*Submitted to the Annals of Statistics*

# UNIFIED TREATMENT OF THE ASYMPTOTICS OF ASYMMETRIC KERNEL DENSITY ESTIMATORS

By Till Hoffmann and Nick Jones

*Imperial College London*

**1. Introduction.** Kernel density estimation with fixed bandwidth has become the most common method to estimate unknown densities of continuous random variables. The standard kernel density estimator (KDE) is defined as

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{1}$$

where $\{X_i : i = 1, \ldots, n\}$ is a set of $n$ i.i.d. samples of a univariate random variable $X$ drawn from a generating distribution $f$, $K$ is a non-negative function that integrates to unity, and $h$ is the bandwidth which controls the smoothness of the estimate [11]. We assume that $f$ is supported on the real line. Unless otherwise stated $\int$ will refer to an integral over the interval $(-\infty, \infty)$.

Before outlining the remainder of this paper, we briefly review three topics for which we will provide a unified treatment: variable-bandwidth KDEs, shifted KDEs and KDEs with asymmetric kernels.

*Variable-bandwidth KDEs:* Choosing the fixed bandwidth $h$ entails a trade-off. If the bandwidth is small, high-density regions are smoothed appropriately and we can recover small-scale features. Yet low-density regions will be undersmoothed. If the bandwidth is large, low-density regions are smoothed appropriately. Unfortunately, high-density regions will be oversmoothed and small-scale features will be masked. Terrell and Scott considered two variable-bandwidth KDEs to alleviate this problem [12]. *Balloon estimators* have a bandwidth that depends on the *evaluation point*, i.e. the point, $x$, at which $f(x)$ needs to be estimated,

$$\hat{f}(x) = \frac{1}{nh(x)}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h(x)}\right). \tag{2}$$

By contrast, the bandwidth of *sample-smoothing estimators* depends on the

sample associated with each kernel such that

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h(X_i)} K\left(\frac{X_i - x}{h(X_i)}\right). \tag{3}$$

The latter yield properly normalised density estimates, i.e. $\int dx\, \hat{f}(x) = 1$, because each term of the sum in Eq. (3) integrates to unity. In general, balloon estimators are not proper densities themselves.

*Shifted KDEs:* Standard KDEs tend to overestimate densities in regions with few samples and underestimate densities in regions with many samples. Samiuddin and El-Sayyad proposed to shift samples slightly from low-density regions to high-density regions to achieve lower bias [9]. Hall and Minnotte extended this approach to decrease the bias further using data sharpening estimators [3]. KDEs with variable bandwidth and shifted samples were considered by Jones et al. in the context of sample-smoothing estimators [8].

*Asymmetric KDEs:* A literature on using asymmetric kernels for density estimation has developed in parallel: gamma [2], log-normal and Birnbaum-Saunders [6], as well as inverse Gaussian and reciprocal inverse Gaussian [10] eliminate bias on or near the boundary of a (semi-)bounded interval. Whenever a new kernel is being considered, its asymptotic properties have to be carefully derived.

In the following, we provide a more general treatment of shifted sample-smoothing estimators and introduce shifted balloon estimators, and develop a generic framework to obtain the asymptotic properties of a range of kernel density estimators. After discussing mathematical preliminaries in Section 2, we define shifted balloon and sample-smoothing estimators in Sections 3 and 4, respectively. We derive the asymptotic bias, variance and mean (integrated) squared error for each class of estimators. In Section 5, we apply our results to density estimators with gamma, log-normal, Birnbaum-Saunders, inverse Gaussian and reciprocal inverse Gaussian kernels, derive plugin expressions for bandwidth selection, and propose two new density estimators for density estimation on the positive real line. We conclude in Section 6.

**2. Preliminaries.** The estimators in Eqs. (1) to (3) can all be expressed as general weight function estimators defined by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{\infty} W(X_i, x), \tag{4}$$

where $W(y,x)$ is a non-negative function that peaks near $x=y$. Because the samples are i.i.d., the first moment of the estimator is

$$\left\langle \hat{f}(x) \right\rangle = \langle W(X,x) \rangle , \tag{5}$$

where the expectation is with respect to the samples $X$. The second moment is

$$\begin{aligned} \left\langle \hat{f}^2(x) \right\rangle &= \frac{1}{n^2} \left\langle \sum_{i,j=1}^{n} W(X_i,x) W(X_j,x) \right\rangle \\ &= \frac{1}{n} \langle W^2(X,x) \rangle + \frac{n-1}{n} \langle W(X,x) \rangle^2 \end{aligned}$$

such that the variance becomes

$$\begin{aligned} \mathrm{var}\hat{f}(x) &= \left\langle \hat{f}^2(x) \right\rangle - \left\langle \hat{f}(x) \right\rangle^2 \\ &= \frac{1}{n} \left( \langle W^2(X,x) \rangle - \langle W(X,x) \rangle^2 \right). \end{aligned} \tag{6}$$

The canonical criterion for assessing the local quality of density estimates is the mean squared error (MSE) and is defined as

$$\mathrm{MSE}\hat{f}(x) = \left\langle \left( f(x) - \hat{f}(x) \right)^2 \right\rangle$$

for a given evaluation point $x$. We rewrite the MSE in terms of squared bias and variance such that

$$\begin{aligned} \mathrm{MSE}\hat{f}(x) &= \mathrm{bias}^2 \hat{f}(x) + \mathrm{var}\hat{f}(x), \\ \text{where} \quad \mathrm{bias}\hat{f}(x) &= \left\langle \hat{f}(x) \right\rangle - f(x) \\ &= \langle W(X,x) \rangle - f(x) . \end{aligned} \tag{7}$$

The mean integrated squared error (MISE) is defined as

$$\mathrm{MISE}\hat{f} = \int dx \, \mathrm{MSE}\hat{f}(x)$$

and will be our criterion for the global quality of density estimates.

The balloon estimator in Eq. (2), the sample-smoothing estimator in Eq. (3), and the estimators we consider in the following two sections can be expressed in terms of a kernel. Following Terrell and Scott, we formalise the notion of a kernel in the following definition [12].

