb{E}_f\big[\widehat{\mathbb{E}_f f^{(2j)}}\big] - \|f^{(j)}\|^2\big)^2 \sim \left(\frac{1\times 3\times\cdots\times(2j-1)}{t^{j+1/2}\sqrt{2\pi}N} - \frac{t}{2}\|f^{(j+1)}\|^2\right)^2.
$$

A similar argument (see [53]) shows that the variance is of the order $O(N^{-2}\times t^{-2j-1/2})$, which is of lesser order than the squared bias. This implies that the leading order term in the asymptotic mean square error of $\widehat{\mathbb{E}_f f^{(2j)}}$ is given by the asymptotic squared bias. There is no need to derive the asymptotic expansion of $\mathbb{E}_f[\widehat{\|f^{(j)}\|^2}]$, because inspection of (26) and (25) shows that $\widehat{\|f^{(j)}\|^2}$ exactly equals $(-1)^j\widehat{\mathbb{E}_f f^{(2j)}}$ when the latter is evaluated at $2t_j$. In other words,

$$
\begin{aligned}
(-1)^j\mathbb{E}_f\big[\widehat{\|f^{(j)}\|}^2\big] &= -\frac{1\times 3\times 5\times\cdots\times(2j-1)}{(2t)^{j+1/2}\sqrt{2\pi}N}\\
&\quad+t\|f^{(j+1)}\|^2 + O(1+N^{-1}).
\end{aligned}
$$

Again, the leading term of the asymptotic mean square error of $\widehat{\|f^{(j)}\|^2}$ is given by the leading term of the squared bias of $\widehat{\|f^{(j)}\|^2}$. Thus, equalizing the asymptotic

mean squared error of both estimators is the same as equalizing their respective asymptotic squared biases. This yields the equation

$$\left(\frac{1\times 3\times\cdots\times(2j-1)}{(2t)^{j+1/2}\sqrt{2\pi}N}-t\|f^{(j+1)}\|^2\right)^2 = \left(\frac{1\times 3\times\cdots\times(2j-1)}{t^{j+1/2}\sqrt{2\pi}N}-\frac{t}{2}\|f^{(j+1)}\|^2\right)^2.$$

The positive solution of the equation yields the desired ${}_*t_j$. □

Thus, for example,

$$ {}_*t_2=\left(\frac{8+\sqrt{2}}{24}\frac{3}{N\sqrt{\pi/2}\|f^{(3)}\|^2}\right)^{2/7} \tag{28}$$

is our bandwidth choice for the estimation of $\|f''\|^2$. We estimate each ${}_*t_j$ by

$$ {}_*\hat{t}_j=\left(\frac{1+1/2^{j+1/2}}{3}\frac{1\times 3\times 5\times\cdots\times(2j-1)}{N\sqrt{\pi/2}\|\widehat{f^{(j+1)}}\|^2}\right)^{2/(3+2j)}. \tag{29}$$

Computation of $\|\widehat{f^{(j+1)}}\|^2$ requires estimation of ${}_*t_{j+1}$ itself, which in turn requires estimation of ${}_*t_{j+2}$, and so on, as seen from formulas (26) and (29). We are faced with the problem of estimating the infinite sequence $\{{}_*t_{j+k}, k\ge 1\}$. It is clear, however, that given ${}_*t_{l+1}$ for some $l>0$ we can estimate all $\{{}_*t_j, 1\le j\le l\}$ recursively, and then estimate ${}_*t$ itself from (38). This motivates the *l-stage direct plug-in bandwidth selector* [26, 48, 53], defined as follows.

1. For a given integer $l>0$, estimate ${}_*t_{l+1}$ via (27) and $\|f^{(l+2)}\|^2$ computed by assuming that f is a normal density with mean and variance estimated from the data. Denote the estimate by ${}_*\hat{t}_{l+1}$.
2. Use ${}_*\hat{t}_{l+1}$ to estimate $\|f^{(l+1)}\|^2$ via the plug-in estimator (26) and ${}_*\hat{t}_l$ via (29). Then use ${}_*\hat{t}_l$ to estimate ${}_*\hat{t}_{l-1}$ and so on until we obtain an estimate of ${}_*\hat{t}_2$.
3. Use the estimate of ${}_*\hat{t}_2$ to compute ${}_*\hat{t}$ from (38).

The l-stage direct plug-in bandwidth selector thus involves the estimation of l functionals $\{\|f^{(j)}\|, 2\le j\le l+1\}$ via the plug-in estimator (26). We can describe the procedure in a more abstract way as follows. Denote the functional dependence of ${}_*\hat{t}_j$ on ${}_*\hat{t}_{j+1}$ in formula (29) as

$${}_*\hat{t}_j=\gamma_j({}_*\hat{t}_{j+1}).$$

It is then clear that ${}_*\hat{t}_j=\gamma_j(\gamma_{j+1}({}_*\hat{t}_{j+2}))=\gamma_j(\gamma_{j+1}(\gamma_{j+2}({}_*\hat{t}_{j+3})))=\cdots$. For simplicity of notation, we define the composition

$$\gamma^{[k]}(t)=\underbrace{\gamma_1(\cdots\gamma_{k-1}(\gamma_k}_{k\text{ times}}(t))\cdots), \qquad k\ge 1.$$

Inspection of formulas (29) and (38) shows that the estimate of ${}_*t$ satisfies

$$
\begin{aligned}
{}_*\hat{t} &= \xi\, {}_*\hat{t}_1 = \xi\gamma^{[1]}({}_*\hat{t}_2) = \xi\gamma^{[2]}({}_*\hat{t}_3) = \cdots = \xi\gamma^{[l]}({}_*\hat{t}_{1+l}),\\
\xi &= \left(\frac{6\sqrt{2}-3}{7}\right)^{2/5} \approx 0.90.
\end{aligned}
$$

Then, for a given integer $l > 0$, the l-stage direct plug-in bandwidth selector consists of computing

$$
{}_*\hat{t} = \xi\gamma^{[l]}({}_*t_{l+1}),
$$

where ${}_*t_{l+1}$ is estimated via (27) by assuming that f in $\|f^{(l+2)}\|^2$ is a normal density with mean and variance estimated from the data. The weakest point of this procedure is that we assume that the true f is a Gaussian density in order to compute $\|f^{(l+2)}\|^2$. This assumption can lead to arbitrarily bad estimates of ${}_*t$, when, for example, the true f is far from being Gaussian. Instead, we propose to find a solution to the nonlinear equation

$$
t = \xi\gamma^{[l]}(t), \tag{30}
$$

for some l, using either fixed point iteration or Newton's method with initial guess $t = 0$. The fixed point iteration version is formalized in the following algorithm.

ALGORITHM 1 (Improved Sheather–Jones). Given $l > 2$, execute the following steps:

1. initialize with $z_0 = \varepsilon$, where ε is machine precision, and $n = 0$;
2. set $z_{n+1} = \xi\gamma^{[l]}(z_n)$;
3. if $|z_{n+1} - z_n| < \varepsilon$, stop and set ${}_*\hat{t} = z_{n+1}$; otherwise, set $n := n + 1$ and repeat from step 2;
4. deliver the Gaussian kernel density estimator (1) evaluated at ${}_*\hat{t}$ as the final estimator of f, and ${}_*\hat{t}_2 = \gamma^{[l-1]}(z_{n+1})$ as the bandwidth for the optimal estimation of $\|f''\|^2$.

Numerical experience suggests the following. First, the fixed-point algorithm does not fail to find a root of the equation $t = \xi\gamma^{[l]}(t)$. Second, the root appears to be unique. Third, the solutions to the equations

$$
t = \xi\gamma^{[5]}(t)
$$

and

$$
t = \xi\gamma^{[l+5]}(t)
$$

for any $l > 0$ do not differ in any practically meaningful way. In other words, there were no gains to be had by increasing the stages of the bandwidth selection rule beyond $l = 5$. We recommend setting $l = 5$. Finally, the numerical procedure for

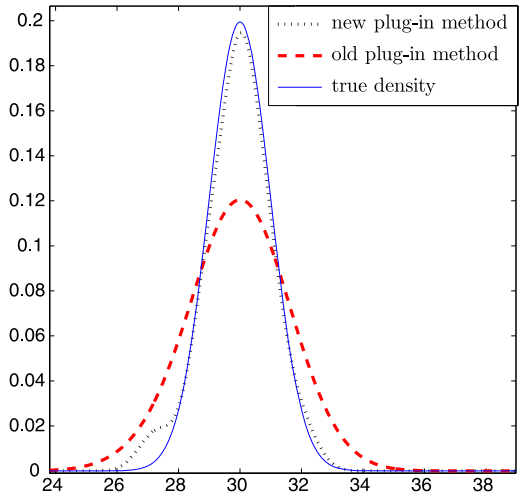

FIG. 3. *The Improved Sheather–Jones bandwidth selection rule in Algorithm 1 leads to improved performance compared to the original plug-in rule that uses the normal reference rule.*

the computation of $\gamma^{[5]}(t)$ is fast when implemented using the Discrete Cosine Transform [4].

The plug-in method described in Algorithm 1 has superior practical performance compared to existing plug-in implementations, including the particular *solve-the-equation* rule of Sheather and Jones [48, 53]. Since we borrow many of the fruitful ideas described in [48] (which in turn build upon the work of Hall, Park and Marron [17, 45]), we call our new algorithm the Improved Sheather–Jones (ISJ) method.

To illustrate the significant improvement of the plug-in method in Algorithm 1, consider, for example, the case where f is a mixture of two Gaussian densities with a common variance of 1 and means of -30 and 30.

Figure 3 shows the right mode of f, and the two estimates resulting from the old plug-in rule [48] and the plug-in rule of Algorithm 1. The left mode is not displayed, but looks similar. The integrated squared error using the new plug-in bandwidth estimate, $\|f - \hat{f}(\cdot; {}_*\hat{t})\|^2$, is one 10th of the error using the old bandwidth selection rule.

5.1. *Experiments with normal reference rules.* The result of Figure 3 is not an isolated case, in which the normal reference rules do not perform well. We performed a comprehensive simulation study in order to compare the Improved Sheather–Jones (ISJ) (Algorithm 1) with the original (vanilla) Sheather–Jones (SJ) algorithm [48, 53].

