BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION OF FAT-TAILED AND SKEWED DISTRIBUTIONS

Abstract. Applied researchers using kernel density estimation have worked with optimal bandwidth rules that invariably assumed that the reference density is Normal (optimal only if the true underlying density is Normal). We offer four new optimal bandwidth rules-of-thumb based on other infinitely supported distributions: Logistic, Laplace, Student's t and Asymmetric Laplace. Additionally, we propose a psuedo rule-of-thumb (ROT) bandwidth based on a Gram-Charlier expansion of the unknown reference density that is linked to the empirical skewness and kurtosis of the data. The intellectual investment needed to implement these new optimal bandwidths is practically zero. We discuss the behavior of these bandwidths as it links to differences in skewness and kurtosis to the Normal reference ROT. We further propose model selection criteria for bandwidth choice when the true underlying density is unknown. The performance of these new ROT bandwidths are assessed in a variety of Monte Carlo simulations as well as a timely example on stock market trading.

2020 Mathematics Subject Classification 62G07 (Density estimation), 05C78 (Graph labelling), 65D10 (Numerical smoothing, curve fitting). Word Count: 8683

1. Introduction

> "Many profest Christians are like to foolish builders, who build by guess, and by rule-of-thumb (as we use to speak), and not by Square and Rule", James Durham (1685).

Three centuries later and counting, thumbs still rule, and the use of rules-of-thumb still characterizes much of human activity, perhaps because human agents need to optimize also with respect to the costs of information gathering and processing, not to mention that they have deadlines to meet. In kernel density estimation, rule-of-thumb (ROT) bandwidths are ubiquitous. They are so ubiquitous in fact that Silverman's (1986) proposed ROT (which is derived specifically for a Normal kernel and a Normal reference density) is used even when other kernels are deployed (for example Guerre, Perrigne & Vuong 2000). While it is well known and well studied that data-driven or plug-in bandwidths deliver superior

Date: September 12, 2022.
Key words and phrases. Density Estimation, Kurtosis, Skewness, AMISE, Optimal Bandwidth.

asymptotic performance (Li & Racine 2007, Jeong, Im & Kim 2021), the prevalence of ROT bandwidths in applied work remains. For example, in some of the most prestigious academic journals, ROT bandwidths are the norm when presenting data: Boguth, Duchin & Simutin (2021, Figure 1) for the density of the time-series standard deviations of a measure of excess value for a panel of US firms; Brauner et al. (2021, Figure 4) for the density of the posterior median effectiveness across the sensitivity analysis of the estimated instantaneous reproduction number for COVID-19; Känzig (2021, Figure 2) for the density of oil price shocks following OPEC announcements; and Le Quéré et al. (2021, Figure 1) for the density of change in fossil CO_2 emissions in the five years since the adoption of the Paris Climate Agreement across a range of countries.

Such dominance is undoubtedly linked to the fact that the most popular statistical languages deploy ROT bandwidths as the default, a practice that strongly reflects the appeal of ROT bandwidths: they are simple to construct, easy to code by hand, and portable across datasets. Moreover, data-driven methods are known to produce bandwidths in finite samples which lead to undersmoothing, producing higher variance estimates (Loader 1999) which can be problematic for inference (tests of symmetry, independence, correct specification, etc.). To that end we ask, are there better, or more appealing, ROT bandwidths?

Several papers have focused on ROT bandwidth selection when the *kernel* is changed (for example Muller 1984, Abadir & Lawford 2004, Henderson & Parmeter 2012). But to our knowledge, beyond Terrell (1990), applied statisticians and econometricians have not investigated how changing the *reference density* impacts the corresponding ROT bandwidth and how this change might then impact density estimates which are estimated using such a ROT bandwidth. One may counter that this choice simply does not matter. We are cognizant of this argument, but believe it still deserves attention and formal clarification. Assessing how poorly one bandwidth may perform for a given setting when the reference density is incorrect is useful to study. Wand & Jones (1994) demonstrate that the classic Sheather & Jones (1991) plug-in estimator stabilizes once 2-3 iterations of the reference

density estimation have been undertaken. This suggests that for direct plug-in methods (which are in some sense data-determined), the reference choice (in this case Normal) does not impact the bandwidth. However, when a 0 iteration ROT bandwidth is calculated, little insight exists on this point.

In this paper, we construct ROT bandwidths that use as reference densities three common symmetric fat-tailed distributions: Logistic, Student's-t, and Laplace.[1] All three of these densities lead to simple closed form, easy to calculate, ROT bandwidth rules that can be used at a minimum to compare one's results with Silverman's ROT based on the assumption of a Normal reference density. We assess the performance of both the "standard" versions of these ROT bandwidths but also the adaptive variant, which is robust against the presence of outliers that may unduly affect the estimated variance in the data. We also construct a ROT when skewness is present in the data, using a specific version of the Asymmetric Laplace distribution as well as a general Gram-Charlier expansion that can be used to approximate the reference density using the empirical skewness and kurtosis present in the data.

A general finding that we present here is that reference densities with thicker tails (and slimmer bodies) than the Normal will produce smaller bandwidths than that from assuming a Normal density, which leads to less bias and more variance overall for the corresponding estimator. This is intuitive. The tails of a density require larger sample sizes to uncover the structure/shape of the tails, and the same holds for steeper slopes in the density graph. In lack of more data, the practitioner needs a smaller bandwidth to extract useful information from the tails, and to detect the steep slopes. But this does not mean that there exists a monotonic relation between a measure like excess kurtosis and the optimal bandwidth. It

[1]We use the term "fat-tailed" in its most general sense, to refer to distributions that have high/higher than the Normal kurtosis, but whose moments exist. We acknowledge that in certain scientific corners, the term is used to refer only to distributions that have no moments at all. That these densities may be considered common, consider that Katsiampa (2019) uses the Student's-t distribution to study volatility of Bitcoin and Ether cryptocurrencies, Rudy, Kutz & Brunton (2019) study a Lorenz equation, with measurement noise drawn from a Student's-t distribution, Tiwari, Raheem & Kang (2019) find that a skewed Student's-t distribution is the best fit for the residuals based on the model they estimate, while Sun, Yang & Gao (2019) use an Asymmetric Laplace mixture model to study grayscale image segmentation. Applied examples abound of using these distributions as a basis for the construction of the statistical model.

does not (as we show later). Instead, what appears decisive for the size of the bandwidth is a measure like the interquartile range (IQR). By representing how narrow or wide is the central part of a density graph, the IQR carries information about the "steepness" of the graph and hence the value of the 2^{nd} derivative of the density, which is what largely determines the optimal bandwidth in terms of 2^{nd}-order asymptotic mean integrated squared error (AMISE). This is perhaps why, for skewness, we find an unambiguous relationship: the optimal bandwidth decreases with skewness (for a given variance). This is because skewness, especially when it is pronounced, leads also to steep ascents or descents of the density graph.