DEFINITION 1. An order-$p$ kernel is a univariate, non-negative function $K$ such that

$$
\begin{aligned}
m_k &\equiv \int dz\, K(z) z^k \qquad \kappa \equiv \int dz\, K^2(z) \\
m_k &= \begin{cases} 1 & \text{if } k = 0,\, p \\ 0 & \text{if } 0 < k < p \end{cases} \qquad 0 < \kappa < \infty,
\end{aligned} \tag{8}
$$

where $m_k$ is the $k^{\text{th}}$ moment of the kernel.

**3. Shifted balloon estimator.** We now provide our modified version of the balloon estimator.

DEFINITION 2. A shifted balloon estimator is a kernel density estimator whose bandwidth depends on the point $x$ at which it is evaluated, i.e.

$$
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x)} K\left( \frac{X_i - x - h^p(x)\, \delta(x)}{h(x)} \right), \tag{9}
$$

where $p$ is the order of the kernel $K$, and $\delta(x)$ is a shift which remains bounded as $n \to \infty$, i.e. $\lim_{n\to\infty} |\delta(x)| < \infty$.

The kernel is thus shifted with respect to the sample by a small amount $h^p(x)\, \delta(x)$ which depends on the evaluation point and vanishes as the bandwidth decreases. Shifted balloon estimators do not yield properly-normalised density estimates in general, similar to their unshifted counterparts.

We need to consider the bias and variance to evaluate the MSE in the limit of small bandwidth.

THEOREM 1. *The mean and variance of a shifted balloon estimator with order-$p$ kernel are*

$$
\left\langle \hat{f}(x) \right\rangle = A_0(x) + A_p(x) + o\left(h^p(x)\right) \tag{10}
$$

$$
\text{and} \quad \operatorname{var} \hat{f}_h(x) = \frac{f(x)\, \kappa}{nh(x)} + o\left(n^{-1} h^{-1}(x)\right), \tag{11}
$$

$$
\text{where} \quad A_k(x) = h^k(x) \sum_{j=k}^{\infty} \frac{f^{(j)}(x)}{k!\,(j-k)!} \left[h^p(x)\, \delta(x)\right]^{j-k} \tag{12}
$$

*and $f^{(j)}(x)$ denotes the $j^{th}$ derivative of $f$ evaluated at $x$. The factorial is defined such that $0! \equiv 1$.*

A proof is given in Appendix A. Substituting Eqs. (12) and (10) into Eq. (7) and dropping higher-order terms in $h(x)$, the bias is

$$\text{bias}\hat{f}(x) = h^p(x)\,\delta(x)\,f'(x) + \frac{h^p(x)}{p!}f^{(p)}(x) + o\left(h^p(x)\right), \tag{13}$$

which is the same as the expression derived by Terrell and Scott if we let $\delta(x) = 0$ [12]. We thus see from Eq. (13) that any dependence of the bias on the first derivative of the generating distribution is due to the shift between the kernel mean and the associated sample as Chen noted in the context of density estimators using gamma kernels [2]. Adding Eq. (11) and the square of Eq. (13) yields the MSE

$$\text{MSE}\hat{f}(x) = \left[h^p(x)\,\delta(x)\,f'(x) + \frac{h^p(x)}{p!}f^{(p)}(x)\right]^2 + \frac{f(x)\,\kappa}{nh(x)} + o\left(n^{-1}h^{-1}(x) + h^{2p}(x)\right).$$

**4. Shifted sample-smoothing estimators.** In the following, we discuss our modified version of the sample-smoothing estimator.

Definition 3. A shifted sample-smoothing estimator is a kernel density estimator whose bandwidth depends on the samples $X_i$, i.e.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h(X_i)}K\left(\frac{X_i - x + h^p(X_i)\,\delta(X_i)}{h(X_i)}\right), \tag{14}$$

where $p$ is the order of the kernel $K$, and $\delta(X_i)$ is a shift which remains bounded as $n \to \infty$, i.e. $\lim_{n\to\infty}|\delta(X_i)| < \infty$.

Again, the kernel is shifted with respect to the sample by a small amount but the magnitude of the shift depends on the sample and not the evaluation point. Shifted sample-smoothing estimators are guaranteed to yield properly-normalised density estimates.

As in the previous section, we consider the mean and the variance assuming the bandwidth is small. See Appendix A for a proof of the following theorem.

Theorem 2. *If*

$$z = \frac{y - x + h^p(y)\,\delta(y)}{h(y)} \tag{15}$$

*is a monotonic function of $y$ for all $x$, the mean and variance of a shifted sample-smoothing estimator with order-p kernel are*

$$\left\langle \hat{f}(x) \right\rangle = B_0(x) + B_p(x) + o\left(h^p(x)\right)$$

$$\text{(16)} \quad \textit{and} \quad \text{var}\hat{f}(x) = \frac{f(x)\kappa}{nh(x)} + o\left(n^{-1}h^{-1}(x)\right),$$

$$\textit{where} \quad B_k = \sum_{j=k}^{\infty} \frac{1}{k!\,(j-k)!} \frac{d^j}{dx^j}\left(f(x)\, h^{k+p(j-k)}(x)\, (-\delta(x))^{j-k}\right).$$

The condition in Eq. (15) is satisfied by most kernels although exceptions exist as we will see in the next section. Using the same argument as in the previous section and dropping higher-order terms in $h(x)$, we have

$$\text{(17)} \quad \text{bias}\hat{f}(x) = -\frac{d}{dx}\left(f(x)\, h^p(x)\, \delta(x)\right) + \frac{1}{p!}\frac{d^p}{dx^p}\left(h^p(x)\, f(x)\right) + o\left(h^p(x)\right).$$

The expressions agree with Terrell and Scott if we let $\delta(x) = 0$ [12]. Jones et al. provide an alternative derivation for second-order kernels, i.e. $p = 2$ [8]. The MSE is

$$\text{MSE}\hat{f}(x) = \left[-\frac{d}{dx}\left(f(x)\, h^p(x)\, \delta(x)\right) + \frac{1}{p!}\frac{d^p}{dx^p}\left(h^p(x)\, f(x)\right)\right]^2 + \frac{\kappa f(x)}{nh(x)} + o\left(n^{-1}h^{-1}(x) + h^{2p}(x)\right).$$

**5. Application.** Having derived the asymptotic properties of shifted balloon and shifted sample-smoothing estimators allows us to obtain the asymptotic properties of a range of kernel density estimators easily. Before applying our method, we need to make a distinction that is immaterial for symmetric kernels. Recall the definition of the general weight function estimator which we repeat here for ease of reference:

$$\text{(18)} \quad \hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} W(X_i, x).$$

There are two fundamentally different choices for the weight function.