Table 1 shows the average results over 10 independent trials for a number of different test cases. The second column displays the target density and the third

TABLE 1

Results over 10 *independent simulation experiments. In all cases the domain was assumed to be* $\mathbb{R}$. *Many test problems are taken from* [42]. *In the table* $\mathsf{N}(\mu, \sigma^2)$, *denotes a Gaussian density with mean* μ *and variance* σ^2

Case	**Target density** $f(x)$	N	**Ratio**
1 (claw)	$\frac{1}{2}\mathsf{N}(0,1) + \sum_{k=0}^{4} \frac{1}{10}\mathsf{N}(\frac{k}{2}-1, (\frac{1}{10})^2)$	10^3	0.72
		10^4	0.94
2 (strongly skewed)	$\sum_{k=0}^{7} \frac{1}{8}\mathsf{N}(3((\frac{2}{3})^k - 1), (\frac{2}{3})^{2k})$	10^3	0.69
		10^4	0.84
3 (kurtotic unimodal)	$\frac{2}{3}\mathsf{N}(0,1) + \frac{1}{3}\mathsf{N}(0, (\frac{1}{10})^2)$	10^2	0.78
		10^3	0.93
4 (double claw)	$\frac{49}{100}\mathsf{N}(-1, (\frac{2}{3})^2) + \frac{49}{100}\mathsf{N}(1, (\frac{2}{3})^2)$	10^5	0.35
	$+ \frac{1}{350}\sum_{k=0}^{6} \mathsf{N}(\frac{k-3}{2}, (\frac{1}{100})^2)$	10^6	0.10
5 (discrete comb)	$\frac{2}{7}\sum_{k=0}^{2} \mathsf{N}(\frac{12k-15}{7}, (\frac{2}{7})^2) + \frac{1}{21}\sum_{k=8}^{10} \mathsf{N}(\frac{2k}{7}, (\frac{1}{21})^2)$	10^3	0.45
		10^4	0.27
6 (asymmetric	$\frac{46}{100}\sum_{k=0}^{1} \mathsf{N}(2k-1, (\frac{2}{3})^2) + \sum_{k=1}^{3} \frac{1}{300}\mathsf{N}(-\frac{k}{2}, (\frac{1}{100})^2)$	10^4	0.68
double claw)	$+ \sum_{k=1}^{3} \frac{7}{300}\mathsf{N}(\frac{k}{2}, (\frac{7}{100})^2)$	10^6	0.24
7 (outlier)	$\frac{1}{10}\mathsf{N}(0,1) + \frac{9}{10}\mathsf{N}(0, (\frac{1}{10})^2)$	10^3	1.01
		10^5	1.00
8 (separated bimodal)	$\frac{1}{2}\mathsf{N}(-12, \frac{1}{4}) + \frac{1}{2}\mathsf{N}(12, \frac{1}{4})$	10^2	0.33
		10^3	0.64
9 (skewed bimodal)	$\frac{3}{4}\mathsf{N}(0,1) + \frac{1}{4}\mathsf{N}(\frac{3}{2}, (\frac{1}{3})^2)$	10^3	1.02
		10^4	1.00
10 (bimodal)	$\frac{1}{2}\mathsf{N}(0, (\frac{1}{10})^2) + \frac{1}{2}\mathsf{N}(5,1)$	10^2	0.31
		10^3	0.70
11	Log-Normal with $\mu = 0$ and $\sigma = 1$	10^3	0.82
		10^4	0.80
12 (asymmetric claw)	$\frac{1}{2}\mathsf{N}(0,1) + \sum_{k=-2}^{2} \frac{2^{1-k}}{31}\mathsf{N}(k+\frac{1}{2}, (\frac{2^{-k}}{10})^2)$	10^3	0.76
		10^4	0.59
13 (trimodal)	$\frac{1}{3}\sum_{k=0}^{2} \mathsf{N}(80k; (k+1)^4)$	10^2	0.21
		10^3	0.17
14 (5-modes)	$\frac{1}{5}\sum_{k=0}^{4} \mathsf{N}(80k; (k+1)^2)$	10^3	0.07
		10^4	0.18
15 (10-modes)	$\frac{1}{10}\sum_{k=0}^{9} \mathsf{N}(100k; (k+1)^2)$	10^3	0.12
		10^4	0.07
16 (smooth comb)	$\sum_{k=0}^{5} \frac{2^{5-k}}{63}\mathsf{N}(\frac{65-96/2^k}{21}; \frac{(32/63)^2}{2^{2k}})$	10^4	0.40
		10^5	0.34

column shows the sample size used for the experiments. The last column shows our criterion for comparison:

$$R = \frac{\|\hat{f}(\cdot;\, {}_*\hat{t}) - f\|^2}{\|\hat{f}(\cdot;\, t_{\mathrm{SJ}}) - f\|^2},$$

that is, the ratio of the integrated squared error of the new ISJ estimator to the integrated squared error of the original SJ estimator. Here, t_{SJ} is the bandwidth computed using the original Sheather–Jones method [48, 53].

The results in Table 1 show that the improvement in the integrated squared error can be as much as ten-fold, and the ISJ method outperforms the SJ method in almost all cases. The evidence suggests that discarding the normal reference rules, widely employed by most plug-in rules, can significantly improve the performance of the plug-in methods.

The multi-modal test cases 12 through 16 in Table 1 and Figure 3 demonstrate that the new bandwidth selection procedure passes the *bi-modality test* [10], which consists of testing the performance of a bandwidth selection procedure using a bimodal target density, with the two modes at some distance from each other. It has been demonstrated in [10] that, by separating the modes of the target density enough, existing plug-in selection procedures can be made to perform arbitrarily poorly due to the adverse effects of the normal reference rules. The proposed plug-in method in Algorithm 1 performs much better than existing plug-in rules, because it uses the theoretical ideas developed in [48], except for the detrimental normal reference rules. A Matlab implementation of Algorithm 1 is freely available from [4], and includes other examples of improved performance.

Algorithm 1 can be extended to bandwidth selection in higher dimensions. For completeness we describe the two-dimensional version of the algorithm in Appendix E. The advantages of discarding the normal reference rules persist in the two-dimensional case. In other words, the good performance of the proposed method in two dimensions is similar to that observed in the univariate case. For example, Figure 4 shows the superior performance of the ISJ method compared to a plug-in approach using the normal reference rule [52, 53], and with kernels assumed to have a diagonal covariance matrix with a single smoothing parameter: $\Sigma = tI$. We estimate the bivariate density, $\frac{1}{4}\sum_{k=1}^{4} \mathsf{N}(\boldsymbol{\mu}_k, I)$, from a sample of size $N = 400$, where

$$\boldsymbol{\mu}_1 = (0, 0), \qquad \boldsymbol{\mu}_2 = (0, 50), \qquad \boldsymbol{\mu}_3 = (50, 0), \qquad \boldsymbol{\mu}_4 = (50, 50).$$

Note that using a plug-in rule with a normal reference rule causes significant oversmoothing. The integrated squared error for the ISJ method is 10 times smaller than the corresponding error for the plug-in rule that uses a normal reference rule [52, 53].

5.2. *Bandwidth selection for the diffusion estimator.* We now discuss the bandwidth choice for the diffusion estimator (9). In the following argument we

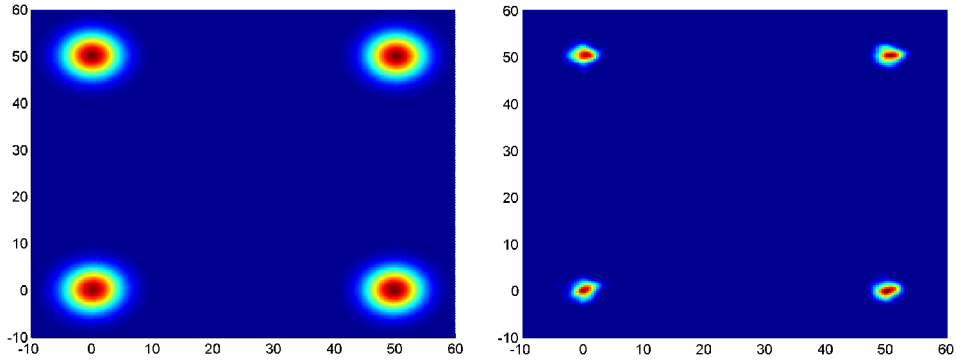

FIG. 4. *Right panel*: *plug-in rule with normal reference rule*; *left panel*: *the Improved Sheather–Jones method*; *the normal reference rule causes significant over-smoothing*.

assume that f is as many times continuously differentiable as needed. Computation of t^* in (23) requires an estimate of $\|Lf\|^2$ and $\mathbb{E}_f[\sigma^{-1}(X)]$. We estimate $\mathbb{E}_f[\sigma^{-1}(X)]$ via the unbiased estimator $\frac{1}{N}\sum_{i=1}^N \sigma^{-1}(X_i)$. The identity $\|Lf\|^2 = \mathbb{E}_f L^*Lf(X)$ suggests two possible plug-in estimators. The first one is

$$\begin{aligned}\widehat{\mathbb{E}_f L^*Lf} &:= \frac{1}{N}\sum_{j=1}^N L^*Lg(x;t_2)\Big|_{x=X_j} \\ &= \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N L^*L\kappa(x,X_i;t_2)\Big|_{x=X_j},\end{aligned} \tag{31}$$

where $g(x;t_2)$ is the diffusion estimator (9) evaluated at t_2, and $\mathscr{X} \equiv \mathbb{R}$. The second estimator is

$$\begin{aligned}\widehat{\|Lf\|}^2 &:= \|Lg(\cdot;t_2)\|^2 \\ &= \left\|\frac{\partial g}{\partial t}(\cdot;t_2)\right\|^2 \\ &= \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \int_{\mathbb{R}} \frac{\partial\kappa}{\partial t}(x,X_i;t_2)\frac{\partial\kappa}{\partial t}(x,X_j;t_2)\,dx \\ &= \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N L^*L\kappa(x,X_i;2t_2)\Big|_{x=X_j},\end{aligned} \tag{32}$$

where the last line is a simplification that follows from the Chapman–Kolmogorov equation (16). The optimal t_2^* is derived in the same way that ${}_*t_2$ is derived for the Gaussian kernel density estimator. That is, t_2^* is such that both estimators $\widehat{\mathbb{E}_f L^*Lf}$ and $\widehat{\|Lf\|}^2$ have the same asymptotic mean square error. This leads to the following proposition.