The closest work to ours is Marron & Wand (1992) who examine exact mean integrated squared error for a range of Normal mixture distributions. The Normal mixture framework allows easy calculation of the optimal bandwidth (in theory) and so exact error rates can be calculated. Our work here differs because we focus on the asymptotic performance of the optimal bandwidth as a generic smoothing device rather than as the correct bandwidth.[2]

Finally, we also advocate for a model selection approach for bandwidth choice, rather than optimization of a statistical criterion, that leverages recent results of McCloud & Parmeter (2020). McCloud & Parmeter (2020) show how to calculate the number of effective parameters that a bandwidth imparts on the corresponding kernel density estimator. This allows for metrics, such as the Akaike Information Criteria (AIC), to be developed to select a kernel density estimator across a variety of bandwidths.

Beyond the theoretical results, we provide a detailed set of simulations comparing the various ROT bandwidths to assess their finite sample performance along with the model selection results. The main takeaway is that once we move away from Normality (either due to skewness or excess kurtosis) Silverman's Normal-Normal ROT stops being optimal (in a MSE sense), and instead one of the new ROTs becomes a better choice, often Asymmetric Laplace. To further investigate how useful these ROTs are, we also conduct a detailed set

[2]Moreover, Marron & Wand's (1992) approach would not yield usable ROT bandwidths as they require estimation of a large number of parameters of the Normal mixture, which for small samples may be estimated quite imprecisely.

of simulations using non-standard distributions. In these cases too, where we have various degrees of excess kurtosis, asymmetry, multimodality and peakedness,[3] the Asymmetric Laplace ROT appears best suited among the class of ROT bandwidths under consideration (evaluated through AMISE). Finally, we note that our ROT bandwidths, which can be implemented virtually at no cost, outperform in many cases data-driven and plug-in methods, which are not necessarily computationally cheap. So the ROT bandwidths we construct represent a clear improvement in the efficiency and the reliability of kernel density estimation. The Asymmetric Laplace ROT is perhaps the most serious contender to replace Silverman's ROT as the default ROT bandwidth for most data samples encountered in practice.

Our model selection results tell a somewhat different story. For underlying distributions that are either symmetric or thin-tailed, use of AIC suggests Silverman's traditional ROT is the most common winner. However, consistent with our AMISE findings, once considerable skewness or kurtosis enters the data, our new ROTs also appear as viable candidates from a model selection standpoint.

We complement our simulations with two empirical illustrations. The first looks at annual snowfall totals for Buffalo, New York. There is a debate in the literature as to whether the distribution contains one or three modes (Scott 1992) and our model selection criteria suggest the two bandwidths which produce a unimodal density. In the second illustration, we look at daily stock returns for GameStop. This video-games company stock became the battleground of an intergeneration clash between investors in the beginning of 2021. This density exhibits both excess kurtosis and skewness, making it an interesting exemplar for our proposed methods. The criterion suggest bandwidths that capture the peakedness in the estimated density without excess variability or spurious modes.

[3]Peakedness captures the probability of a given deviation around a point and is commonly associated with the height of the mode of a distribution. Here we define peakedness as in Birnbaum (1948): Let X_1 and X_2 be real random variables with real constants a_1 and a_2. X_1 is more peaked about a_1 than X_2 around a_2 if $P(|X_1 - a_1| \geq t) \leq P(|X_2 - a_2| \geq t)\ \forall t \geq 0$.

2. Optimal Bandwidths for the Canonical Kernel Density Estimator

Our focus here will be on the 2nd-order kernel density estimator of the density of x, $f(x)$, defined as (Wand & Jones 1994, Li & Racine 2007, Henderson & Parmeter 2015):

$$\widehat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right), \tag{1}$$

where $k(\cdot)$ is the kernel smoothing function and h is the smoothing parameter.[4]

ROT optimal bandwidths are derived as the minimizers of AMISE which stems from the sum of squared bias and variance of $\widehat{f}(x)$. For the derivations that follow, we will assume that the data sample X is independent and identically distributed (iid).

The bias of the 2nd-order kernel density estimator in (1) is (Henderson & Parmeter 2012)

$$\text{Bias}\left\{\widehat{f}(x)\right\} = E\left[\widehat{f}(x)\right] - f(x) \approx \frac{\kappa_2(k)}{2} h^2 f^{(2)}(x), \tag{2}$$

where $\kappa_j(k) = \int u^2 k(u) du$ is the j^{th} moment of the kernel and $f^{(2)}(x)$ is the second derivative of the unknown density. The variance is (Henderson & Parmeter 2012)

$$\text{Var}\left[\widehat{f}(x)\right] \approx \frac{f(x) R(k)}{nh}. \tag{3}$$

We use the notation $R(g(x)) = \int g(x)^2 dx$ and $R(g^{(2)}(x))$ denotes the roughness of a function.

By combining Equations (2) and (3), the AMISE for the 2nd-order kernel density estimator is (Henderson & Parmeter 2012):

$$\text{AMISE}\left\{\widehat{f}(x)\right\} = \frac{\kappa_2^2(k)}{4} \int \left(h^2 f^{(2)}(x)\right)^2 dx + \frac{R(k)}{nh}. \tag{4}$$

2.1. **Derivation of the Optimal Bandwidth.**

To derive a general form for the optimal bandwidth, we differentiate AMISE in Equation (4) with respect to h and set the derivative

[4] See Rosenblatt (1956) and Parzen (1962) for the original exposition on the kernel density estimator.

equal to zero:

$$h_{opt}^{k-f} = \left[\frac{R\left(k\right)}{\kappa_2^2(k)}\right]^{1/5} \left[R\left(f^{(2)}\right)\right]^{-1/5} n^{-1/5}. \tag{5}$$

Here our use of the $k-f$ superscript is nonstandard, but one that we believe will be useful with the various bandwidths we will be discussing later. Written as above, the expression clearly separates the components that multiplicatively determine the optimal bandwidth: the first depends on the choice of the kernel; the second, on the choice of the reference density; the third carries the effect on h_{opt}^{k-f} from the size of the sample.

We will focus on the Normal (Gaussian) kernel as our focus is on considering different reference densities. We also provide bandwidth factors also for the Epanechnikov kernel that will allow practitioners to implement a different kernel-reference density combination. For the Normal kernel we have that $\kappa_2(k) = 1$ and $R\left(k\right) = (2\sqrt{\pi})^{-1} \approx 0.282$. With these values the kernel-contributed factor in the optimal bandwidth expression Equation (5) becomes 0.776. Also, in order to make the expression representative of a parametric distribution family, we will use the roughness of a distribution with unitary variance $\sigma^2 = 1$, denoting it $R\left(f_1^{(2)}\right)$. By standard rules of density transformation, differentiation and integration, we have that $R\left(f_1^{(2)}\right) = \sigma^5 R\left(f^{(2)}\right)$. Consequently we obtain

$$h_{opt}^{N-f} = 0.776 \left[R\left(f_1^{(2)}\right)\right]^{-1/5} \sigma n^{-1/5}. \tag{6}$$

The adaptive variant for finite samples that is robust with respect to the sample standard deviation is (Silverman 1986, pg.47)

$$h_{opt}^{N-f} = 0.776 \left[R\left(f_1^{(2)}\right)\right]^{-1/5} A n^{-1/5}, \qquad A = \min\left\{\hat{\sigma}, \frac{\widehat{\mathrm{IQR}}}{\mathrm{IQR}(\sigma=1)}\right\}. \tag{7}$$

The adaptive variant compares two measures of dispersion: the sample standard deviation, and the mark-up increase of the sample interquartile range (IQR) over the IQR of the

assumed reference density with unitary standard deviation. This ROT uses the minimum of the two.