Definition 4. A weight function $W(y, x)$ is called *proper* if and only if

$$\int dy\, W(y, x) = 1.$$

It is called *improper* otherwise.

Estimators with proper weight functions are guaranteed to be densities themselves because each term in Eq. (18) integrates to unity. Sample-smoothing estimators belong to this class. Estimators with improper weight functions, such as balloon estimators, do not in general yield normalised densities. If the weight function is symmetric with respect to exchange of its argument, the distinction becomes immaterial. Standard kernels with fixed bandwidth as defined in Eq. (1) are examples of symmetric weight functions if the kernel is an even function of its argument.

5.1. *Density estimation using improper weight functions.* Improper weight functions have recently become popular for estimating densities with (semi-)bounded support to avoid boundary bias. See Section 2.11 in [13] for a detailed discussion of boundary bias. In particular, Brown and Chen used beta-distribution kernels for non-parametric regression on a finite interval [1]. Chen proposed the use of gamma-distribution kernels to estimate densities of positive random variables [2], which spurred further research into positive density estimation using Birnbaum-Saunders and log-normal kernels [6] as well as inverse Gaussian and reciprocal inverse Gaussian kernels [10].

Intuitively, the contribution from any weight function becomes ever more sharply peaked as the bandwidth of the estimator decreases. Because the global properties become less important, we should be able to approximate general weight functions by simpler kernels that peak where the sample and evaluation point are close. We will see that the improper estimators in the literature are asymptotically equivalent to shifted balloon estimators as defined in Eq. (9). This observation enables us to derive their asymptotic properties using the general results from Section 3 instead of having to derive them independently. We consider the gamma-kernel density estimator proposed by Chen [2] in detail and discuss other kernels in Section 5.3.

After reparametrisation, Chen's improper estimator is given by

$$\hat{f}\left(x\right) = \sum_{i=1}^{n} \mathrm{G}\left(X_i; 1 + \frac{x}{\sigma^2}, \sigma^2\right),$$

where $\sigma$ is a bandwidth parameter, and $\mathrm{G}\left(t; k, \theta\right) = \frac{\theta^{-k}}{\Gamma(k)} x^{k-1} \exp\left(-\frac{t}{\theta}\right)$ denotes the probability density function of a gamma random variable $t$ with shape parameter $k$ and scale parameter $\theta$. If the bandwidth $\sigma$ is small, the shape parameter of the gamma distribution is large and it can be approximated by a Gaussian with the same mean [7], i.e.

$$\mathrm{G}\left(X_i; 1 + \frac{x}{\sigma^2}, \sigma^2\right) \to \mathcal{N}\left(X_i; x + \sigma^2, \sigma^2\left(x + \sigma^2\right)\right).$$

The approximation takes the form of a shifted balloon estimator with Gaussian kernel and

$$h(x) = \sigma\sqrt{x+\sigma^2}$$
$$h^2(x)\,\delta(x) = \sigma^2.$$

Having represented Chen's improper estimator as a shifted balloon estimator, we can find the bias and variance using Eqs. (13) and (11):

$$\begin{aligned}\operatorname{bias}\hat{f}(x) &= \sigma^2 f'(x) + \frac{\sigma^2\left(x+\sigma^2\right)}{2} f''(x) + o\left(\sigma^2\right)\\ &= \sigma^2\left(f'(x) + \frac{x}{2} f''(x)\right) + o\left(\sigma^2\right)\\ \text{and}\quad \operatorname{var}\hat{f}(x) &= \frac{f(x)}{2n\sigma\sqrt{\pi\left(x+\sigma^2\right)}} + o\left(n^{-1}\sigma^{-1}\right)\\ &= \frac{f(x)}{2n\sigma\sqrt{\pi x}} + o\left(n^{-1}\sigma^{-1}\right),\end{aligned}$$

where we have used that Gaussians are second-order kernels with $\kappa = \frac{1}{2\sqrt{\pi}}$. The expressions agree with the ones derived by Chen. The MSE is

$$\operatorname{MSE}\hat{f}(x) = \sigma^4\left(f'(x) + \frac{x}{2} f''(x)\right)^2 + \frac{f(x)}{2n\sigma\sqrt{\pi x}} + o\left(\sigma^4 + n^{-1}\sigma^{-1}\right).$$

Extending the work by Chen [2], we approximate $f(x)$ by a reference distribution with known functional form which enables us to obtain a plugin expression for selecting the bandwidth. The parametric form of the reference distribution is largely a free choice; we assume that $f(x)$ is approximately log-normal with logarithmic mean $\mu$ and variance $\Sigma^2$ for convenience. Integrating yields

$$\operatorname{MISE}\hat{f} = \sigma^4\frac{\exp\left(\frac{9\Sigma^2}{4} - 3\mu\right)\left(12 + 20\Sigma^2 + 9\Sigma^4\right)}{128\sqrt{\pi}\Sigma^4} + \frac{\exp\left(\frac{\Sigma^2}{8} - \frac{\mu}{2}\right)}{2n\sigma\sqrt{\pi}} \tag{19}$$

and setting the derivative of the MISE with respect to $\sigma$ to zero gives the optimal scale parameter

$$\sigma^* = \frac{2^{4/5}\Sigma\exp\left(\frac{\mu}{2} - \frac{17\Sigma^2}{40}\right)}{\left(12 + 20\Sigma^2 + 9\Sigma^4\right)^{1/5}} n^{-1/5}. \tag{20}$$

The expression allows us to obtain a bandwidth by estimating $\mu$ and $\Sigma$ from the samples without the need for computationally-expensive cross-validation. Plugin bandwidth estimators are particularly useful for preliminary data exploration. An example density estimate is shown in Figure 1.