PROPOSITION 3. *The estimators* $\widehat{\mathbb{E}_f L^*Lf}$ *and* $\widehat{\|Lf\|^2}$ *have the same asymptotic mean square error when*

$$t_2^* = \left(\frac{8+\sqrt{2}}{24}\frac{-3\sqrt{2}\mathbb{E}_f[\sigma^{-1}(X)]}{8\sqrt{\pi}N\mathbb{E}_f[L^*L^2 f(X)]}\right)^{2/7}. \tag{33}$$

PROOF. Although the relevant calculations are lengthier, the arguments here are exactly the same as the ones used in Proposition 1. In particular, we have the same assumptions on t about its dependence on N. For simplicity of notation, the operators L^* and L are here assumed to apply to the first argument of the kernel κ:

$$\begin{aligned}
&\mathbb{E}_f[\widehat{\mathbb{E}_f L^*Lf}]\\
&\quad= \mathbb{E}_f \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N L^*L\kappa(x, X_i; t)\Big|_{x=X_j}\\
&\quad= \frac{1}{N}\int f(x)L^*L\kappa(x, X_i; t)\Big|_{X_i=x}\,dx\\
&\qquad+ \frac{N-1}{N}\iint f(y)f(x)L^*L\kappa(x, y; t)\,dy\,dx\\
&\quad= \frac{3\sqrt{2}\mathbb{E}_f[\sigma^{-1}(X)]}{8\sqrt{\pi}t^{5/2}N} + O(N^{-1}t^{-3/2})\\
&\qquad+ \iint f(y)f(x)L^*L\kappa(x, y; t)\,dy\,dx + O(N^{-1})\\
&\quad= \frac{3\sqrt{2}\mathbb{E}_f[\sigma^{-1}(X)]}{8\sqrt{\pi}t^{5/2}N} + \int f(y)\int L^*Lf(x)\kappa(x, y; t)\,dx\,dy\\
&\qquad+ O\big(N^{-1}(1+t^{-3/2})\big)\\
&\quad= \frac{3\sqrt{2}\mathbb{E}_f[\sigma^{-1}(X)]}{8\sqrt{\pi}t^{5/2}N} + \|Lf\|^2 + t\int f(y)L^*L^2 f(y)\,dy\\
&\qquad+ O\big(N^{-1}(1+t^{-3/2}) + t^2\big),
\end{aligned}$$

where we have used a consequence of Lemma 1,

$$\int f(x)L^*L\kappa(x, X_i; t)\Big|_{X_i=x}\,dx \sim \frac{3\sqrt{2}\mathbb{E}_f[\sigma^{-1}(X)]}{8\sqrt{\pi}t^{5/2}}, \qquad t\downarrow 0,$$

and a consequence of the detailed balance equation (15),

$$\begin{aligned}
\int L^*Lf(x)\kappa(x, y; t)\,dx &= \int \frac{p(x)L^*Lf(x)}{p(y)}\kappa(y, x; t)\,dx\\
&= L^*Lf(y) + tL^*L^*Lf(y) + O(t^2).
\end{aligned}$$

Therefore, the squared bias has asymptotic behavior ($N \to \infty$)

$$(\mathbb{E}_f[\widehat{\mathbb{E}_f L^* L f}] - \|Lf\|^2)^2 \sim \left(\frac{3\sqrt{2}\mathbb{E}_f[\sigma^{-1}(X)]}{8\sqrt{\pi} t^{5/2} N} + t \int f(y) L^* L^2 f(y)\, dy \right)^2.$$

Since estimator $\widehat{\|Lf\|^2}$ equals $\widehat{\mathbb{E}_f L^* L f}$ when the latter is evaluated at $2t_2$, the asymptotic squared bias of $\widehat{\|Lf\|^2}$ follows immediately, and we simply repeat the arguments in the proof of Proposition 1 to obtain the desired t_2^*. $\square$

Note that t_2^* has the same rate of convergence to 0 as ${}_*t_2$ in (28). In fact, since the Gaussian kernel density estimator is a special case of the diffusion estimator (9) when $p(x) = a(x) = 1$, the plug-in estimator (32) for the estimation of $\|Lf\|^2$ reduces to the plug-in estimator for the estimation of $\frac{1}{4}\|f''\|^2$. In addition, when $p(x) = a(x) = 1$, the t_2^* in (33) and ${}_*t_2$ in (28) are identical. We thus suggest the following bandwidth selection and estimation procedure for the diffusion estimator (9).

ALGORITHM 2.

1. Given the data $X_1, \ldots, X_N$, run Algorithm 1 to obtain the Gaussian kernel density estimator (1) evaluated at ${}_*\hat{t}$ and the optimal bandwidth $\sqrt{{}_*\hat{t}_2}$ for the estimation of $\|f''\|^2$. This is the pilot estimation step.
2. Let $p(x)$ be the Gaussian kernel density estimator from step 1, and let $a(x) = p^\alpha(x)$ for some $\alpha \in [0, 1]$.
3. Estimate $\|Lf\|^2$ via the plug-in estimator (32) using $\hat{t}_2^* = {}_*\hat{t}_2$, where ${}_*\hat{t}_2$ is computed in step 1.
4. Substitute the estimate of $\|Lf\|^2$ into (23) to obtain an estimate for t^*.
5. Deliver the diffusion estimator (9) evaluated at $\hat{t}^*$ as the final density estimate.

The bandwidth selection rule that we use for the diffusion estimator in Algorithm 2 is a single stage direct plug-in bandwidth selector, where the bandwidth t_2^* for the estimation of the functional $\|Lf\|^2$ is approximated by ${}_*\hat{t}_2$ (which is computed in Algorithm 1), instead of being derived from a normal reference rule. In the next section, we illustrate the performance of Algorithm 2 using some well-known test cases for density estimation.

REMARK 5 (Random variable generation). For applications of kernel density estimation, such as the smoothed bootstrap, efficient random variable generation from the diffusion estimator (9) is accomplished via the Euler method as applied to the stochastic differential equation (8) (see [34]).

ALGORITHM 3.

1. Subdivide the interval $[0, \hat{t}^*]$ into n equal intervals of length $\delta t = \hat{t}^*/n$ for some large n.

2. Generate a random integer I from 1 to N uniformly.
3. For $i = 1, \ldots, n$, repeat

$$Y_i = Y_{i-1} + \mu(Y_{i-1})\delta t + \sigma(Y_{i-1})\sqrt{\delta t}\, Z_i,$$

where $Z_1, \ldots, Z_n \sim_{\text{i.i.d.}} \mathsf{N}(0, 1)$, and $Y_0 = X_I$.
4. Output Y_n as a random variable with approximate density (9).

Note that since we are only interested in the approximation of the statistical properties of Y_n, there are no gains to be had from using the more complex Milstein stochastic integration procedure [34].

6. Numerical experiments. In this section, we provide a simulation study of the diffusion estimator. In implementing Algorithm 2, there are a number of issues to consider. First, the numerical solution of the PDE (7) is a straightforward application of either finite difference or spectral methods [36]. A Matlab implementation using finite differences and the stiff ODE solver `ode15s.m` is available from the first author upon request. Second, we compute $\|Lg(\cdot; \hat{t}_2^*)\|^2$ in Algorithm 2 using the approximation

$$\|Lg(\cdot; t)\|^2 = \left\| \frac{\partial g}{\partial t}(\cdot; t) \right\|^2 \approx \|g(\cdot; t + \varepsilon) - g(\cdot; t)\|^2/\varepsilon^2, \qquad \varepsilon \ll 1,$$

where $g(\cdot; t)$ and $g(\cdot; t + \varepsilon)$ are the successive output of the numerical integration routine (`ode15s.m` in our case). Finally, we selected $\alpha = 1$ or $a(x) = p(x)$ in Algorithm 2 without using any clipping of the pilot estimate. For a small simulation study with $\alpha = 0$, see [5].

We would like to point out that simulation studies of existing variable-location scale estimators [27, 46, 51] are implemented assuming that the target p.d.f. f and any functionals of f are known *exactly* and no pilot estimation step is employed. In addition, in these simulation studies the bandwidth is chosen so that it is the global minimizer of the exact MISE. Since in practical applications the MISE and all functionals of f are not available, but have to be estimated, we proceed differently in our simulation study. We compare the estimator of Algorithm 2 with the Abramson's popular adaptive kernel density estimator [1]. The parameters ${}_*t$ and ${}_*t_2$ of the diffusion estimator are estimated using the new bandwidth selection procedure in Algorithm 1. The implementation of Abramson's estimator in the Stata language is given in [33]. Briefly, the estimator is given by

$$\hat{f}_{\mathsf{A}}(x) = \frac{1}{N\sqrt{t}\lambda_i} \sum_{i=1}^{N} \phi\left(\frac{x - X_i}{\sqrt{t}\lambda_i}\right),$$

where $\lambda_i^2 = G/\hat{f}(X_i; t_p)$, $G = (\prod_{i=1}^{N} \hat{f}(X_i; t_p))^{1/N}$, and the bandwidths $\sqrt{t}$ and $\sqrt{t_p}$ are computed using Least Squares Cross Validation (LSCV) [38].

Our criterion for the comparison is the numerical approximation to

$$\text{Ratio} = \frac{\|g(\cdot;\hat{t}^*) - f\|^2}{\|\hat{f}_{\mathsf{A}} - f\|^2},$$

that is, the ratio of the integrated squared error of the diffusion estimator to the integrated squared error of the alternative kernel density estimator.

Table 2, column 4 (ratio I) shows the average results over 10 independent trials for a number of different test cases. The second column displays the target density and the third column shows the sample size used for the experiments. In the table $\mathsf{N}(\mu, \sigma^2)$, denotes a Gaussian density with mean μ and variance σ^2. Most test problems are taken from [42]. For each test case, we conducted a simulation run with both a relatively small sample size and a relatively large sample size wherever possible. The table shows that, unlike the standard variable location-scale estimators [27, 51], the diffusion estimator does not require any clipping procedures in order to retain its good performance for large sample sizes.