In both ROT expressions, the higher the roughness of the reference density, the smaller the optimal bandwidth. Terrell (1990) also noted this in his search to determine the maximal *smoothness* by finding the density with the smallest roughness, which turns out to be (see Terrell 1990, Theorem 1), for a 2^{nd}-order kernel, the Beta(4, 4) distribution, leading to $R\left(f_1^{(2)}\right) \approx 0.144$. In this regard, no density which has at least two derivatives can have roughness smaller than this distribution, which then implies an *upper* bound on the bandwidth for optimal smoothing (based on AMISE).

2.2. **Roughness and Excess Kurtosis.** The inverse monotonic relation between roughness and the optimal bandwidth obtained just above, makes us ask (with our minds towards the other direction, that of increasing roughness and narrower bandwidths), are there any characteristics of a distribution that will signal higher roughness?

The answer is "perhaps". For symmetric densities, the existence of positive excess kurtosis signals higher roughness than the Normal density. But the relation between positive excess kurtosis and roughness is not monotonic. Instead we find an (inverse) monotonic relation between the IQR and roughness. A smaller IQR leads to higher roughness, and if the IQRs are close, then higher excess kurtosis leads to higher roughness. But the more critical characteristic is the IQR, and we will see a specific example where a distribution with high excess kurtosis has smaller roughness than a distribution that has much smaller excess kurtosis but lower IQR. For skewed distributions, it appears that irrespective of how the IQR evolves, higher skewness (in absolute terms) increases roughness.

In order to understand the relationship between excess kurtosis and roughness, consider an infinitely supported continuous random variable with density $f_1(x)$ symmetric around

zero. Its excess kurtosis will equal

$$\gamma_2 = \mu_4 - 3 \implies \gamma_2 + 3 = \int_{-\infty}^{\infty} t^4 f_1(t) dt.$$

By twice applying integration by parts to the integral, we are led to the following alternative expression,

$$\gamma_2 + 3 = \frac{1}{30} \int_{-\infty}^{\infty} t^6 \cdot f_1^{(2)}(t) dt = \frac{1}{15} \int_{0}^{\infty} t^6 \cdot f_1^{(2)}(t) dt,$$

while the roughness of a symmetric (around zero) distribution is

$$R[f_1^{(2)}] = \int_{-\infty}^{\infty} \left[f_1^{(2)}(t) \right]^2 dt = 2 \int_{0}^{\infty} f_1^{(2)}(t) \cdot f_1^{(2)}(t) dt.$$

As we move towards the tails, the graph of a density with infinite support becomes necessarily convex, and so $f^{(2)}(x)$ will be positive, but smaller and smaller as we move along extreme values. But the contribution of the tails in the excess kurtosis coefficient will be the accumulation of products $x^6 \cdot f_1^{(2)}(x)$ where x^6 increases for higher values of x. On the other hand, the contribution of the tails to the roughness will be the accumulation of the products $f_1^{(2)}(x) \cdot f_1^{(2)}(x)$ that fall much faster in value as x increases. Distributions with fat tails/high excess kurtosis will have relatively higher values of $f_1^{(2)}(x)$ as we move towards the tails (because this puts the brakes on the reduction of the value of the density, hence a slower decay), and this indeed creates a tendency to have increased roughness also. But the excess kurtosis coefficient will increase much faster than roughness for the same series of $f_x^{(2)}(x)$ values, on account of the factor x^6.

On the other hand, closer to the central region of the distribution, say in the $(0, 1)$ interval, the factor x^6 will disproportionately shrink the contribution of $f_1^{(2)}(x)$ in the excess kurtosis coefficient. Now, if the density has a small IQR, this will produce a steeper initial decline to connect the central region with the tail region. But a "steeper decline" means high values for $f_1^{(2)}(x)$. Here, the value of $f_1^{(2)}(x)$ may be negative (for concave parts of the density

graph), in which case the contribution of the central region to γ_2 will be negative, while it will be positive for roughness. But even if $f_1^{(2)}(x)$ is everywhere positive (i.e the density is everywhere convex, such as the Laplace distribution for example), still the dampening effect of x^6 in the central region will make its contribution to excess kurtosis small.

The above discussion is in accord with the findings and the forceful argument of Westfall (2014), that excess kurtosis is a meaningful measure for tail extremity but not for "peakedness". In symmetric distributions, peakedness is monotonically connected to the IQR: the higher the peakedness, the smaller the IQR, and hence the higher the $f_1^{(2)}(x)$ values, and so the higher the roughness. So roughness relates primarily to peakedness (that proxies steepness) and then to tail fatness, while excess kurtosis relates primary to tail fatness and then, if at all, to peakedness. A connection between the two exists, but not one that would lead to a monotonic relation.

To see this in a more tangible way, consider another distribution with density g_1 (with g_1 sharing the same characteristic with f_1 as regards to its support: its mean and variance) and the difference of their excess kurtosis and of their roughness:

$$\begin{aligned} 15(\gamma_2(f_1) - \gamma_2(g_1)) &= \int_0^\infty t^6 \cdot [f_1^{(2)}(t) - g_1^{(2)}(t)]dt, \\ \frac{1}{2}[R(f_1^{(2)}) - R(g_1^{(2)})] &= \int_0^\infty \left(\left[f_1^{(2)}(t)\right]^2 - \left[g_1^{(2)}(t)\right]^2 \right) dt \\ &= \int_0^\infty \left(f_1^{(2)}(t) + g_1^{(2)}(t) \right) \left(f_1^{(2)}(t) - g_1^{(2)}(t) \right) dt \end{aligned}$$

Suppose that $\gamma_2(f_1) > \gamma_2(g_1)$. Then $\exists\, t_0 : t > t_0 \implies f_1^{(2)}(t) > g_1^{(2)}(t)$. f_1 will have fatter tails than g_1, and so eventually its 2^{nd} derivative should be positive and larger than the 2^{nd} derivative of g_1. Does this imply that we will get the same inequality for roughness?

Not necessarily. If both distributions have convex densities, then their 2^{nd} derivatives are everywhere positive. Closer to the origin we will necessarily have $g_1^{(2)}(x) > f_1^{(2)}(x)$ so the difference term in the roughness expression will be negative, tending to reduce the

overall value of the roughness difference.[5] Moreover, the contributions to the roughness for values of the variable near the tails will be small, and so, in principle, we could obtain a negative difference, $R(f_1^{(2)}) < R(g_1^{(2)})$. Considering alternative scenarios with regards to the concavity/convexity of the densities, leads to analogous ambiguous results.