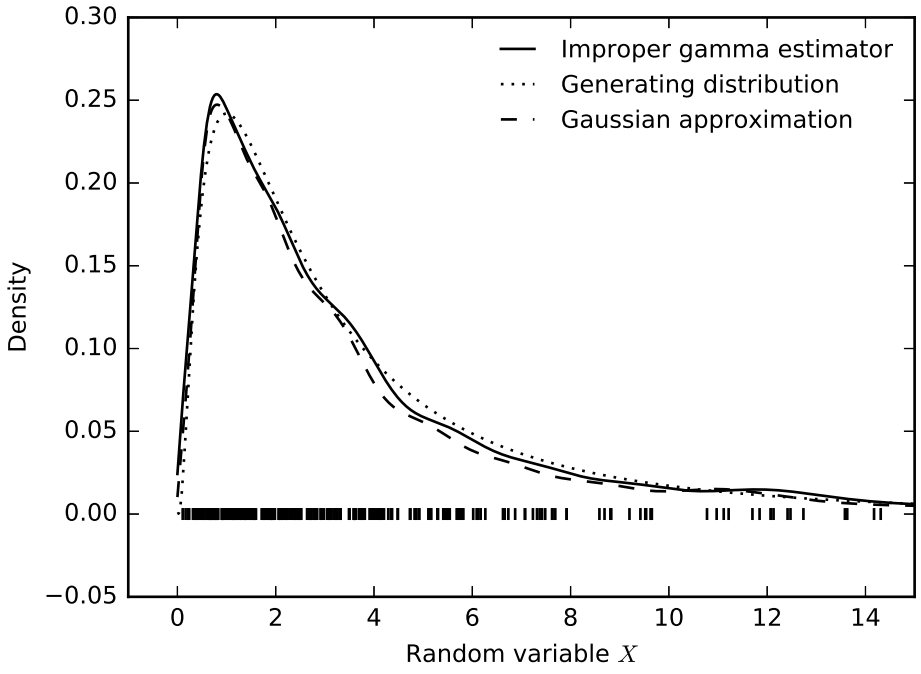


FIG 1. *Density estimate of* $n = 300$ *samples drawn from a log-normal distribution with logarithmic mean* $\mu = 1$ *and logarithmic variance* $\Sigma^2 = 1$ *using an improper gamma estimator (solid black line) and a Gaussian approximation (dashed black line). The generating distribution is shown as a dotted black line and individual samples are shown as markers on the horizontal axis. The bandwidth for both estimators was obtained using the plugin expression in Eq.* (20) *which minimises the MISE in Eq.* (19).

5.2. *Density estimation using proper weight functions.* Improper weight functions have the unappealing property that they do not yield properly normalised density estimates in general. Jeon and Kim proposed a density estimator using a proper gamma kernel to alleviate this problem [5]. After reparametrisation, their estimator is given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{\infty} \mathrm{G}\left(x; 1 + \frac{X_i}{\sigma^2}, \sigma^2\right), \tag{21}$$

i.e. the role of the evaluation point and the samples have been swapped. The kernel can be approximated by a normal distribution if the bandwidth is small and the estimator takes the form of a shifted sample-smoothing estimator

$$\mathrm{G}\left(x; 1 + \frac{X_i}{\sigma^2}, \sigma^2\right) \to \mathcal{N}\left(x; X_i + \sigma^2, \sigma^2\left(X_i + \sigma^2\right)\right).$$

Using Eqs. (17) and (16), the bias and variance are

$$\begin{aligned}
\mathrm{bias}\hat{f}(x) &= -\sigma^2 f'(x) + \frac{1}{2}\frac{d^2}{dx^2}\left(\sigma^2\left(x+\sigma^2\right) f(x)\right) + o\left(\sigma^2\right) \\
&= \frac{\sigma^2 x}{2} f''(x) + o\left(\sigma^2\right) \\
\text{and} \quad \mathrm{var}\hat{f}(x) &= \frac{f(x)}{2n\sigma\sqrt{\pi x}} + o\left(n^{-1}\sigma^{-1}\right).
\end{aligned}$$

Jeon and Kim obtained an upper bound for the MSE of the estimator defined in Eq. (21) but we can provide an explicit form:

$$\mathrm{MSE}\hat{f}(x) = \frac{\sigma^4}{4}\left[x f''(x)\right]^2 + \frac{f(x)}{2n\sigma\sqrt{\pi x}} + o\left(\sigma^4 + n^{-1}\sigma^{-1}\right).$$

Employing the log-normal distribution as a reference distribution, and optimising the MISE with respect to $\sigma$ yields the plugin expression

$$\sigma^* = \frac{2^{4/5}\Sigma \exp\left(\frac{\mu}{2} - \frac{17\Sigma^2}{40}\right)}{\left(12 + 4\Sigma^2 + \Sigma^4\right)^{1/5}} n^{-1/5}.$$

As discussed in Appendix B, cross-validation is straightforward for Jeon and Kim's estimator. A comparison between the plugin expression and leave-one-out cross-validation is shown in Figure 2. The results are encouraging even for a modest sample size of $n = 300$. The MISE profile obtained from cross-validation is more variable because it depends on each individual sample whereas the asymptotic MISE profile only depends on the logarithmic mean and variance of the whole sample.