TABLE 2

Results over 10 *independent simulation experiments. In all cases the domain was assumed to be* $\mathbb{R}$

Case	**Target density** $f(x)$	N	**Ratio I**	**Ratio II**
1	$\frac{1}{2}\mathsf{N}(0, (\frac{1}{10})^2) + \frac{1}{2}\mathsf{N}(5, 1)$	10^3	0.9	0.82
		10^5	0.23	0.48
2	$\frac{1}{2}\mathsf{N}(0, 1) + \sum_{k=0}^{4} \frac{1}{10}\mathsf{N}(\frac{k}{2} - 1, (\frac{1}{10})^2)$	10^3	0.65	0.99
		3×10^5	0.11	0.51
3	$\sum_{k=0}^{7} \frac{1}{8}\mathsf{N}(3((\frac{2}{3})^k - 1), (\frac{2}{3})^{2k})$	10^3	1.05	0.75
		10^5	0.15	0.45
4	$\frac{49}{100}\mathsf{N}(-1, (\frac{2}{3})^2) + \frac{49}{100}\mathsf{N}(1, (\frac{2}{3})^2) + \frac{1}{350}\sum_{k=0}^{6} \mathsf{N}(\frac{k-3}{2}, (\frac{1}{100})^2)$	10^3	0.94	0.63
		10^5	0.46	0.76
5	$\frac{2}{7}\sum_{k=0}^{2} \mathsf{N}(\frac{12k-15}{7}, (\frac{2}{7})^2) + \frac{1}{21}\sum_{k=8}^{10} \mathsf{N}(\frac{2k}{7}, (\frac{1}{21})^2)$	10^3	0.54	2.24
		10^5	0.12	0.84
6	$\frac{46}{100}\sum_{k=0}^{1} \mathsf{N}(2k - 1, (\frac{2}{3})^2) + \sum_{k=1}^{3} \frac{1}{300}\mathsf{N}(-\frac{k}{2}, (\frac{1}{100})^2)$	10^4	0.83	0.93
	$+ \sum_{k=1}^{3} \frac{7}{300}\mathsf{N}(\frac{k}{2}, (\frac{7}{100})^2)$	10^5	0.55	0.68
7	$\frac{1}{2}\mathsf{N}(-2, \frac{1}{4}) + \frac{1}{2}\mathsf{N}(2, \frac{1}{4})$	10^3	0.51	0.51
		10^5	0.41	0.89
8	$\frac{3}{4}\mathsf{N}(0, 1) + \frac{1}{4}\mathsf{N}(\frac{3}{2}, (\frac{1}{3})^2)$	10^3	0.59	0.53
		10^6	0.79	1.01
9	Log-Normal with $\mu = 0$ and $\sigma = 1$	10^3	0.17	0.85
		10^5	0.12	0.51
10	$\frac{1}{2}\mathsf{N}(0, 1) + \sum_{k=-2}^{2} \frac{2^{1-k}}{31}\mathsf{N}(k + \frac{1}{2}, (\frac{2^{-k}}{10})^2)$	10^3	0.88	0.98
		10^4	0.30	0.85

TABLE 3
Practical performance of the boundary bias correction of the diffusion estimator for the test cases: (1) *exponential distribution with mean equal to unity*; (2) *test cases* 1 *through* 8, *truncated to the interval* $(-\infty, 0]$

Test case	**Exp(1)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Ratio	0.52	0.38	0.74	0.25	0.70	0.38	0.74	0.56	0.46

Next, we compare the practical performance of the proposed diffusion estimator with the performance of higher-order kernel estimators. We consider the sinc kernel estimator defined as

$$\hat{f}_{\mathsf{sinc}}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{t}}K\left(\frac{x - X_i}{\sqrt{t}}\right), \qquad K(x) = \frac{\sin(x)}{\pi x},$$

where again $\sqrt{t}$ is selected using LSCV. Table 2, column 5 (ratio II) shows that the results are broadly similar and our method is favored in all cases except test case 5. Higher-order kernels do not yield proper density estimators, because the kernels take on negative values. Thus, an important advantage of our method and all second order kernel methods is that they provide nonnegative density estimators. As pointed out in [53], the good asymptotic performance of higher-order kernels is not guaranteed to carry over to finite sample sizes in practice. Our results confirm this observation.

In addition, we make a comparison with the novel polynomial boundary correction method of Hall and Park [20]. The results are given in Table 3, where we use some of the test cases defined in Table 1, truncated to the interval $(-\infty, 0]$. Table 3 shows that for finite sample sizes the practical performance of our approach is competitive. We now give the implementation details. Let β be the point of truncation from above, which is assumed to be known in advance. Then, the Hall and Park estimator is

$$\hat{f}_\alpha(x;t) = \frac{1}{N\int_{-\infty}^{\beta}\phi((x-y)/h)\,dy}\sum_{i=1}^{N}\phi\left(\frac{x - X_i + \alpha(x)}{\sqrt{t}}\right), \qquad x \le \beta, \tag{34}$$

where $\alpha(x) = t\frac{\hat{f}_0'(x)}{\hat{f}_0(x)}\rho(\frac{x-a}{h})$; $\hat{f}_0(x)$ is equivalent to $\hat{f}_\alpha(x)$ when $\alpha(x) \equiv 0$, and $\hat{f}_0'(x)$ is an estimator of $f'(x)$; $\rho(u) = \frac{1}{\phi(u)}\int_{-\infty}^{u} v\phi(v)\,dv$. We use LSCV to select a suitable bandwidth $\sqrt{t}$. The denominator in (34) adjusts for the deficit of probability mass in the neighborhood of the end-point, but note that theoretically (34) does not integrate to unity and therefore random variable generation from (34) is not straightforward. In addition, our estimator more easily handles the case with two end-points. On the positive side, Hall and Park [20] note that their estimator

preserves positivity and has excellent asymptotic properties, which is an advantage over many other boundary kernels.

Finally, we give a two-dimensional density estimation example, which to the best of our knowledge cannot be handled satisfactorily by existing methods [19, 31] due to the boundary bias effects. The two-dimensional version of equation (2) is

$$\frac{\partial \hat{f}}{\partial t}(\mathbf{x}; t) = \frac{1}{2}\left(\frac{\partial^2 \hat{f}}{\partial x_1^2}(\mathbf{x}; t) + \frac{\partial^2 \hat{f}}{\partial x_2^2}(\mathbf{x}; t)\right) \qquad \forall t > 0, \mathbf{x} \in \mathscr{X},$$

$$\hat{f}(\mathbf{x}; 0) = \Delta(\mathbf{x}),$$

$$\mathbf{n} \cdot \nabla \hat{f}(\mathbf{x}; t) = 0 \qquad \forall t > 0,$$

where $\mathbf{x} = (x_1, x_2)$ belongs to the set $\mathscr{X} \subseteq \mathbb{R}^2$, the initial condition $\Delta(\mathbf{x})$ is the empirical density of the data, and in the Neumann boundary condition $\mathbf{n}$ denotes the unit outward normal to the boundary $\partial\mathscr{X}$ at $\mathbf{x}$. The particular example which we consider is the density estimation of 600 uniformly distributed points on the domain $\mathscr{X} = \{\mathbf{x} : x_1^2 + (4x_2)^2 \leq 4\}$. We assume that the domain of the data $\mathscr{X}$ is known prior to the estimation. Figure 5 shows $\hat{f}(\mathbf{x}; \hat{t}^*)$ on $\mathscr{X} = \{\mathbf{x} : x_1^2 + (4x_2)^2 \leq 4\}$, that is, it shows the numerical solution of the two-dimensional PDE at time $\hat{t}^* = 0.13$ on the set $\mathscr{X}$. The bandwidth was determined using the bandwidth selection procedure described in Appendix E. We emphasize the satisfactory way in which the p.d.f. $\hat{f}(\mathbf{x}; \hat{t}^*)$ handles any boundary bias problems. It appears that currently existing methods [19, 22, 31, 32] cannot handle such two-dimensional (boundary) density estimation problems either because the geometry of the set $\mathscr{X}$ is too complex, or because the resulting estimator is not a bona-fide p.d.f.

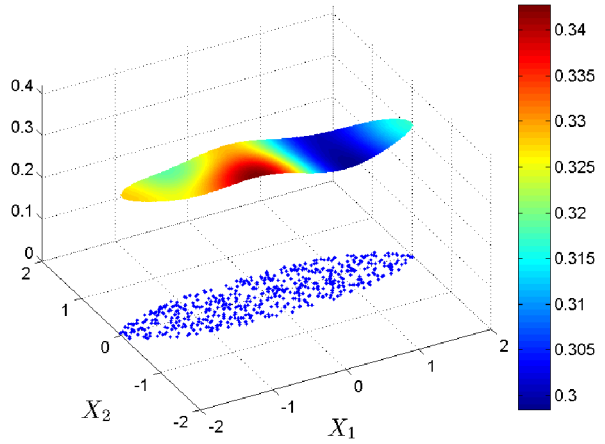

FIG. 5. *A two-dimensional example with* 600 *points generated uniformly within an ellipse.*

7. Conclusions and future research. We have presented a new kernel density estimator based on a linear diffusion process. The key idea is to construct an adaptive kernel by considering the most general linear diffusion with its stationary density equal to a pilot density estimate. The resulting diffusion estimator unifies many of the existing ideas about adaptive smoothing. In addition, the estimator is consistent at boundaries. Numerical experiments suggest good practical performance. As future research, the proposed estimator can be extended in a number of ways. First, we can construct kernel density estimators based on Lévy processes, which will have the diffusion estimator as a special case. The kernels constructed via a Lévy process could be tailored for data for which smoothing with the Gaussian kernel density estimator or diffusion estimator is not optimal. Such cases arise when the data is a sample from a heavy-tailed distribution. Second, more subtle and interesting smoothing models can be constructed by considering nonlinear parabolic PDEs. One such candidate is the quasilinear parabolic PDE with diffusivity that depends on the density exponentially:

$$\frac{\partial}{\partial t}g(x;t) = \frac{\partial}{\partial x}\left(e^{-\alpha g(x;t)}\frac{\partial}{\partial x}g(x;t)\right), \qquad \alpha > 0.$$

Another viable model is the semilinear parabolic PDE

$$\frac{\partial}{\partial t}\left(e^{u(x;t)}\right) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(x;t),$$

where $u(x;t) = \log(g(x;t))$ is the logarithm of the density estimator. The Cauchy density $\frac{t}{\pi(x^2+t^2)}$ is a particular solution and thus the model could be useful for smoothing heavy-tailed data. All such nonlinear models will provide adaptive smoothing without the need for a pilot run, but at the cost of increased model complexity.

APPENDIX A: GAUSSIAN KERNEL DENSITY ESTIMATOR PROPERTIES

In this appendix, we present the technical details for the proofs of the properties of the diffusion estimator. In addition, we include a description of our plug-in rule in two dimensions.

We use $\|\cdot\|$ to denote the Euclidean norm on $\mathbb{R}$.

THEOREM 3. *Let $t = t_N$ be such that $\lim_{N\to\infty} t_N = 0$ and $\lim_{N\to\infty} N\sqrt{t_N} = \infty$. Assume that f'' is a continuous square-integrable function. The integrated squared bias and integrated variance of the Gaussian kernel density estimator* (1) *have asymptotic behavior*

$$\|\mathbb{E}_f[\hat{f}(\cdot;t)] - f\|^2 = \tfrac{1}{4}t^2\|f''\|^2 + o(t^2), \qquad N\to\infty, \tag{35}$$

and

$$\int \mathrm{Var}_f[\hat{f}(x;t)]\,dx = \frac{1}{2N\sqrt{\pi t}} + o\big((N\sqrt{t})^{-1}\big), \qquad N\to\infty, \tag{36}$$

respectively. The first-order asymptotic approximation of MISE, denoted AMISE, is thus given by

$$\text{AMISE}\{\hat{f}\}(t) = \frac{1}{4}t^2\|f''\|^2 + \frac{1}{2N\sqrt{\pi t}}. \tag{37}$$

The asymptotically optimal value of t is the minimizer of the AMISE

$$*t = \left(\frac{1}{2N\sqrt{\pi}\|f''\|^2}\right)^{2/5}, \tag{38}$$

giving the minimum value

$$\text{AMISE}\{\hat{f}\}(*t) = N^{-4/5}\frac{5\|f''\|^{2/5}}{4^{7/5}\pi^{2/5}}. \tag{39}$$

For a simple proof, see [53].