2.3. **A Gram-Charlier approximation for Roughness.** The previous discussion is a warning against using without caution density approximations in order to assess the relationship between roughness and excess kurtosis. Case in point, the use of a 2^{nd}-order Gram-Charlier type A series expansion of the density of a general distribution:

$$f(x) \approx \phi\left(\frac{x-\mu}{\sigma}\right)\left[1+\frac{\gamma_1}{3!}He_3\left(\frac{x-\mu}{\sigma}\right)+\frac{\gamma_2}{4!}He_4\left(\frac{x-\mu}{\sigma}\right)\right], \tag{8}$$

where μ is the mean, σ is the standard deviation and γ_1 and γ_2 are the skewness and excess kurtosis coefficients respectively with $He_3(x) = x^3 - 3x$ and $He_4(x) = x^4 - 6x^2 + 3$ the 3^{rd} and 4^{th} order Hermite polynomials (as used in probability theory). Setting $\mu = 0, \sigma = 1$, computing the second derivative of the above expression, squaring it and integrating, the roughness of a distribution with unitary variance is approximated by[6]

$$R[f_1^{(2)}] \approx R[\phi^{(2)}] + \frac{105}{2^8\sqrt{\pi}}\gamma_1^2 + \frac{35}{2^7\sqrt{\pi}}\gamma_2 + \frac{1155}{2^{13}\sqrt{\pi}}\gamma_2^2. \tag{9}$$

The first correction term will be zero for symmetric distributions. From the above approximation, one would conclude that when excess kurtosis is positive, it has a positive monotonic relation with roughness. But this is not true, and as we have said, we will provide a specific counterexample shortly.

[5]The necessity of the reversal of the inequality comes from the fact that the integral of the 2^{nd} derivative of a density over the support is zero for symmetric densities with infinite support. So it is not possible that $f_1^{(2)}(x) > g_1^{(2)}(x)$ for the whole range.

[6]We note here that Dharmani (2015) independently derived (his equation (7) on page 6) our series skewness/kurtosis formula. Our derivation, together with the mathematical derivations in Sections 3, 4 and 5 are included in a Technical Appendix that is available from the authors upon request.

What the above approximation *is* valid for, is to conclude that excess kurtosis and skewness lead to higher roughness than that of the Normal distribution.[7] So, with distributions that have fatter tails than the Normal, or are skewed, the optimal bandwidth should be smaller than that produced by the Silverman Normal-Normal ROT. This motivates the construction of ROT bandwidths for fat-tailed and skewed distributions, a task to which we now turn.

3. Optimal Bandwidths Under Alternative Symmetric Distributions

We will consider three well known symmetric distributions that have positive excess kurtosis: Logistic, Laplace, and Student's t with 5 degrees of freedom. Their excess kurtosis coefficients are, respectively, 1.2 for the Logistic, 3 for the Laplace, and 6 for $t(5)$. We chose them to cover a range from mild to high excess kurtosis values. Our goal is to calculate h_{opt}^{N-f} for these distributions as reference densities, always for a 2^{nd}-order Gaussian kernel, and to determine if any of these may be suitable for general use as a ROT bandwidth, against the benchmark ROT bandwidth that is based on the Normal distribution for which we have roughness and optimal bandwidth (see Silverman 1986, pg. 45),

$$R\left(\phi^{(2)}\right) = \left(2\sqrt{\pi}\right)^{-1}(3/4) \approx 0.2115, \qquad h_{opt}^{N-N} = 1.059\sigma n^{-1/5}.$$

As has been discussed in the Introduction, we are unaware of a ROT bandwidth that attempts to account for the presence of excess kurtosis (or skewness) in the data. Silverman himself may have contributed to this by asserting that his Normal-Normal ROT optimal bandwidth, especially its "adaptive" variant, fared well even in the presence of skewness and excess kurtosis in the data, providing only summary indicative results to support this assertion (Silverman 1986, pp. 46-48). While asymptotically the differences will dissipate, it is instructive to learn if, and how, finite sample differences appear. This analysis is also

[7]We note that transforming the Gram-Charlier type A expansion into an Edgeworth one and including even four additional higher order terms as prioritized by the latter, would not affect the result for symmetric distributions, because these additional terms are multiplied by the 3rd moment which is zero.

useful more generally as the bandwidths we derive are optimal when the true density is in fact Logistic, Laplace, or Student's $t(5)$, respectively.

3.1. **Logistic.** The most widely used Logistic distribution has density

$$f(x) = \frac{1}{s}\frac{e^{-x/s}}{\left(1+e^{-x/s}\right)^2},\ s>0. \tag{10}$$

The roughness of the Logistic density is $R\left(f^{(2)}\right) = 1/42s^5$, its variance is $\sigma^2 = s^2\pi^2/3$. To standardize, we set $s = \sqrt{3}/\pi$. The standardized roughness is

$$R\left(f_1^{(2)}\right) = \frac{\pi^5}{14\cdot 3^{7/2}} \approx 0.467.$$

This is more than double the roughness of the standard Normal density (which is 0.2115). For the adaptive variant of the ROT bandwidth, the interquartile range of the Logistic distribution with unitary variance is IQR = 1.211.

3.2. **Laplace.** The Laplace distribution has density

$$f(x) = \frac{1}{2b}\exp\{-|x|/b\}.$$

The roughness of the standardized Laplace density is $R\left(f^{(2)}\right) = 1/4b^5$. The variance of the Laplace distribution is $\sigma^2 = 2b^2$, so to standardize it we need to set $b = 1/\sqrt{2}$ and we obtain

$$R\left(f_1^{(2)}\right) = \sqrt{2} \approx 1.414,$$

almost seven times higher than the roughness of the standard Normal, and three times higher than the roughness of the Logistic. The interquartile range for unitary variance is here IQR $= \ln 4/\sqrt{2} \approx 0.980$.

3.3. **Student's t.** Let p denote the (integer) degrees of freedom. The density of the Student's-t distribution can be written as

$$f(x) = A_0 \cdot (p + x^2)^{-(p+1)/2}, \quad A_0 = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{p^{(p+1)/2}}{\sqrt{p\pi}}. \tag{11}$$

The roughness of the distribution is

$$R\left(f^{(2)}\right) = \frac{(2p+3)!!}{(2p+8)!!} \cdot \left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}\right)^2 \cdot \frac{3\,(p+3)^2\,(p+1)^2}{p^{5/2}}.$$

Here $(2p+3)!! = 3 \cdot 5 \cdots (2p+3)$ and $(2p+8)!! = 2 \cdot 4 \cdots (2p+8)$. This basic version of the Student's-t distribution, has implicitly a scale parameter that is set equal to unity, and cannot have variance lower than 3 (for integer degrees of freedom). The variance of the $t(p)$ distribution is $\sigma^2 = p/(p-2)$. The appropriate scaling to obtain the roughness for unitary variance is to multiply the expression by $[p/(p-2)]^{5/2}$, and we obtain

$$R\left(f_1^{(2)}\right) = \frac{(2p+3)!!}{(2p+8)!!} \cdot \left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}\right)^2 \cdot \frac{3\,(p+3)^2\,(p+1)^2}{(p-2)^{5/2}}.$$

For $p = 5$, we obtain

$$R\left(f_1^{(2)} \mid p=5\right) = \frac{(13)!!}{(18)!!} \cdot \left(\frac{\Gamma(3)}{\Gamma(5/2)}\right)^2 \cdot \frac{3 \cdot 8^2 \cdot 6^2}{3^{5/2}} = 0.730.$$

But this is *smaller* than the roughness of Laplace that has *half* the excess kurtosis of $t(5)$ (for same variance). This is the counterexample that shows that the relation between excess kurtosis and roughness is not monotonic. The IQR of $t(5)$ (scaled to have unitary variance) is IQR ≈ 1.1257.