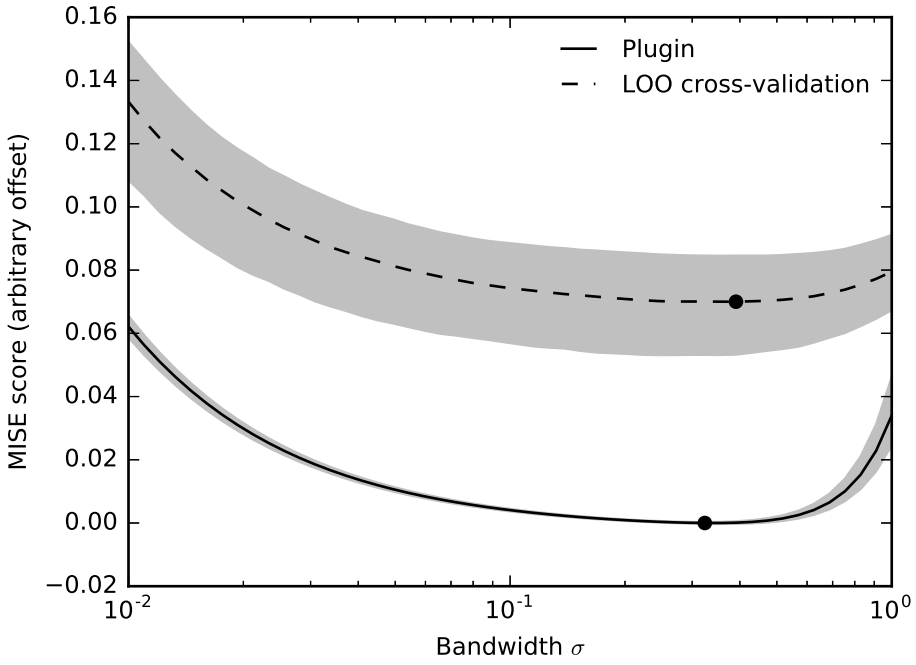


FIG 2. *MISE profiles as a function of bandwidth $\sigma$ using a Gaussian approximation (solid line) and leave-one-out (LOO) cross-validation (dashed line). The shaded bands correspond to the 90% confidence interval from one thousand independent samples of size $n = 300$ drawn from a log-normal distribution with $\mu = \Sigma = 1$. The offset between the two curves is arbitrary.*

| Distribution | $\langle t \rangle$ | Variance | Parametrisation | $h^2(x)\,\delta(x)$ | $h(x)$ |
|---|---|---|---|---|---|
| $\mathrm{G}(t;k,\theta) = \frac{\theta^{-k} t^{k-1}}{\Gamma(k)} \exp\left(-\frac{t}{\theta}\right)$ | $k\theta$ | $k\theta^2$ | $\mathrm{G}\left(y; \frac{x}{\sigma^2}+1, \sigma^2\right)$ | $\sigma^2$ | $\sigma\sqrt{x+\sigma^2}$ |
| $\mathrm{LN}\left(t;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log t-\mu)^2}{2\sigma^2}\right)$ | $\mathrm{e}^{\mu+\frac{\sigma^2}{2}}$ | $\langle t \rangle^2 \left(\mathrm{e}^{\sigma^2}-1\right)$ | $\mathrm{LN}\left(y; \log x, \sigma^2\right)$ | $x\left(\mathrm{e}^{\frac{\sigma^2}{2}}-1\right)$ | $x\mathrm{e}^{\frac{\sigma^2}{2}}\sqrt{\mathrm{e}^{\sigma^2}-1}$ |
| $\mathrm{BS}(t;\sigma,\lambda) = \frac{(1+t\lambda)}{2\sigma\sqrt{2\pi\lambda t^3}} \exp\left(-\frac{(t\lambda-1)^2}{2t\lambda\sigma^2}\right)$ | $\frac{2+\sigma^2}{2\lambda}$ | $\frac{\sigma^2\left(4+5\sigma^2\right)}{4\lambda^2}$ | $\mathrm{BS}\left(y;\sigma,x^{-1}\right)$ | $\frac{x\sigma^2}{2}$ | $x\sigma\sqrt{1+5\sigma^2/4}$ |
| $\mathrm{IG}(t;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left(-\frac{\lambda(t-\mu)^2}{2t\mu^2}\right)$ | $\mu$ | $\frac{\mu^3}{\lambda}$ | $\mathrm{IG}\left(y;x,\sigma^{-2}\right)$ | 0 | $\sigma x^{3/2}$ |
| $\mathrm{RIG}(t;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi t}} \exp\left(-\frac{\lambda(\mu t-1)^2}{2t\mu^2}\right)$ | $\frac{1}{\mu}+\frac{1}{\lambda}$ | $\frac{1}{\lambda\mu}+\frac{2}{\lambda^2}$ | $\mathrm{RIG}\left(y;\frac{1}{x-\sigma^2},\sigma^{-2}\right)$ | 0 | $\sigma\sqrt{x+\sigma^2}$ |

TABLE 1. *List of probability distributions with their mean and variance. The fourth column shows the parametrisation used for improper weight functions, where $x$ is the evaluation point and $y$ is the associated sample. The parametrisation for proper weight functions can be obtained by swapping $x$ and $y$. The fifth column lists the shift between the mean of the kernel and the sample. The effective bandwidth is shown in the last column.*

| | Improper weight function | | | Proper weight function | |
|---|---|---|---|---|---|
| Kernel with ref. | $n \times \mathrm{var}_\sigma$ | $\mathrm{bias}_\sigma$ | $n^{1/5} \times \sigma*$ | $\mathrm{bias}_\sigma$ | $n^{1/5} \times \sigma*$ |
| G [2, 5] | $\frac{f(x)}{2\sigma\sqrt{\pi x}}$ | $\sigma^2\left(f'(x)+\frac{xf''(x)}{2}\right)$ | $\frac{2^{4/5}\Sigma\exp\left(\frac{\mu}{2}-\frac{17\Sigma^2}{40}\right)}{\left(12+20\Sigma^2+9\Sigma^4\right)^{1/5}}$ | $\frac{\sigma^2}{2}xf''(x)$ | $\frac{2^{4/5}\Sigma\exp\left(\frac{\mu}{2}-\frac{17\Sigma^2}{40}\right)}{\left(12+4\Sigma^2+\Sigma^4\right)^{1/5}}$ |
| LN [6] | $\frac{f(x)}{2\sqrt{\pi}\sigma x}$ | $\frac{\sigma^2 x}{2}\left(f'(x)+xf''(x)\right))$ | $\frac{2^{4/5}\Sigma\exp\frac{\Sigma^2}{20}}{\left(12+4\Sigma^2+\Sigma^4\right)^{1/5}}$ | $\frac{\sigma^2}{2}\frac{d}{dx}\left(x\frac{d}{dx}\left(xf(x)\right)\right)$ | $\frac{2^{4/5}\Sigma\exp\frac{\Sigma^2}{20}}{\left(12+4\Sigma^2+\Sigma^4\right)^{1/5}}$ |
| BS [6] | same as log-normal | | | | |
| IG [10] | $\frac{f(x)}{2\sqrt{\pi}\sigma x^{3/2}}$ | $\frac{\sigma^2}{2}x^3f''(x)$ | $\frac{2^{4/5}\Sigma\exp\left(\frac{7\Sigma^2}{40}-\frac{\mu}{2}\right)}{\left(12+68\Sigma^2+225\Sigma^4\right)^{1/5}}$ | method not applicable | |
| RIG [10] | $\frac{f(x)}{2\sigma\sqrt{\pi x}}$ | $\frac{\sigma^2}{2}xf''(x)$ | $\frac{2^{4/5}\Sigma\exp\left(\frac{\mu}{2}-\frac{17\Sigma^2}{40}\right)}{\left(12+4\Sigma^2+\Sigma^4\right)^{1/5}}$ | $\sigma^2\left(f'(x)+\frac{xf''(x)}{2}\right)$ | $\frac{2^{4/5}\Sigma\exp\left(\frac{\mu}{2}-\frac{17\Sigma^2}{40}\right)}{\left(12+20\Sigma^2+9\Sigma^4\right)^{1/5}}$ |