APPENDIX B: PROOF OF LEMMA 1

We seek to establish the behavior of the solution of (11) and (10) as $t \downarrow 0$. We use the Wentzel–Kramers–Brillouin–Jeffreys (WKBJ) method described in [2, 8, 29, 43]. In the WKBJ method, we look for an asymptotic expansion of the form

$$\kappa(x, y; t) \sim e^{-1/(2t)s^2(x,y)} \sum_{m=0}^{\infty} t^{m-1/2} C_m(x, y), \qquad t \downarrow 0, \tag{40}$$

where $\{C_m(x, y)\}$ and $s(x, y)$ are unknown functions. To determine $s(x, y)$ and $\{C_m(x, y)\}$, we substitute the expansion into (10) and, after canceling the exponential term, equate coefficients of like powers of t. This matching of the powers of t leads to solvable ODEs, which determine the unknown functions. Eliminating the leading order $O(t^{-5/2})$ term gives the ODE for s

$$a(x)\left[\frac{\partial}{\partial x}s(x, y)\right]^2 - p(x) = 0. \tag{41}$$

Setting the next highest order $O(t^{-3/2})$ term in the expansion to zero gives the ODE

$$\begin{aligned} 0 = {} & 2a(x)s(x, y)\frac{\partial s}{\partial x}\frac{dp}{dx}p(x)C_0(x, y) - 2a(x)s(x, y)\frac{\partial s}{\partial x}p^2(x)\frac{\partial C_0}{\partial x} \\ & + p^3(x)C_0(x, y) + s^2(x, y)p^3(x)C_1(x, y) \\ & - \frac{da}{dx}p^2(x)s(x, y)\frac{\partial s}{\partial x}C_0(x, y) \\ & + a(x)s^2(x, y)\left(\frac{\partial s}{\partial x}\right)^2 p^2(x)C_1(x, y) - a(x)\left(\frac{\partial s}{\partial x}\right)^2 p^2(x)C_0(x, y) \\ & - a(x)s(x, y)\frac{\partial^2 s}{\partial x^2}p^2(x)C_0(x, y). \end{aligned} \tag{42}$$

To determine a unique solution to (41), we impose the condition $s(x,x)=0$, which is necessary, but not sufficient, to ensure that $\lim_{t\downarrow 0}\kappa(x,y;t)=\delta(x-y)$. This gives the solution

$$s(x,y)=\int_y^x \sqrt{\frac{p(s)}{a(s)}}\,ds.$$

Substituting this solution into (42) and simplifying gives an equation without $C_1(x,y)$,

$$C_0(x,y)p(x)\frac{da}{dx}+4a(x)p(x)\frac{\partial C_0}{\partial x}-3C_0(x,y)\frac{dp}{dx}a(x)=0, \tag{43}$$

whence we have the general solution $C_0(x,y)=h(y)p^{3/4}(x)a^{-1/4}(x)$ for some as yet unknown function of y, $h(y)$. To determine $h(y)$, we require that the kernel $\widetilde{\kappa}(x,y;t)$ satisfies the detailed balance equation (15). This ensures that $\widetilde{\kappa}(x,y;t)$ also satisfies (11). It follows that $C_0(x,y)$ has to satisfy $p(y)C_0(x,y)=p(x)C_0(y,x)$, which after rearranging gives

$$h(x)(a(x)p(x))^{1/4}=h(y)(a(y)p(y))^{1/4}.$$

A separation of variables argument now gives $h(y)(a(y)p(y))^{1/4}=\text{const.}$, and hence

$$C_0(x,y)=\text{const.}(a(y)p(y))^{-1/4}p^{3/4}(x)a^{-1/4}(x).$$

We still need to determine the arbitrary constant. The constant is chosen so that

$$\lim_{t\downarrow 0}\int_{-\infty}^{\infty}\widetilde{\kappa}(x,y;t)\,dx=1,$$

which ensures that $\lim_{t\downarrow 0}\widetilde{\kappa}(x,y;t)=\delta(x-y)$. This final condition yields

$$C_0(x,y)=\frac{p(x)}{\sqrt{2\pi}(a(y)p(y)a(x)p(x))^{1/4}},$$

and hence

$$\widetilde{\kappa}(x,y;t)=\frac{p(x)}{\sqrt{2\pi t}[p(x)a(x)a(y)p(y)]^{1/4}}\exp\left\{-\frac{1}{2t}\left[\int_y^x\sqrt{\frac{p(s)}{a(s)}}\,ds\right]^2\right\}.$$

REMARK 6. Matching higher powers of t gives first order linear ODEs for the rest of the unknown functions $\{C_m(x,y), m\geq 1\}$. The ODE for each $C_m(x,y), m=1,2,3,\ldots$ is

$$as'(C_m/p)'+\left(\frac{(as')'}{2p}+(m-1/2)\right)C_m=\big(a(C_{m-1}/p)'\big)',\qquad C_m(y,y)=0,$$

where all derivatives apply to the variable x and y is treated as a constant. Thus, in principle, all functions $\{C_m(x,y)\}$ can be uniquely determined.

It can be shown (see [8]) that the expansion (40) is valid under the conditions that a, p and all their derivatives are bounded from above, and $p(x) \geq p_0 > 0$, $a(x) \geq a_0 > 0$. Here, we only establish the validity of the leading order approximation $\widetilde{\kappa}$ under the milder conditions (17). We do not attempt to prove the validity of the higher order terms in (40) under the weaker conditions. The proof of the following lemma uses arguments similar to the ones given in [8].

LEMMA 2. *Let $a(x)$ and $p(x)$ satisfy conditions (17). Then, for all $t \in (0, t_0]$, where $t_0 > 0$ is some constant independent of x and y, there holds*

$$|\kappa(x, y; t) - \widetilde{\kappa}(x, y; t)| \leq \text{const.}\, C_0(x, y) t^{1/4} e^{-s^2(x,y)/(2t)} \qquad \forall x, y.$$

To prove the lemma, we first begin by proving the following auxiliary results.

PROPOSITION 4. *Define*

$$\ell(z) = \ell(z; x, y, t, \tau) = \frac{s^2(x, z)}{2(t - \tau)} + \frac{s^2(z, y)}{2\tau}.$$

Then for $\tau \in (0, t)$, we have

$$\ell(z) \geq \frac{s^2(x, y)}{2t}.$$

Moreover, there exists a unique $z_0 = z_0(x, y, t, \tau)$ for which $\ell(z_0) = \frac{s^2(x,y)}{2t}$, and $\ell(z)$ is increasing for $z > z_0$ and decreasing for $z < z_0$.

PROOF. We have

$$\ell(z) = \frac{1}{2(t - \tau)} \left(\int_z^x \sigma^{-1}(s)\, ds \right)^2 + \frac{1}{2\tau} \left(\int_y^z \sigma^{-1}(s)\, ds \right)^2,$$

and hence

$$\ell'(z) = \frac{-\sigma^{-1}(z)}{t - \tau} \int_z^x \sigma^{-1}(s)\, ds + \frac{\sigma^{-1}(z)}{\tau} \int_y^z \sigma^{-1}(s)\, ds. \tag{44}$$

For $x \neq y$, $\ell'(y) > 0$, $\ell'(x) < 0$, and therefore by the continuity of ℓ', there exists $z_0 \in (x, y) : \ell'(z_0) = 0$. For $x = y$, set $z_0 = x$. Setting $z = z_0$ in (44),

$$\frac{1}{t - \tau} \int_{z_0}^x \sigma^{-1}(s)\, ds = \frac{1}{\tau} \int_y^{z_0} \sigma^{-1}(s)\, ds. \tag{45}$$

Therefore, $\int_{z_0}^x \sigma^{-1}(s)\, ds = \frac{t-\tau}{\tau} \int_y^{z_0} \sigma^{-1}(s)\, ds$ and adding $\int_y^{z_0} \sigma^{-1}(s)\, ds$ to both sides we obtain

$$\int_y^x \sigma^{-1}(s)\, ds = \frac{t}{\tau} \int_y^{z_0} \sigma^{-1}(s)\, ds,$$

from which we see that (45) is also equal to $\frac{1}{t}\int_y^x \sigma^{-1}(s)\,ds$. Hence, by substitution $\ell(z_0) = \frac{1}{2t}(\int_y^x \sigma^{-1}(s)\,ds)^2$, as required. Finally, note that if $F(z) = \ell(z) - \frac{t}{2\tau(t-\tau)}(\int_{z_0}^z \sigma^{-1}(s)\,ds)^2$, then $F'(z) = 0$ for all z. Hence, $F(z) = F(z_0) = \ell(z_0)$ and

$$\ell(z) = \ell(z_0) + \frac{t}{2\tau(t-\tau)}\left(\int_{z_0}^z \sigma^{-1}(s)\,ds\right)^2. \tag{46}$$

As a consequence of Proposition 4, we have the following result. □

PROPOSITION 5. *Assuming* $\lim_{z\to\pm\infty}\int_{z_0}^z \sigma^{-1}(s)\,ds = \pm\infty$, *we have the following equality*:

$$\begin{aligned}\int_0^t \sqrt{\int_{-\infty}^{\infty}\left(\frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}}\frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\sigma(z)}\sqrt{\tau}}\right)^2 dz}\,d\tau \\ = 2\pi^{-1/4}t^{1/4}\Gamma^2(3/4)e^{-s^2(x,y)/(2t)} \\ = c_2 t^{1/4}e^{-s^2(x,y)/(2t)},\end{aligned}$$

where c_2 *is a constant* [*indeed* $c_2 = 2\pi^{-1/4}\Gamma^2(3/4)$].

PROOF. We have

$$\begin{aligned}&\int_{-\infty}^{\infty}\frac{e^{-s^2(x,z)/(t-\tau)}}{t-\tau}\frac{e^{-s^2(z,y)/\tau}}{\sigma(z)\tau}\,dz \\ &\quad= \frac{1}{(t-\tau)\tau}e^{-2\ell(z_0)}\int_{-\infty}^{\infty}\sigma^{-1}(z)e^{-(\int_{z_0}^z \sigma^{-1}(s)\,ds)^2/(\tau(t-\tau)/t)}\,dz \\ &\quad= \frac{1}{\sqrt{t(t-\tau)\tau}}e^{-2\ell(z_0)}\int_{-\infty}^{\infty}e^{-v^2}\,dv,\end{aligned}$$

with the change of variable $v(z) = \frac{1}{\sqrt{\tau(t-\tau)/t}}\int_{z_0}^z \sigma^{-1}(s)\,ds$. Then the result follows from the fact that $\int_0^t(\tau(t-\tau))^{-1/4}\,d\tau = 2\pi^{-1/2}t^{1/2}\Gamma^2(3/4)$.