We collect in Table 1 our results, where we also include the value of each distribution at the mode, $f_1(\text{mode})$, and of course, the factors for the optimal bandwidths per case. For convenient reference, we have also included the optimal bandwidths when the Epanechnikov kernel is used, that has $\kappa_2(k) = 0.2$ and $R(k) = 0.6$, resulting in a kernel-contributed factor in the optimal bandwidth of 1.719.

TABLE 1. Optimal bandwidths for fat-tailed symmetric distributions. All distributions have unitary variance. Bandwidth factors are computed for the Normal kernel ($N-f$), and the Epanechnikov kernel ($E-f$) and need to be multiplied by $\hat{\sigma}n^{-1/5}$.

Distribution	γ_2	f_1(mode)	IQR	$R(f_1^{(2)})$	h_{opt}^{N-f}	h_{opt}^{E-f}
Normal	0	0.399	1.349	0.211	1.059	2.345
Logistic	1.20	0.453	1.211	0.467	0.903	2.001
Student's-$t(5)$	6	0.490	1.125	0.730	0.826	1.830
Laplace	3	0.707	0.980	1.414	0.724	1.604

The table shows clearly the monotonic relation between the density value at the peak and the IQR, and of the IQR with roughness (and so with the optimal bandwidth). Note also that the roughness of the Student's $t(5)$ is much higher than the roughness of the Logistic, even though their IQRs are relatively close; here the high excess kurtosis of $t(5)$ amplifies the increase in roughness. But between $t(5)$ and Laplace, the effect of the excess kurtosis is no match for the effect of the much increased peak and reduced IQR, which makes the roughness of the Laplace almost double that of the $t(5)$. Apart from that, we see that the optimal bandwidth narrows considerably as we move along the lines compared to the Normal-Normal ROT bandwidth: it is 15% smaller for the Logistic, 22% smaller for $t(5)$ and 32% smaller for the Laplace distribution.[8]

A way to understand why IQR will not necessarily decrease as excess kurtosis increases, is to consider a 4^{th}-order Cornish-Fisher expansion of the quantile function, to express the IQR of a distribution. There one would find that, first, as excess kurtosis increases, the IQR tends unambiguously to decrease from the expansion terms that include the excess kurtosis coefficient, but also, that the IQR tends to *increase* as the 6^{th} moment increases, which is also present. But the 6^{th} moment increases when the 4^{th} increases (which relates to excess kurtosis). If the structure of the distribution is such that the 6^{th} moment increases

[8]The roughness values can be used to compute optimal bandwidths with the use of other kernels if so desired, while the IQR values can be used to construct the adaptive variants of these ROT bandwidths.

disproportionately compared to the increase of the 4^{th}, then we may end up seeing higher excess kurtosis, but higher IQR also, and hence lower roughness.

Next we will see that an even stronger determinant of roughness than IQR is skewness, where even though IQR increases, the roughness increases because skewness increases.

4. An Optimal Bandwidth When the Distribution is Skewed

In this section, we construct a ROT optimal bandwidth for data samples that exhibit skewness, which is a frequent phenomenon. Unlike excess kurtosis, skewness is usually dependent on a shape parameter of the distribution even after we standardize the variance (the shape parameter allows symmetry when it takes some specific value). This means that in order to compute the ROT optimal bandwidth, we will need not only to estimate the standard deviation of the sample, but also the sample skewness coefficient, in order to recover an estimate for the shape parameter.

We will base our ROT on a version of the Asymmetric Laplace distribution, which is generated as follows: Consider two independent Exponential random variables Z_1 and Z_2 with scale parameters $\sigma_1 = \theta/\tau$, and $\sigma_2 = \theta/(1-\tau)$, respectively, for $\theta > 0$, and $\tau \in (0,1)$. Then the distribution of their difference $X = Z_1 - Z_2$ is[9]

$$f(x) = \frac{\tau(1-\tau)}{\theta} \cdot \begin{cases} \exp\left\{\frac{1-\tau}{\theta}x\right\} & x \leq 0 \\ \exp\left\{-\frac{\tau}{\theta}x\right\} & x > 0. \end{cases} \tag{12}$$

For $\tau = 1/2$, we recover the Laplace distribution with scale parameter $b = 2\theta$. The mean and variance of this distribution are

$$E(X) = \frac{\theta(1-2\tau)}{\tau(1-\tau)}, \quad \text{Var}(X) = \frac{\theta^2[\tau^2 + (1-\tau)^2]}{\tau^2(1-\tau)^2},$$

[9]This distribution has been considered by Poiraud-Casanova & Thomas-Agnan (2000) in relation to quantile regression and estimation.

while its skewness coefficient is

$$\gamma_1 = 2\frac{(1-\tau)^3 - \tau^3}{[\tau^2 + (1-\tau)^2]^{3/2}} \in (-2, 2). \tag{13}$$

By computing the sample skewness, we can solve the above for τ, and use it to compute the roughness and the optimal bandwidth. Note that if $\widehat{\tau} \neq 1/2$, the density we will use produces a non-zero mean, but this will not affect the roughness or the optimal bandwidth.

For the distribution standardized to have unitary variance, the roughness is

$$R\left(f_1^{(2)}\right) = \frac{(1-\tau)^3 + \tau^3}{2\tau^3(1-\tau)^3} \cdot \left[\tau^2 + (1-\tau)^2\right]^{5/2}.$$

We see that τ and $(1-\tau)$ are exchangeable in the above expression, meaning that $R\left(f_1^{(2)} \mid \tau\right) = R\left(f_1^{(2)} \mid 1-\tau\right)$.

In Table 2, we present the values for skewness, τ, mode, IQR, standardized roughness, and optimal bandwidths for values of skewness in $[0, 1.9]$.

Given that τ and $(1-\tau)$ are exchangeable up to a sign change (so the absolute value of skewness is symmetric around $\tau = 0.5$) in the expressions for the skewness, it follows that $\tau(-\gamma_1) = 1 - \tau(\gamma_1)$. So we need only present the values for positive skewness. If our sample has, say, skewness $\widehat{\gamma}_1 = -0.9$, the corresponding value for the distribution parameter will be $\widehat{\tau} = 1 - 0.387 = 0.613$. The roughness and the optimal bandwidth depend only on the absolute value of the skewness.