TABLE 2. *List of kernels with their asymptotic variance. The bias and optimal plugin bandwidth $\sigma^*$ are shown separately for proper and improper estimators.*

5.3. *Application to estimators with other kernels.* We use the methods developed above to derive the asymptotic properties of estimators with proper and improper weight functions. The gamma, log-normal (LN), Birnbaum-Saunders (BS), inverse Gaussian (IG) and reciprocal inverse Gaussian (RIG) distributions are listed in Table 1. To serve as a kernel, we parametrise the distributions in terms of the evaluation point $x$, associated sample $y$, and bandwidth parameter $\sigma$. In the limit $\sigma \to 0$, all kernels converge to a normal distribution [7] with mean $x + h^2(x)\,\delta(x)$ and variance $h^2(x)$ such that our method is applicable. The parametrisation for proper estimators can be obtained by exchanging the role of the evaluation point and the associated sample.

The asymptotic bias and variance for proper and improper estimators are listed in Table 2. The expressions for the asymptotics of improper estimators agree with the expressions in the original papers after reparametrisation. Because improper BS and LN estimators have a higher-order dependence on the evaluation point than the improper gamma estimator, they have smaller bias near the boundary which has been confirmed by simulations [6]. The same applies to the improper IG estimator because its bias is cubic in the evaluation point.

The proper estimators with LN, BS, and RIG kernels are new contributions although the LN kernel is equivalent to a log-transform of the samples followed by standard kernel density estimation using Gaussian kernels (see Section 2.11 in [11]), which is widely used. The method we have developed is not applicable to proper estimators with RIG kernels because the assumption that $z(y, x) = \frac{y-x}{\sigma y^{3/2}}$ is monotonic in $y$ is violated. All estimators achieve the same $n^{-4/5}$ convergence rate as Gaussian KDEs because their MISE is of the form $\sigma^4 E + (\sigma n)^{-1} F$, where $E$ and $F$ are numbers that do not depend on $n$ or $\sigma$.

**6. Discussion.** We have extended balloon and sample-smoothing estimators by a shift parameter and derived their asymptotic properties. Our approach facilitates the unified study of a wide range of density estimators which are subsumed under these two general classes of kernel density estimators. We have demonstrated our method by obtaining the asymptotics of estimators with improper gamma, log-normal, Birnbaum-Saunders, inverse Gaussian and reciprocal inverse Gaussian. Recently, an estimator with proper gamma kernel was proposed [5]; we have obtained the first explicit expression for its bias. We have proposed estimators with proper Birnbaum-Saunders and reciprocal inverse Gaussian kernels which have the distinct advantage of yielding normalised density estimates. Plugin expressions for

bandwidth estimation are provided for both proper and improper estimators using the log-normal distribution as a reference distribution.

Given the plethora of density estimators, which ones should be used? If the generating density is expected to have a large gradients, the proper gamma estimator and the improper inverse Gaussian (IG) or reciprocal inverse Gaussian (RIG) estimators are suitable because their bias is independent of the first derivative. The improper log-normal, Birnbaum-Saunders, IG and RIG estimators as well as the proper gamma estimator have relatively low bias near the origin and are appropriate for samples that are clustered near zero. Estimators whose effective bandwidth has a higher-order dependence on the evaluation point should be used for heavily-skewed data to avoid spurious wiggles in the tails of the distribution. We believe that proper estimators are preferable because they are guaranteed to yield normalised densities.

In future work, we would like to apply our approach to further kernels and conduct a full simulation study of the finite-sample properties of the estimators.

## APPENDIX A: PROOFS

LEMMA 1. *The integral*

$$L = \int \frac{dy}{h(x)} \phi\left(\frac{y - x - h^p(x)\,\delta(x)}{h(x)}\right) \theta(y)$$

*is given by*

$$L = \sum_{k=0}^{\infty} A_k(x) \int dz\, \phi(z)\, z^k,$$

$$\textit{where} \quad A_k(x) = h^k(x) \sum_{j=k}^{\infty} \frac{\theta^{(j)}(x)}{k!\,(j-k)!} \left[h^p(x)\,\delta(x)\right]^{j-k}.$$

PROOF. We make the change of variables

$$z = \frac{y - x - h^p(x)\,\delta(x)}{h(x)}$$

such that, holding $x$ constant,

$$dz = \frac{dy}{h(x)}$$

$$\text{and} \quad L = \int dz\, \phi(z)\, \theta\left(x + h^p(x)\,\delta(x) + z h(x)\right).$$

Expanding $\theta$ as a Taylor series about $x$ yields

$$L = \int dz\, \phi(z) \sum_{j=0}^{\infty} \frac{\theta^{(j)}(x)}{j!} \left(h^p(x)\, \delta(x) + z h(x)\right)^j .$$