Given these two auxiliary results, we proceed with the proof of Lemma 2. Writing

$$\kappa^*(x,y;t) = \frac{\partial}{\partial t}\widetilde{\kappa}(x,y;t) - L\widetilde{\kappa}(x,y;t) = -\frac{e^{-s^2(x,y)/(2t)}}{\sqrt{t}}LC_0(x,y),$$

we define inductively the following sequence of function $\{\rho_j\}$, starting with $\rho_0 = 0$:

$$\rho_{j+1}(x,y;t) = -\kappa^*(x,y;t) - \int_0^t\int_{-\infty}^{\infty}\kappa^*(x,z;t-\tau)\rho_j(z,y;\tau)\,dz\,d\tau,$$

$$j = 1, 2, \ldots.$$

Note in particular that $\rho_1 = -\kappa^*$. We will show that there exists a limit of $\{\rho_j\}$. We begin by proving via induction that for $j \geq 1$, $x, y \in \mathbb{R}$, $t \in (0, t_0]$, where

$$t_0 = \min\left\{\left(\frac{\sqrt{2\pi}}{2c_1c_2}\right)^{4/3}, 1\right\},$$

there holds

$$|\rho_{j+1}(x, y, t) - \rho_j(x, y, t)| \leq \frac{c_3}{2^j}|LC_0(x, y)|t^{1/4}e^{-s^2(x,y)/(2t)}, \tag{47}$$

where $c_3 = 2c_1c_2/\sqrt{2\pi}$. First, we calculate for $j = 1$

$$\rho_2(x, y, t) = -\kappa^*(x, y, t) + \int_0^t \int_{-\infty}^{\infty} \kappa^*(x, z, t-\tau)\kappa^*(z, y, \tau)\, dz\, d\tau.$$

Therefore, we have the following bound:

$$\begin{aligned}
&|\rho_2(x, y, t) - \rho_1(x, y, t)| \\
&\quad \leq \int_0^t \int_{-\infty}^{\infty} |\kappa^*(x, z, t-\tau)\kappa^*(z, y, \tau)|\, dz\, d\tau \\
&\quad = \int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} \frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\tau}} |LC_0(x, z)LC_0(z, y)|\, dz\, d\tau \\
&\quad = \int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} \frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\sigma(z)\tau}} \\
&\qquad\qquad \times \sqrt{\sigma(z)}|LC_0(x, y)| \frac{|Lq(z)|}{\sqrt{2\pi}(a(z)p(z))^{1/4}}\, dz\, d\tau \\
&\quad = \frac{1}{\sqrt{2\pi}}|LC_0(x, y)| \int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} \\
&\qquad\qquad \times \frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\sigma(z)\tau}} \frac{|Lq(z)|}{q(z)}\, dz\, d\tau \\
&\quad \leq \frac{1}{\sqrt{2\pi}}|LC_0(x, y)|c_1c_2t^{1/4}e^{-s^2(x,y)/(2t)},
\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality, Proposition 5 and assumption (17). We thus have

$$|\rho_2(x, y, t) - \rho_1(x, y, t)| \leq \frac{c_3}{2}|LC_0(x, y)|t^{1/4}e^{-s^2(x,y)/(2t)}.$$

Next, assume the induction statement is true for $2, 3, \ldots, j-1$. Then

$$\begin{aligned}
&|\rho_{j+1}(x, y, t) - \rho_j(x, y, t)| \\
&\quad \leq \int_0^t \int_{-\infty}^{\infty} |\kappa^*(x, z, t-\tau)||\rho_j(z, y, \tau) - \rho_{j-1}(z, y, \tau)|\, dz\, d\tau
\end{aligned}$$

$$\le \int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} |LC_0(x,z)| \frac{c_3}{2^{j-1}} |LC_0(z,y)| \times \tau^{1/4} e^{-s^2(z,y)/(2\tau)}\, dz\, d\tau$$

$$\le \frac{c_3}{2^{j-1}} |LC_0(x,y)| \int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} \frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\sigma(z)\tau}} \times \tau^{3/4} \frac{|Lq(z)|}{\sqrt{2\pi} q(z)}\, dz\, d\tau$$

$$\le \frac{c_3}{2^{j-1}} |LC_0(x,y)| t^{1/4} e^{-s^2(x,y)/(2t)} t_0^{3/4} \frac{c_1 c_2}{\sqrt{2\pi}}.$$

The last line follows from the Cauchy–Schwarz inequality and the fact that $\tau^{3/4} \le t^{3/4} \le t_0^{3/4}$. Since $t_0^{3/4} \frac{c_1 c_2}{\sqrt{2\pi}} \le \frac{1}{2}$, we obtain

$$|\rho_{j+1}(x,y,t) - \rho_j(x,y,t)| \le \frac{c_3}{2^j} |LC_0(x,y)| t^{1/4} e^{-s^2(x,y)/(2t)}.$$

This establishes (47). Next, we have the bound for all $j \ge 1$:

$$\begin{aligned} |\rho_j(x,y,t)| &\le |\rho_1(x,y,t)| + \sum_{j=1}^{\infty} \frac{c_3}{2^j} |LC_0(x,y)| t^{1/4} e^{-s^2(x,y)/(2t)} \\ &\le |LC_0(x,y)| \left(\frac{1}{\sqrt{t}} + c_3 t^{1/4} \right) e^{-s^2(x,y)/(2t)} \\ &\le |LC_0(x,y)| \frac{2}{\sqrt{t}} e^{-s^2(x,y)/(2t)}. \end{aligned} \tag{48}$$

In the light of (48) and (47), the pointwise limit

$$\rho(x,y,t) = \lim_{j \to \infty} \rho_j(x,y,t)$$

exists on $\mathbb{R} \times \mathbb{R} \times (0, t_0)$. In addition, $\rho(x,y,t)$ satisfies the limiting equation

$$0 = \kappa^*(x,y,t) + \rho(x,y,t) + \int_0^t \int_{-\infty}^{\infty} \kappa^*(x,z,t-\tau) \rho(z,y,\tau)\, dz\, d\tau,$$

and indeed

$$\kappa(x,y;t) - \widetilde{\kappa}(x,y;t) = \int_0^t \int_{-\infty}^{\infty} \widetilde{\kappa}(x,z,t-\tau) \rho(z,y,\tau)\, dz\, d\tau. \tag{49}$$

In order to see this, we can apply directly the arguments of Section 5 of [8] in the case $N = 0$; see also Section 1.3 of [14]. Hence, we can take the limit in (48) to conclude

$$|\rho(x,y,t)| \le 2 |LC_0(x,y)| t^{-1/2} e^{-s^2(x,y)/(2t)} \tag{50}$$

for $t \in (0, t_0]$. The claim of the lemma then follows from

$$
\begin{aligned}
&|\kappa(x, y; t) - \widetilde{\kappa}(x, y; t)| \\
&\quad \le \int_0^t \int_{-\infty}^{\infty} \widetilde{\kappa}(x, z, t-\tau)|\rho(z, y, \tau)|\, dz\, d\tau \\
&\quad \le 2\int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} C_0(x, z) \frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\tau}} |LC_0(z, y)|\, dz\, d\tau \\
&\quad \le \frac{2}{\sqrt{2\pi}} C_0(x, y) \int_0^t \int_{-\infty}^{\infty} \frac{e^{-s^2(x,z)/(2(t-\tau))}}{\sqrt{t-\tau}} \frac{e^{-s^2(z,y)/(2\tau)}}{\sqrt{\sigma(z)\tau}} \frac{|Lq(z)|}{q(z)}\, dz\, d\tau \\
&\quad \le 2C_0(x, y) t^{1/4} e^{-s^2(x,y)/(2t)} \frac{c_1 c_2}{\sqrt{2\pi}} = c_3 C_0(x, y) t^{1/4} e^{-s^2(x,y)/(2t)}. \qquad \square
\end{aligned}
$$

APPENDIX C: PROOF OF THEOREM 1

Note that (18) is given by $\int_{-\infty}^{\infty} \kappa(x, y; t) f(y)\, dy - f(x)$, and from (11) we have

$$
\begin{aligned}
\frac{\partial}{\partial t} g(x; t) &= \int_{\mathscr{X}} f(y) L^* \kappa(x, y; t)\, dy \\
&= -\frac{1}{2} \frac{d}{dy} \left(\frac{f(y)}{p(y)} \right) a(y) \kappa(x, y; t) \Big|_{y \in \partial \mathscr{X}} + \int_{\mathscr{X}} \kappa(y, x; t) Lf(x)\, dx.
\end{aligned}
$$

Given that $\mathscr{X} \equiv \mathbb{R}$, Lemma 1 gives $\kappa(x, y; t)|_{y \in \partial \mathscr{X}} \sim \widetilde{\kappa}(x, y; t)|_{y=-\infty}^{y=\infty}$, $t \downarrow 0$. The last term is zero since for fixed x,

$$
\lim_{y \to \pm\infty} \left[\int_y^x \sqrt{\frac{p(s)}{a(s)}}\, ds \right]^2 = \infty,
$$

and hence $\lim_{y \to \pm\infty} \widetilde{\kappa}(x, y; t) = 0$. We have

$$
g(x; t) = g(x; 0) + t \frac{\partial}{\partial t} g(x; t) \Big|_{t=0} + O(t^2),
$$

because $g(x; t), t > 0$ is smooth (see, e.g., Theorem IV $\cdot$ 10 $\cdot$ 1 in [35]). Therefore,

$$
g(x; t) = f(x) + tLf(x) + O(t^2),
$$

and (18) and (19) follow. We now proceed to demonstrate (20). First, the second moment has the behavior

$$
\begin{aligned}
&\mathbb{E}_f[\kappa^2(x, Y; t)] \\
&\quad = \int_{\mathscr{X}} f(y) \kappa^2(x, y; t)\, dy \sim \int_{\mathscr{X}} f(y) \widetilde{\kappa}^2(x, y; t)\, dy \\
&\quad \sim \frac{p^2(x)}{2\pi t \sqrt{p(x)a(x)}} \int_{-\infty}^{\infty} \frac{f(y)}{\sqrt{p(y)a(y)}} e^{-1/2[\sqrt{2/t} \int_x^y \sqrt{p(s)/a(s)}\, ds]^2}\, dy.
\end{aligned}
$$

We can simplify the last expression by the change of variable $u = \sqrt{\frac{2}{t}} \times \int_x^y \sqrt{\frac{p(s)}{a(s)}}\,ds$. This gives

$$\frac{p^2(x)}{2\pi\sqrt{2t}\sqrt{p(x)a(x)}}\int_{-\infty}^{\infty}\frac{f(y(u,t))}{p(y(u,t))}e^{-u^2/2}\,du,$$

where $y(u,t) = y(u,0) + \sqrt{t}\frac{\partial y}{\partial\sqrt{t}}|_{t=0} + O(t) = x + u\sqrt{\frac{ta(x)}{2p(x)}} + O(t)$ is a Taylor expansion of $y(u,t)$ at $\sqrt{t} = 0$. Therefore, $\frac{f(y(u,t))}{p(y(u,t))} \sim \frac{f(x)}{p(x)}$ as $t \downarrow 0$, and

$$\frac{p^2(x)}{2\pi\sqrt{2t}\sqrt{p(x)a(x)}}\int_{-\infty}^{\infty}\frac{f(y(u,t))}{p(y(u,t))}e^{-u^2/2}\,du \sim \frac{1}{2\sqrt{\pi t}}f(x)\sqrt{\frac{p(x)}{a(x)}}, \qquad t \downarrow 0.$$

Hence, from (9) we have

$$\begin{aligned}\operatorname{Var}_f[g(x;t)] &= \frac{1}{N}\mathbb{E}_f[\kappa^2(x,Y;t)] - \frac{1}{N}\mathbb{E}_f[\kappa(x,Y;t)]^2\\ &\sim \frac{f(x)}{2N\sqrt{\pi t}\sigma(x)}, \qquad t \downarrow 0,\end{aligned}$$

from which (21) and (20) follow.