We see in Table 2 that as the strength of skewness increases, so does the peak, the IQR, *but also the roughness*. Intuitively, skewness leads to the long tail "stretching" relatively more than how much the short tail contracts, leading to higher IQR. It also leads to a steeper slope near the origin that leads to increased roughness. Skewness appears to be the stronger of the various characteristics of a distribution that we have considered, as regards to its unambiguous effect on roughness, and hence on the optimal bandwidth. In the specific

TABLE 2. Optimal bandwidths for the Asymmetric Laplace distribution. All distributions have unitary variance. Bandwidth factors are computed for the Normal kernel ($N-f$), and the Epanechnikov kernel ($E-f$) and need to be multiplied by $\hat{\sigma}n^{-1/5}$.

γ_1	τ	f_1(mode)	IQR	$R(f_1^{(2)})$	h_{opt}^{N-f}	h_{opt}^{E-f}
0.0	0.500	0.707	0.980	1.414	0.724	1.604
0.1	0.488	0.707	0.980	1.421	0.723	1.602
0.2	0.476	0.708	0.981	1.442	0.721	1.598
0.3	0.464	0.709	0.981	1.478	0.718	1.590
0.4	0.452	0.710	0.982	1.529	0.713	1.579
0.5	0.440	0.712	0.983	1.597	0.707	1.565
0.6	0.427	0.715	0.985	1.692	0.699	1.547
0.7	0.415	0.717	0.987	1.802	0.690	1.528
0.8	0.401	0.721	0.989	1.962	0.678	1.502
0.9	0.387	0.725	0.992	2.162	0.665	1.473
1.0	0.373	0.730	0.995	2.411	0.651	1.442
1.1	0.358	0.735	0.999	2.744	0.634	1.405
1.2	0.342	0.742	1.003	3.197	0.615	1.362
1.3	0.325	0.749	1.009	3.821	0.594	1.315
1.4	0.306	0.758	1.016	4.756	0.568	1.258
1.5	0.285	0.770	1.025	6.205	0.539	1.193
1.6	0.262	0.783	1.036	8.555	0.505	1.119
1.7	0.234	0.801	1.051	13.229	0.463	1.026
1.8	0.200	0.825	1.066	24.204	0.410	0.909
1.9	0.150	0.863	1.082	71.362	0.330	0.732

reference distribution, skewness adds roughness on top of the roughness of the symmetric Laplace distribution.

5. Efficiency of Alternative Parametric Families

Beyond the construction of these optimal bandwidths, we can also investigate how well the estimation of the density compares across alternative densities when alternative "optimal" bandwidths are used.

5.1. **Alternative ROT Efficiency.** Plugging the optimal bandwidth expression equation (5) into the AMISE expression (4), the minimized AMISE under correct specification of the

reference density is

$$\text{AMISE}_{opt}^{k-f} = \frac{5}{4} R(k)^{4/5} (\kappa_2(k))^{2/5} R\left(f_1^{(2)}\right)^{1/5} \sigma^{-1} n^{-4/5}. \tag{14}$$

Using the Normal kernel values produces

$$\text{AMISE}_{opt}^{N-f} = 0.36 R\left(f_1^{(2)}\right)^{1/5} \sigma^{-1} n^{-4/5}. \tag{15}$$

It is clear that the higher the roughness, the higher the *minimized* AMISE.

Since roughness exceeds the Normal roughness when we have excess kurtosis, we get that distributions with slimmer bodies and thicker tails are harder to estimate from the standpoint of mean squared error. The same holds for skewed distributions compared to symmetric ones, especially to their symmetric version in their own distribution family.

But this does not mean that going for the lower AMISE assuming correct specification, is beneficial if, in this way, we misspecify the reference density: misspecification leads to even higher AMISE, as we show next.

Looking at losses of efficiency due to misspecification of the reference density, are in some sense more informative than similar efficiency losses predicated on kernel choice (Epanechnikov 1969) since this is purely at the user's discretion. The AMISE formula when an alternative optimized, yet misspecified, bandwidth is used (say the Laplace optimal bandwidth for data from a Normal density), becomes (see Technical Appendix)

$$\text{AMISE}_{k-g}^{k-f} = R(k)^{4/5} \kappa_2^{2/5} \left(R\left(g_1^{(2)}\right)^{1/5} + \frac{R\left(f_1^{(2)}\right)}{4R\left(g_1^{(2)}\right)^{4/5}} \right) \sigma^{-1} n^{-4/5}. \tag{16}$$

TABLE 3. Efficiency of parametric family optimal bandwidths for potential misspecification.

		Assumed distribution			
Roughness	Correct	Normal	Logistic	Student's-$t(5)$	Laplace
0.211	Normal	1	1.044	1.100	1.214
0.467	Logistic	1.060	1	1.015	1.081
0.730	Student's-$t(5)$	1.164	1.017	1	1.031
1.414	Laplace	1.463	1.126	1.040	1

We are interested in

$$\begin{aligned}
\text{Eff}_{k-g}^{k-f} =& \frac{\text{AMISE}_{k-g}^{k-f}}{\text{AMISE}_{opt}^{k-f}} = \frac{R\left(g_1^{(2)}\right)^{1/5} + \frac{R\left(f_1^{(2)}\right)}{4R\left(g_1^{(2)}\right)^{4/5}}}{\frac{5}{4}R\left(f_1^{(2)}\right)^{1/5}} \\
=& \frac{4}{5}\left[\frac{R\left(g_1^{(2)}\right)}{R\left(f_1^{(2)}\right)}\right]^{1/5} + \frac{1}{5}\left[\frac{R\left(f_1^{(2)}\right)}{R\left(g_1^{(2)}\right)}\right]^{4/5} \\
=& \frac{4}{5}R_1^{1/5} + \frac{1}{5}R_1^{-4/5}, \qquad (17)
\end{aligned}$$

where we have defined the relative roughness ratio $R_1 = R\left(g_1^{(2)}\right)/R\left(f_1^{(2)}\right)$, where g_1 is the assumed density and f_1 the true density. We mention here that the comparison of AMISE between different choices of the reference density is independent of the kernel and hence whether we were to use a Normal kernel or the Epanechnikov kernel (as an example), would have no effect on our results. It is easy to determine that the relative efficiency is minimized when $g_1^{(2)} = f_1^{(2)}$ and takes the value 1, while $\text{Eff}_{k-g}^{k-f} > 1$ whenever $g_1^{(2)} \neq f_1^{(2)}$. But the effects of misspecification are not symmetric, and this can be seen in Table 3 where we tabulate the efficiency ratios of the distributions we are considering (the variance in all cases is standardized to unity).