Using the binomial theorem and exchanging the order of summation gives

$$\begin{aligned} L &= \int dz\, \phi(z) \sum_{j=0}^{\infty} \frac{\theta^{(j)}(x)}{j!} \sum_{k=0}^{j} \binom{j}{k} z^k h^{k+p(j-k)}(x)\, \delta^{j-k}(x) \\ &= \sum_{k=0}^{\infty} A_k(x) \int dz\, \phi(z)\, z^k, \\ \text{where} \quad A_k(x) &= h^k(x) \sum_{j=k}^{\infty} \frac{\theta^{(j)}(x)}{k!\,(j-k)!} \left[h^p(x)\, \delta(x)\right]^{j-k} . \end{aligned}$$

□

Proof of Theorem 1. From Eqs. (5) and (9), the first moment of a shifted balloon estimator is

$$\left\langle \hat{f}(x) \right\rangle = \int \frac{dy}{h(x)} K\left(\frac{y - x - h^p(x)\, \delta(x)}{h(x)}\right) f(y) .$$

Let $\phi(z) = K(z)$ and $\theta(y) = f(y)$. Then by Lemma 1

$$\left\langle \hat{f}(x) \right\rangle = \sum_{k=0}^{\infty} A_k(x) \int dz\, K(z)\, z^k, \tag{22}$$

$$\text{where} \quad A_k(x) = h^k(x) \sum_{j=k}^{\infty} \frac{f^{(j)}(x)}{k!\,(j-k)!} \left[h^p(x)\, \delta(x)\right]^{j-k} .$$

Recall that by Definition 1, the integral in Eq. (22) vanishes for $0 < k < p$ and is equal to unity for $k = 0, p$ such that

$$\left\langle \hat{f}(x) \right\rangle = A_0(x) + A_p(x) + o\left(h^p(x)\right) .$$

According to Eqs. (6) and (9), the variance of a shifted balloon estimator is

$$\operatorname{var} \hat{f}(x) = \frac{1}{n} \int \frac{dy}{h^2(x)} K^2\left(\frac{y - x - h^p(x)\, \delta(x)}{h(x)}\right) f(y) - \frac{\left\langle \hat{f}(x) \right\rangle^2}{n}. \tag{23}$$

We use Lemma 1 with $\phi(z) = K^2(z)$and $\theta(y) = f(y)/h(x)$ such that the variance becomes

$$\mathrm{var}\hat{f}(x) = \frac{f(x)\kappa}{nh(x)} + o\left(n^{-1}h^{-1}(x)\right).$$

The second term in Eq. (23) is absorbed by $o\left(n^{-1}h^{-1}(x)\right)$ because of its zero-order dependence on $h(x)$. □

LEMMA 2. *The integral*

$$L = \int \frac{dy}{h(y)} \phi\left(\frac{y - x + h^p(y)\delta(y)}{h(y)}\right) \theta(y)$$

*is given by*

$$L = \sum_{k=0}^{\infty} B_k(x) \int dz\, \phi(z)\, z^k,$$

$$\textit{where} \quad B_k(x) = \sum_{j=k}^{\infty} \frac{1}{k!(j-k)!} \frac{d^j}{dx^j}\left(\theta(x) h^{k+p(j-k)}(x) (-\delta(x))^{j-k}\right)$$

*if*

$$z(y,x) = \frac{y - x + h^p(y)\delta(y)}{h(y)} \tag{24}$$

*is a monotonic function of $y$ for all $x$ such that the inverse $y(z,x)$ exists.*

PROOF. We make the change of variables in Eq. (24) such that, holding $x$ constant,

$$\begin{aligned} dz &= \left(\frac{\frac{d}{dy}\left(y - x + h^p(y)\delta(y)\right)}{h(y)} - \frac{y - x + h^p(y)\delta(y)}{h(y)} \frac{h'(y)}{h(y)}\right) dy \\ &= \frac{dy}{h(y)} \frac{d}{dy}\left(y - x + h^p(y)\delta(y) - zh(y)\right) \end{aligned}$$

$$\text{and} \quad L = \int dz\, \phi(z) \times \frac{\theta(y(z,x))}{\left[\frac{d}{dy}\left(y - x + h^p(y)\delta(y) - zh(y)\right)\right]_{y=y(z,x)}}. \tag{25}$$

For fixed $z$, we express the fraction in the integrand of Eq. (25) as the contour integral

$$\ell = \frac{1}{2\pi i} \oint_{\gamma} d\tau \frac{\theta(\tau)}{\tau - x + h^p(\tau)\delta(\tau) - zh(\tau)}, \tag{26}$$

where $\tau$ is an integration variable. The contour $\gamma$ encloses the points defined by $\tau = y(z,x)$ and $\tau = x$. The simple pole at $\tau = y(z,x)$ has residue (see Theorem 8.15 in [4])

$$\frac{\theta(y(z,x))}{\left[\frac{d}{dy}\left(y - x + h^p(y)\,\delta(y) - zh(y)\right)\right]_{y=y(z,x)}}$$

such that the contour integral in Eq. (26) is indeed equal to the second term in Eq. (25) by Cauchy's residue theorem.

Expanding $\ell$ as a power series about $\tau = x$ yields

$$\begin{aligned}\ell &= \frac{1}{2\pi i}\oint_\gamma d\tau\,\frac{\theta(y)}{\tau - x}\sum_{j=0}^{\infty}\left(\frac{zh(\tau) - h^p(x)\,\delta(\tau)}{\tau - x}\right)^j \\ &= \frac{1}{2\pi i}\oint_\gamma d\tau\,\theta(y)\sum_{j=0}^{\infty}\sum_{k=0}^{j}\binom{j}{k}\frac{z^k h^{k+p(j-k)}(\tau)\left(-\delta(\tau)\right)^{j-k}}{(\tau - x)^{j+1}}.\end{aligned}$$