APPENDIX D: CONSISTENCY AT BOUNDARY

As in [53], we consider the case where the support of f is $[0,\infty]$. The consistency of the estimator near $x = 0$ is analyzed by considering the pointwise bias of estimator (9) at a point x_N such that x_N is $O(\sqrt{t_N})$ away from the boundary, that is, x_N is approaching the boundary at the same rate at which the bandwidth is approaching 0. We then have the following result, which shows that the diffusion estimator (9), and hence its special case (3), is consistent at the boundaries.

PROPOSITION 6. *Let $\mathscr{X} \equiv [0,\infty]$, and assume that $x = x_N = \alpha\sqrt{t_N}$ for some constant $\alpha \in [0,1]$, where $\lim_{N\to\infty} t_N = 0$ and $\lim_{N\to\infty} N\sqrt{t_N} = \infty$. Then for the diffusion estimator (9) we have*

$$\mathbb{E}_f g(x_N;t) = f(x_N) + O\big(\sqrt{t_N}\big), \qquad N \to \infty.$$

Hence, the diffusion estimator (9) is consistent at the boundaries.

PROOF. First, we differentiate both sides of $\mathbb{E}_f g(x;t) = \int_0^1 f(y)\kappa(x;y;t)\,dy$ with respect to t and use (11) to obtain

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}_f g(x;t) &= \int_0^\infty f(y)\frac{\partial}{\partial t}\kappa(x;y;t)\,dy\\ &= \int_0^\infty f(y)L^*\kappa(x;y;t)\,dy\\ &= -\frac{1}{2}\left(\frac{f(y)}{p(y)}\right)'a(y)\kappa(x;y;t)\Big|_{y=0}^{y=\infty} + \int_0^\infty \kappa(x;y;t)Lf(y)\,dy.\end{aligned}$$

Second, we show that $\kappa(\alpha\sqrt{t_N}; 0; t_N) = O(t^{-1/2})$ and $\lim_{y\to\infty}\kappa(\alpha\sqrt{t_N}; y; t_N) = o(1)$, and $\int_0^1 \kappa(x; y; t_N)Lf(y)\,dy = O(1)$ as $N \to \infty$. To this end, we consider the small bandwidth behavior of κ. It is easy to verify using Lemma 1 that the *boundary kernel*

$$\kappa_B(x, y; t) = \widetilde{\kappa}(x, y; t) + \widetilde{\kappa}(x, -y; t)$$

satisfies

$$\frac{\partial}{\partial t}\kappa_B(x, y; t) = L^*\kappa_B(x, y; t) + O\big(e^{-s^2(x,y)/(2t)}t^{-1/2}\big), \qquad t \downarrow 0,$$

on $x, y \in \mathbb{R}$ with initial condition $\kappa_B(x, y; 0) = \delta(x - y)$. In addition, the boundary kernel satisfies the condition $\frac{\partial}{\partial y}\kappa_B(x, y; t)|_{y=0} = 0$, and therefore κ_B describes the small bandwidth asymptotics of the solution of the PDE (11) on the domain $x, y \in [0, \infty)$ with boundary condition $\frac{\partial}{\partial y}\kappa(x, y; t)|_{y=0} = 0$. Hence, we have

$$\kappa\big(\alpha\sqrt{t}; 0; t\big) \sim \kappa_B\big(\alpha\sqrt{t}; 0; t\big) = \text{const.}\, t^{-1/2}e^{O(\sqrt{t})}, \qquad t \downarrow 0,$$

and

$$\lim_{y\to\infty}\kappa_B\big(\alpha\sqrt{t}; y; t\big) = 0, \qquad t > 0.$$

Therefore,

$$\frac{\partial}{\partial t}\mathbb{E}_f g(x_N; t_N) = o(1) - O(t_N^{-1/2}), \qquad N \to \infty,$$

or

$$\frac{\mathbb{E}_f g(x_N; t_N) - \mathbb{E}_f g(x_N; 0)}{t_N} + O(t_N) = O(t_N^{-1/2}) + O(1), \qquad N \to \infty,$$

which, after rearranging, gives

$$\mathbb{E}_f g(x_N; t_N) = f(x_N) + O\big(\sqrt{t_N}\big), \qquad N \to \infty. \qquad \square$$

APPENDIX E: BANDWIDTH SELECTION IN HIGHER DIMENSIONS

Algorithm 1 can be extended to two dimensions for the estimation of a p.d.f. $f(\mathbf{x})$ on $\mathbb{R}^2$. Assuming a Gaussian kernel

$$\phi(\mathbf{x}, \mathbf{y}; t) = \frac{1}{2\pi t}e^{-(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})/(2t)},$$

where $\mathbf{x} = [x_1, x_2]^T$ and $\mathbf{y} = [y_1, y_2]^T$, the asymptotically optimal squared bandwidth is given by ([53], page 99)

$$t^* = \big(2\pi N(\psi_{0,2} + \psi_{2,0} + 2\psi_{1,1})\big)^{-1/3},$$

where

$$\begin{aligned}\psi_{i,j} &= (-1)^{i+j}\int_{\mathbb{R}^2} f(\mathbf{x})\,\frac{\partial^{2(i+j)}}{\partial x_1^{2i}\,\partial x_2^{2j}} f(\mathbf{x})\,d\mathbf{x}, \qquad i,j\in\mathbb{N}^+,\\ &= \int\left(\frac{\partial^{(i+j)}}{\partial x_1^{i}\,\partial x_2^{j}} f(\mathbf{x})\right)^2 d\mathbf{x}.\end{aligned} \tag{51}$$

Note that our definition of ψ differs slightly from the definition of ψ in [53]. Here the partial derivatives under the integral sign are applied $2(i+j)$ times, while in [53] they are applied $(i+j)$ times. Similar to the one-dimensional case, there are two viable plug-in estimators for $\psi_{i,j}$. The first one is derived from the first line of (51):

$$\widetilde{\psi}_{i,j} = \frac{(-1)^{i+j}}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\frac{\partial^{2(i+j)}}{\partial x_1^{2i}\,\partial x_2^{2j}}\phi(\mathbf{X}_m,\mathbf{X}_k;t_{i,j}), \tag{52}$$

and the second one is derived from the second line of (51):

$$\begin{aligned}\widehat{\psi}_{i,j} &= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\int\frac{\partial^{(i+j)}}{\partial x_1^{i}\,\partial x_2^{j}}\phi(\mathbf{x},\mathbf{X}_m;t_{i,j})\,\frac{\partial^{(i+j)}}{\partial x_1^{i}\,\partial x_2^{j}}\phi(\mathbf{x},\mathbf{X}_k;t_{i,j})\,d\mathbf{x}\\ &= \frac{(-1)^{i+j}}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{N}\frac{\partial^{2(i+j)}}{\partial x_1^{2i}\,\partial x_2^{2j}}\phi(\mathbf{X}_m,\mathbf{X}_k;2t_{i,j}).\end{aligned} \tag{53}$$

The asymptotic expansion of the squared bias of estimator $\widetilde{\psi}_{i,j}$ is given by ([53], page 113)

$$\begin{aligned}&(\mathbb{E}_f[\widetilde{\psi}_{i,j}]-\psi_{i,j})^2\\ &\quad\sim\left(\frac{q(i)q(j)}{N t_{i,j}^{i+j+1}}+\frac{t_{i,j}}{2}(\psi_{i+1,j}+\psi_{i,j+1})\right)^2, \qquad N\to\infty,\end{aligned} \tag{54}$$

where

$$q(j)=\begin{cases}(-1)^j\dfrac{1\times 3\times 5\times\cdots\times(2j-1)}{\sqrt{2\pi}}, & j\ge 1,\\ \dfrac{1}{\sqrt{2\pi}}, & j=0.\end{cases}$$

Thus, we have

$$\begin{aligned}&(\mathbb{E}_f[\widehat{\psi}_{i,j}]-\psi_{i,j})^2\\ &\quad\sim\left(\frac{q(i)q(j)}{N(2t_{i,j})^{i+j+1}}+t_{i,j}(\psi_{i+1,j}+\psi_{i,j+1})\right)^2, \qquad N\to\infty.\end{aligned} \tag{55}$$

For both estimators the squared bias is the dominant term in the asymptotic mean squared error, because the variance is of the order $O(N^{-2}t^{-2i-2j-1})$. It follows that both estimators will have the same leading asymptotic mean square error term provided that

$$t_{i,j} = \left(\frac{1+2^{-i-j-1}}{3}\frac{-2q(i)q(j)}{N(\psi_{i+1,j}+\psi_{i,j+1})}\right)^{1/(2+i+j)}. \tag{56}$$

We estimate $t_{i,j}$ via

$$\hat{t}_{i,j} = \left(\frac{1+2^{-i-j-1}}{3}\frac{-2q(i)q(j)}{N(\widehat{\psi}_{i+1,j}+\widehat{\psi}_{i,j+1})}\right)^{1/(2+i+j)}. \tag{57}$$

Thus, estimation of $\psi_{i,j}$ requires estimation of $\psi_{i,j+1}$ and $\psi_{i+1,j}$, which in turn requires estimation of $\psi_{i+2,j}, \psi_{i+1,j+1}, \psi_{i,j+2}$ and so on applying formula (57), recursively. Observe that to estimate all $\psi_{i,j}$ for which $i+j=k$, that is, $\{\psi_{i,j} : i+j=k\}$, we need estimates of all $\{\psi_{i,j} : i+j=k+1\}$. For example, from formula (57) we can see that estimation of $t_{2,0}, t_{1,1}, t_{0,2}$ requires estimation of $t_{3,0}, t_{2,1}, t_{1,2}, t_{0,3}$.