What we observe in Table 3 is that *oversmoothing is worse than undersmoothing, in terms of efficiency loss.* For example, if the true density is Normal, but we use the Laplace ROT bandwidth (undersmoothing), the efficiency ratio is 1.214. But when the true density is Laplace and we use the Normal ROT bandwidth (oversmoothing), the efficiency ratio climbs

to 1.463. This holds for all entries in the Table. This provides formal motivation to not use the Normal ROT bandwidth when the data exhibit excess kurtosis or narrow IQR, but a ROT based on a density with higher roughness, even if we end up undersmoothing. The informal motivation for undersmoothing comes from Silverman himself (Silverman 1986, pg.43), who sensibly observed that it is better to undersmooth than oversmooth because the eye of the observer can smooth much more easily than "unsmooth".[10] In fact mentally "unsmoothing" is practically impossible because there is no information left in the oversmoothed density to guide such "unsmoothing". This problem is similar to the case of a linear against a non-linear graph: a non-linear graph guides the eye towards its linear ("smoothed") version, but the same linear graph may be the result of approximating any one of very different non-linear graphs. The counterargument in favor of oversmoothing is, to quote Terrell (1990, pg. 472), "An undersmoothed density estimate tends to display features such as asymmetries and multiple modes that could have come about by chance. By using the most smoothing that is compatible with the scale of the problem, we tend to eliminate accidental features".

5.2. **The Effects of Estimating the Standard Deviation.** We close this section by noting that all its derivations and computations have been conditional on the *sample* standard deviation (or on the assumption that the true standard deviation is known a priori). Essentially we were elaborating on the sample-specific AMISE. A further refinement would be to use the sample standard deviation in the optimal bandwidth expression (because this is how we would be able to actually compute it), but to keep the true standard deviation to accompany the roughness of the true standardized distribution in the correctly specified AMISE, the misspecified AMISE and the relative efficiency expressions, because this is what characterizes the true data generating process. In such a case, Equation (16) would become

[10]That said, we are aware that viewers are consumers and consumers may be lazy – they would prefer to bear the risk of being misled by viewing a smoothed graph, rather than do the mental effort to smooth a ragged one.

(see Technical Appendix):

$$\text{AMISE}_{k-g}^{k-f} = R(k)^{4/5}\kappa_2^{2/5}\left(R\left(g_1^{(2)}\right)^{1/5}\left(\frac{\sigma}{\widehat{\sigma}}\right) + \left(\frac{\widehat{\sigma}}{\sigma}\right)^4 \frac{R\left(f_1^{(2)}\right)}{4R\left(g_1^{(2)}\right)^{4/5}}\right)\sigma^{-1}n^{-4/5}. \tag{18}$$

In this way, one would take into account also the consequences from the inaccuracy in the estimation of σ. Minimizing this AMISE expression with respect to $(\widehat{\sigma}/\sigma)$, we find that it has an optimal value, $(\widehat{\sigma}/\sigma)^* = \left[R\left(g_1^{(2)}\right)/R\left(f_1^{(2)}\right)\right]^{1/5}$, *for which we recover the minimum AMISE under correct specification.* While this is not feasible as we don't know the true standard deviation in the first place, it gives us the following qualitative results: if in reality we are over-smoothing, $R\left(g_1^{(2)}\right) < R\left(f_1^{(2)}\right)$, we would want the sample standard deviation to under-estimate the true value. But if we are under-smoothing, $R\left(g_1^{(2)}\right) > R\left(f_1^{(2)}\right)$, we would want the sample standard deviation to over-estimate the true value, in order to mitigate the loss of efficiency in AMISE terms. In some cases this could be useful guidance in deciding whether to correct the sample standard deviation for bias or not.

5.3. **Model Selection Across Bandwidths.** Rather than select h such that AMISE is minimized, we could turn our attention to a model selection approach. Typically, model selection would seek to optimize some pre-specified criterion and then penalize this based on the complexity of the model, in most cases the number of parameters. As nonparametric kernel density estimators do not have "parameters", it is not obvious how such a correction would proceed. Recently, McCloud & Parmeter (2020) have demonstrated how to calculate the effective parameters from a kernel density estimator for a given bandwidth.[11] This opens up the use of traditional model selection criteria for use with alternative bandwidths.

[11]See also McCloud & Parmeter (2021)

While there are many model selection criteria, we focus on AIC. AIC, for a given bandwidth, selects the model which has the smallest value of

$$2tr(H)/n - 2 * \sum_{i=1}^{n} \ln\left(\widehat{f}(x_i)\right)$$

where H is the $n \times n$ hat matrix for the kernel density estimator with i,j element defined as (McCloud & Parmeter 2020):

$$\frac{K\left(\frac{x_j - x_i}{h}\right)}{\sum_{\ell=1}^{n} K\left(\frac{x_\ell - x_i}{h}\right)}.$$

H has useful geometric properties. It is symmetric with entries bounded between 0 and 1, and rows that sum to 1. The use of $tr(H)$ to calculate the number of parameters is directly linked to the size of the bandwidth; in effect as $h \to \infty$ the number of parameters goes to 1 and for $h \to 0$ the number of parameters goes to n.

Alternative penalities could be used beyond AIC. For instance, in our Online Appendix, we also consider corrected AIC (AIC_c), which was suggested by Loader (1999), which uses $\frac{2(tr(H)+1)}{n-tr(H)-2}$. We further consider GCV ($2\ln(1 - tr(H)/n)$), Rice's T ($\ln(1 - 2tr(H)/n)$) and BIC ($tr(H)\ln(n)$).

6. Simulated Performance of Alternative ROTs

We consider two distinct sets of simulations. First, we generate data from each of the parent distributions used to construct the optimal bandwidths and determine how well each performs when used (potentially) erroneously for a different distribution. We consider 1,000 Monte Carlo simulations for samples sizes $n \in \{50, 100, 200, 400, 800\}$. For each simulation we estimate the kernel density estimator using each of the proposed ROT bandwidths over a grid of 100 points. Our second set of simulations mimics the first, except we chose distributions from the suite of densities discussed in Marron & Wand (1992). Specifically, we focus attention on the Skewed, Kurtotic, Bimodal, Separated Bimodal, Asymmetric Bimodal, Claw

and Asymmetric Claw densities.[12] These densities take an array of non-standard shapes relative to the more familiar parent densities used to derive the ROT bandwidths. We use the command `npudensbw()` in the `np` package (Hayfield & Racine 2008) in `R` to calculate the kernel density estimates for all simulations along with a 2^{nd}-order Gaussian kernel for all computations. We compare each bandwidth selection method via mean squared error (Tables 4–5) and via our model selection criteria (Tables 6–7).

6.1. **Performance via Mean Square Error.**

6.1.1. *Known Parent Distributions.* Table 4 presents the mean (over the 1,000 simulations) of the mean squared error evaluated over the 100 grid points. There are several key takeaways from the results. First, the Normal ROT does well under symmetry and for mildly slim IQR up to the Student's-$t(5)$. When the data are distributed as Logistic, the Normal ROT bandwidths loses (by very little) against the correct ROT. We see distinct losses of the Normal ROT bandwidth when the data are Laplace or when skewness is present. Once the data are kurtotic, we see that the Normal ROT is inferior to the Asymmetric Laplace ROT (as expected).

We also considered adaptive versions of all of our ROT bandwidths. As the results are generally worse than the non-adaptive ones, those results are not presented here (but are available upon request). The main takeaway for applied work is that it is feasible to use the Normal ROT for data which are symmetric and have a mildly slim IQR, but one should consider the use of the Laplace or Asymmetric Laplace ROT when the data have skew or high peakedness.