The second equality follows by the binomial theorem. The $j^{\text{th}}$ term gives rise to a pole of order $j+1$ at $\tau = x$ with residue (see Theorem 8.17 in [4])

$$R_j = \sum_{k=0}^{j}\frac{z^k}{k!\,(j-k)!}\frac{d^j}{dx^j}\left(\theta(x)\,h^{k+p(j-k)}(x)\left(-\delta(x)\right)^{j-k}\right)$$

such that $\ell = \sum_{j=0}^{\infty} R_j$. Substituting back into Eq. (25) and exchanging the order of summation yields

$$\begin{aligned}L &= \int dz\,\phi(z)\sum_{j=0}^{\infty}\sum_{k=0}^{j}\frac{z^k}{k!\,(j-k)!}\frac{d^j}{dx^j}\left(\theta(x)\,h^{k+p(j-k)}(x)\left(-\delta(x)\right)^{j-k}\right) \\ &= \sum_{k=0}^{\infty}B_k(x)\int dz\,\phi(z)\,z^k \\ \text{where}\quad B_k(x) &= \sum_{j=k}^{\infty}\frac{1}{k!\,(j-k)!}\frac{d^j}{dx^j}\left(\theta(x)\,h^{k+p(j-k)}(x)\left(-\delta(x)\right)^{j-k}\right).\end{aligned}$$

□

Proof of Theorem 2. From Eqs. (5) and (14), the first moment of a shifted sample-smoothing estimator is

$$\left\langle \hat{f}(x)\right\rangle = \int\frac{dy}{h(y)}\,K\left(\frac{y - x + h^p(y)\,\delta(y)}{h(y)}\right)f(y).$$

Let $\phi(z) = K(z)$ and $\theta(y) = f(y)$. Then by Lemma 2

$$\left\langle \hat{f}(x) \right\rangle = \sum_{k=0}^{\infty} B_k(x) \int dz\, K(z)\, z^k, \tag{27}$$

$$\text{where} \quad B_k(x) = \sum_{j=k}^{\infty} \frac{1}{k!\,(j-k)!} \frac{d^j}{dx^j} \left( f(x)\, h^{k+p(j-k)}(x)\, (-\delta(x))^{j-k} \right).$$

Recall that by Definition 1, the integral in Eq. (27) vanishes for $0 < k < p$ and is equal to unity for $k = 0,\, p$ such that

$$\left\langle \hat{f}(x) \right\rangle = B_0(x) + B_p(x) + o\left(h^p(x)\right).$$

According to Eqs. (6) and (14), the variance of a shifted sample-smoothing estimator is

$$\text{var}\hat{f}(x) = \frac{1}{n} \int \frac{dy}{h^2(y)} K^2 \left( \frac{y - x + h^p(y)\,\delta(y)}{h(y)} \right) f(y) - \frac{\left\langle \hat{f}(x) \right\rangle^2}{n}. \tag{28}$$

We use Lemma 1 with $\phi(z) = K^2(z)$and $\theta(y) = f(y)/h(y)$ such that the variance becomes

$$\text{var}\hat{f}(x) = \frac{f(x)\,\kappa}{nh(x)} + o\left(n^{-1}h^{-1}(x)\right).$$

The second term in Eq. (28) is absorbed by $o\left(n^{-1}h^{-1}(x)\right)$ because of its zero-order dependence on $h(x)$. □

## APPENDIX B: CROSS-VALIDATION

The quantity

$$M = \int dx\, \hat{f}^2(x) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_i(X_i) \tag{29}$$

is an unbiased estimator of the MISE of the density estimator $\hat{f}(x)$, where $\hat{f}_i(X_i)$ is the density estimated from all samples except $X_i$ evaluated at $X_i$. Details can be found in Section 3.4.3 of [11]. The second term in Eq. (29) is straightforward to evaluate whereas the first requires integration of the density estimate over the whole domain. Fortunately, the integral is tractable

for the proper gamma estimator in Eq. (21):

$$\int dx\, \hat{f}\,(x) = \frac{1}{n^2}\sum_{i,j=1}^{n}\int dx\, \mathrm{G}\left(x; 1+\frac{X_i}{\sigma^2}, \sigma^2\right)\mathrm{G}\left(x; 1+\frac{X_j}{\sigma^2}, \sigma^2\right)$$
$$= \frac{1}{n^2}\sum_{i,j=1}^{n}\frac{\Gamma\left(1+\frac{X_i+X_j}{\sigma^2}\right)2^{-1-\frac{X_i+X_j}{\sigma^2}}}{\Gamma\left(1+X_i/\sigma^2\right)\Gamma\left(1+X_j/\sigma^2\right)}.$$

## REFERENCES


[1] Brown, B. M. and Chen, S. X. (1999). Beta-Bernstein Smoothing for Regression Curves with Compact Support. *Scandinavian Journal of Statistics* **26** 47–59.

[2] Chen, S. X. (2000). Probability Density Function Estimation Using Gamma Kernels. *Annals of the Institute of Statistical Mathematics* **52** 471–480.

[3] Hall, P. and Minnotte, M. C. (2002). High order data sharpening for density estimation. *Journal of the Royal Statistical Society B* **64** 141–157.

[4] Howie, J. M. (2008). *Complex Analysis*. Springer.

[5] Jeon, Y. and Kim, J. (2014). A gamma kernel density estimation for insurance loss data. *Insurance: Mathematics and Economics* **53** 569–579.

[6] Jin, X. and Kawczak, J. (2003). Birnbaum-Saunders and Lognormal Kernel Estimators for Modelling Durations in High Frequency Financial Data. *Annals of Economics and Finance* **4** 103–124.

[7] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions* **1**. Wiley.

[8] Jones, M. C., McKay, I. J. and Hu, T. C. (1994). Variable location and scale kernel density estimation. *Annals of the Institute of Statistical Mathematics* **46** 521–535.

[9] Samuiddin, M. and El-Sayyad, G. M. (1990). On nonparametric kernel density estimates. *Biometrika* **77** 865-874.

[10] Scaillet, O. (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels. *Journal of Nonparametric Statistics* **16** 217–226.

[11] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

[12] Terrell, G. R. and Scott, D. W. (1992). Variable Kernel Density Estimation. *Ann. Statist.* **20** 1236–1265.

[13] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall.



Department of Mathematics
Imperial College London
London, SW7 2AZ
United Kingdom
E-mail: t.hoffmann13@imperial.ac.uk
nick.jones@imperial.ac.uk