For a given integer $k \geq 3$, we define the function $\gamma(t)$ as follows. Given an input $t > 0$:

1. Set $\hat{t}_{i,j} = t$ for all $i+j=k$.
2. Use the set $\{\hat{t}_{i,j} : i+j=k\}$ to compute all functionals $\{\widehat{\psi}_{i,j} : i+j=k\}$ via (53).
3. Use $\{\widehat{\psi}_{i,j} : i+j=k\}$ to compute $\{\hat{t}_{i,j} : i+j=k-1\}$ via (57).
4. If $k=2$ go to step 5; otherwise set $k := k-1$ and repeat from step 2.
5. Use $\{\widehat{\psi}_{i,j} : i+j=2\}$ to output

$$\gamma(t) = \left(2\pi N(\widehat{\psi}_{0,2}+\widehat{\psi}_{2,0}+2\widehat{\psi}_{1,1})\right)^{-1/3}.$$

The bandwidth selection rule simply consists of solving the equation $\gamma(t) = t$ for a given $k \geq 3$ via either the fixed point iteration in Algorithm 1 (ignoring step 4) or by using Newton's method. We obtain excellent numerical results for $k=4$ or $k=5$. Higher values of k did not change the value of t in any significant way, but only increased the computational cost of evaluating the function $\gamma(t)$. Again note that this appears to be the first successful plug-in bandwidth selection rule that does not involve any arbitrary reference rules, but it is purely data-driven. An efficient Matlab implementation of the bandwidth selection rule described here, and using the two-dimensional discrete cosine transform, can be downloaded freely from [4]. The Matlab implementation takes an additional step in which, once a fixed point of $\gamma(t)$ has been found, the final set of estimates $\{\widehat{\psi}_{i,j} : i+j=2\}$ is used to compute the entries $\sqrt{t_{X_1}}$ and $\sqrt{t_{X_2}}$ of the optimal diagonal bandwidth matrix ([53], page 111) for a Gaussian kernel of the form

$$\frac{1}{2\pi\sqrt{t_{X_1}t_{X_2}}}e^{-(x_1-y_1)^2/(2t_{X_1})-(x_2-y_2)^2/(2t_{X_2})}.$$

These entries are estimated via the formulas

$$t_{X_1} = \left(\frac{\widehat{\psi}_{0,2}^{3/4}}{4\pi N \widehat{\psi}_{2,0}^{3/4} (\widehat{\psi}_{1,1} + \sqrt{\widehat{\psi}_{2,0}\widehat{\psi}_{0,2}})} \right)^{1/3}$$

and

$$t_{X_2} = \left(\frac{\widehat{\psi}_{2,0}^{3/4}}{4\pi N \widehat{\psi}_{0,2}^{3/4} (\widehat{\psi}_{1,1} + \sqrt{\widehat{\psi}_{2,0}\widehat{\psi}_{0,2}})} \right)^{1/3}.$$

REFERENCES

[1] ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10** 1217–1223. MR0673656

[2] AZENCOTT, R. (1984). Density of diffusions in small time: Asymptotic expansions. In *Seminar on Probability, XVIII*. *Lecture Notes in Math.* **1059** 402–498. Springer, Berlin. MR0770974

[3] BELLMAN, R. (1961). *A Brief Introduction to Theta Functions*. Holt, Rinehart and Winston, New York. MR0125252

[4] BOTEV, Z. I. (2007). Kernel density estimation using Matlab. Available at http://www.mathworks.us/matlabcentral/fileexchange/authors/27236.

[5] BOTEV, Z. I. (2007). Nonparametric density estimation via diffusion mixing. Technical report, Dept. Mathematics, Univ. Queensland. Available at http://espace.library.uq.edu.au.

[6] CHAUDHURI, P. and MARRON, J. S. (2000). Scale space view of of curve estimation. *Ann. Statist.* **28** 408–428. MR1790003

[7] CHOI, E. and HALL, P. (1999). Data sharpening as a prelude to density estimation. *Biometrika* **86** 941–947. MR1741990

[8] COHEN, J. K., HAGIN, F. G. and KELLER, J. B. (1972). Short time asymptotic expansions of solutions of parabolic equations. *J. Math. Anal. Appl.* **38** 82–91. MR0303086

[9] CSISZÁR, I. (1972). A class of measures of informativity of observation channels. *Period. Math. Hungar.* **2** 191–213. MR0335152

[10] DEVRÔYE, L. (1997). Universal smoothing factor selection in density estimation: Theory and practice. *Test* **6** 223–320. MR1616896

[11] DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York. MR1847783

[12] ETHIER, S. N. and KURTZ, T. G. (2009). *Markov Processes. Characterization and Convergence*. Wiley, New York. MR0838085

[13] FELLER, W. (1952). The parabolic differential equations and the associated semi-groups of transformations. *Ann. of Math. (2)* **55** 468–519. MR0047886

[14] FRIEDMAN, A. (1964). *Partial Differential Equations of Parabolic Type*. Prentice Hall, Englewood Cliffs, NJ. MR0181836

[15] HALL, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika* **77** 523–535. MR1087843

[16] HALL, P., HU, T. C. and MARRON, J. S. (1995). Improved variable window kernel estimates of probability densities. *Ann. Ststist.* **23** 1–10. MR1331652

[17] HALL, P. and MARRON, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115. MR0907270

[18] HALL, P. and MINNOTTE, M. C. (2002). High order data sharpening for density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 141–157. MR1883130
[19] HALL, P. and PARK, B. U. (2002). New methods for bias correction at endpoints and boundaries. *Ann. Statist.* **30** 1460–1479. MR1936326
[20] HALL, P. and PARK, B. U. (2002). New methods for bias correction at endpoints and boundaries. *Ann. Statist.* **30** 1460–1479. MR1936326
[21] HAVRDA, J. H. and CHARVAT, F. (1967). Quantification methods of classification processes: Concepts of structural α entropy. *Kybernetika (Prague)* **3** 30–35. MR0209067
[22] JONES, M. C. and FOSTER, P. J. (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sinica* **6** 1005–1013. MR1422417
[23] JONES, M. C., MARRON, J. S. and PARK, B. U. (1991). A simple root n bandwidth selector. *Ann. Statist.* **19** 1919–1932. MR1135156
[24] JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1993). Simple boundary correction for kernel density estimation. *Statist. Comput.* **3** 135–146.
[25] JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91** 401–407. MR1394097
[26] JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). Progress in data-based bandwidth selection for kernel density estimation. *Comput. Statist.* **11** 337–381. MR1415761
[27] JONES, M. C., MCKAY, I. J. and HU, T. C. (1994). Variable location and scale kernel density estimation. *Ann. Inst. Statist. Math.* **46** 521–535. MR1309722
[28] JONES, M. C. and SIGNORINI, D. F. (1997). A comparison of higher-order bias kernel density estimators. *J. Amer. Statist. Assoc.* **92** 1063–1073. MR1482137
[29] KANNAI, Y. (1977). Off diagonal short time asymptotics for fundamental solutions of diffusion equations. *Comm. Partial Differential Equations* **2** 781–830. MR0603299
[30] KAPUR, J. N. and KESAVAN, H. K. (1987). *Generalized Maximum Entropy Principle (With Applications)*. Standford Educational Press, Waterloo, ON. MR0934205
[31] KARUNAMUNI, R. J. and ALBERTS, T. (2005). A generalized reflection method of boundary correction in kernel density estimation. *Canad. J. Statist.* **33** 497–509. MR2232376
[32] KARUNAMUNI, R. J. and ZHANG, S. (2008). Some improvements on a boundary corrected kernel density estimator. *Statist. Probab. Lett.* **78** 499–507. MR2400863
[33] KERM, P. V. (2003). Adaptive kernel density estimation. *Statist. J.* **3** 148–156.
[34] KLOEDEN, P. E. and PLATEN, E. (1999). *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin.
[35] LADYŽENSKAJA, O. A., SOLONNIKOV, V. A. and URAL'CEVA, N. N. (1967). *Linear and Quasilinear Equations of Parabolic Type. Translations of Mathematical Monographs* **23** xi+648. Amer. Math. Soc., Providence, RI. MR0241822
[36] LARSSON, S. and THOMEE, V. (2003). *Partial Differential Equations with Numerical Methods*. Springer, Berlin. MR1995838
[37] LEHMANN, E. L. (1990). Model specification: The views of fisher and neyman, and later developments. *Statist. Sci.* **5** 160–168. MR1062574
[38] LOADER, C. R. (1999). Bandwidth selection: Classical or plug-in. *Ann. Statist.* **27** 415–438. MR1714723
[39] LOFTSGAARDEN, D. O. and QUESENBERRY, C. P. (1965). A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.* **36** 1049–1051. MR0176567
[40] MARRON, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011–1023. MR0803755
[41] MARRON, J. S. and RUPPERT, D. (1996). Transformations to reduce boundary bias in kernel density-estimation. *J. Roy. Statist. Soc. Ser. B* **56** 653–671. MR1293239
[42] MARRON, J. S. and WAND, M. P. (1992). Exact mean integrated error. *Ann. Statist.* **20** 712–736. MR1165589

[43] MOLCHANOV, S. A. (1975). Diffusion process and Riemannian geometry. *Russian Math. Surveys* **30** 1–63.

[44] PARK, B. U., JEONG, S. O. and JONES, M. C. (2003). Adaptive variable location kernel density estimators with good performance at boundaries. *J. Nonparametr. Stat.* **15** 61–75. MR1958960

[45] PARK, B. U. and MARRON, J. S. (1990). Comparison of data-driven bandwidith selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

[46] SAMIUDDIN, M. and EL-SAYYAD, G. M. (1990). On nonparametric kernel density estimates. *Biometrika* **77** 865. MR1086696

[47] SCOTT, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*. Wiley, New York. MR1191168

[48] SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690. MR1125725

[49] SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London. MR0848134

[50] SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York. MR1391963

[51] TERRELL, G. R. and SCOTT, D. W. (1992). Variable kernel density estimation. *Ann. Statist.* **20** 1236–1265. MR1186249

[52] WAND, M. P. and JONES, M. C. (1994). Multivariate plug-in bandwidth selection. *Comput. Statist.* **9** 97–117. MR1280754

[53] WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London. MR1319818

SCHOOL OF MATHEMATICS AND PHYSICS
UNIVERSITY OF QUEENSLAND
ST. LUCIA, BRISBANE
QUEENSLAND, 4072
AUSTRALIA
E-MAIL: botev@maths.uq.edu.au
grotow@maths.uq.edu.au
kroese@maths.uq.edu.au
URL: http://www.maths.uq.edu.au/~kroese/