6.1.2. *Non-standard Distributions.* Table 5 presents the mean MSE from these simulations. Here we again consider $n \in \{50, 100, 200, 400, 800\}$. We immediately see that for all of the

[12]We note here that none of our ROT bandwidth consider densities with more than a single peak. In practice, it may make sense to determine whether the underlying density is multimodal (e.g., see Hall, Minnotte & Zhang (2004), Henderson, Parmeter & Russell (2008) and/or Minnotte (2010)).

TABLE 4. Mean MSE for parent distributions of proposed ROT procedures across 1,000 Monte Carlo simulations.

Gaussian	Normal	Logistic	Student's-t	Laplace	Asym Laplace	winner
n=50	0.001975	0.002217	0.002423	0.002813	0.002849	Normal
n=100	0.001104	0.001208	0.001306	0.001496	0.001506	Normal
n=200	0.000599	0.000648	0.000699	0.000799	0.000801	Normal
n=400	0.000341	0.000365	0.000391	0.000443	0.000444	Normal
n=800	0.000189	0.000203	0.000217	0.000246	0.000246	Normal
Logistic						
n=50	0.002185	0.002228	0.002343	0.002617	0.002680	Normal
n=100	0.001238	0.001234	0.001283	0.001413	0.001437	Log
n=200	0.000646	0.000635	0.000658	0.000722	0.000729	Log
n=400	0.000349	0.000343	0.000356	0.000391	0.000392	Log
n=800	0.000183	0.000178	0.000183	0.000200	0.000201	Log
t5						
n=50	0.005901	0.006889	0.007446	0.008272	0.008647	Normal
n=100	0.004828	0.005516	0.005884	0.006406	0.006589	Normal
n=200	0.004036	0.004494	0.004729	0.005052	0.005139	Normal
n=400	0.003483	0.003793	0.003947	0.004151	0.004192	Normal
n=800	0.003130	0.003345	0.003448	0.003583	0.003603	Normal
Laplace						
n=50	0.004449	0.003894	0.003723	0.003632	0.003714	Lap
n=100	0.002701	0.002255	0.002095	0.001960	0.001956	aLap
n=200	0.001675	0.001351	0.001231	0.001121	0.001111	aLap
n=400	0.001014	0.000789	0.000703	0.000619	0.000614	aLap
n=800	0.000644	0.000489	0.000429	0.000368	0.000365	aLap
Asym Laplace						
n=50	0.002795	0.002425	0.002301	0.002215	0.002265	Lap
n=100	0.001721	0.001418	0.001303	0.001197	0.001191	aLap
n=200	0.001058	0.000844	0.000760	0.000678	0.000655	aLap
n=400	0.000670	0.000515	0.000454	0.000391	0.000364	aLap
n=800	0.000424	0.000317	0.000275	0.000231	0.000211	aLap

proposed ROT bandwidths under study, the mean MSE decreases by nearly 50% as the sample size doubles. This is consistent with the MSE consistency of the kernel density estimator in general. A noticeable outcome of these simulations is that the proposed Asymmetric Laplace ROT (aLap) has the most "wins" relative to the other ROT proposals. As our theory of the optimal bandwidth dictated, in the presence of extreme skewness or kurtosis, a superior ROT bandwidth exists, namely the Asymmetric Laplace.

TABLE 5. Mean MSE over 1,000 Monte Carlo replications for various densities found in Marron & Wand (1992).

Skewed	Normal	Logistic	Student's-t	Laplace	Asym Laplace	winner
n=50	0.002542	0.002569	0.002687	0.002979	0.003203	Normal
n=100	0.001436	0.001414	0.001464	0.001606	0.001700	Log
n=200	0.000785	0.000757	0.000779	0.000849	0.000893	Log
n=400	0.000429	0.000410	0.000420	0.000456	0.000479	Log
n=800	0.000247	0.000231	0.000235	0.000251	0.000262	Log
Kurtotic						
n=50	0.045156	0.040661	0.038105	0.034318	0.033627	aLap
n=100	0.035636	0.031417	0.029028	0.025514	0.025161	aLap
n=200	0.028707	0.024719	0.022493	0.019272	0.019100	aLap
n=400	0.022723	0.019004	0.016972	0.014100	0.014027	aLap
n=800	0.018000	0.014621	0.012824	0.010350	0.010319	aLap
Bimodal						
n=50	0.002514	0.002423	0.002441	0.002564	0.002576	Log
n=100	0.001609	0.001475	0.001457	0.001503	0.001505	Log
n=200	0.001023	0.000892	0.000860	0.000862	0.000862	Lap
n=400	0.000633	0.000532	0.000506	0.000499	0.000499	Lap
n=800	0.000375	0.000304	0.000284	0.000276	0.000276	Lap
Sep Bimodal						
n=50	0.011279	0.008632	0.007383	0.005888	0.005841	aLap
n=100	0.008237	0.006026	0.005034	0.003882	0.003863	aLap
n=200	0.005813	0.004066	0.003314	0.002467	0.002460	aLap
n=400	0.003946	0.002641	0.002102	0.001512	0.001510	aLap
n=800	0.002630	0.001699	0.001327	0.000931	0.000930	aLap
Asym Bimodal						
n=50	0.003385	0.003294	0.003303	0.003409	0.003440	Log
n=100	0.002220	0.002013	0.001947	0.001922	0.001926	Lap
n=200	0.001476	0.001261	0.001183	0.001122	0.001120	aLap
n=400	0.000960	0.000774	0.000706	0.000646	0.000644	aLap
n=800	0.000612	0.000473	0.000422	0.000377	0.000375	aLap
Claw						
n=50	0.014575	0.014655	0.014767	0.014988	0.015027	Normal
n=100	0.011810	0.011695	0.011646	0.011541	0.011536	aLap
n=200	0.009855	0.009639	0.009483	0.009141	0.009130	aLap
n=400	0.008440	0.008104	0.007829	0.007265	0.007254	aLap
n=800	0.007391	0.006872	0.006465	0.005711	0.005706	aLap
Asym Claw						
n=50	0.006163	0.006209	0.006239	0.006273	0.006285	Normal
n=100	0.005235	0.005106	0.005009	0.004837	0.004818	aLap
n=200	0.004500	0.004224	0.004041	0.003756	0.003721	aLap
n=400	0.003838	0.003465	0.003243	0.002925	0.002889	aLap
n=800	0.003149	0.002744	0.002524	0.002225	0.002192	aLap

6.1.3. *Versus Data Driven Bandwidths.* Although our primary interest is in comparison to existing ROT bandwidths, it is also of interest to see comparisons to data driven methods. We consider the same set of simulations run for Table 4 in Table A1 as well as for Table 5 in Tables A2 and A3.[13] Recall that the former set was for our standard distributions (e.g., Gaussian) while the latter set was for the non-standard distributions (Marron & Wand 1992). We consider the two most popular methods, least-squares cross-validation (LSCV) and direct plug-in (Sheather & Jones (1991), labeled SJ). For completeness, we also report bandwidths obtained from our Gram-Charlier (GC) approximation via Equation 9 (replacing the unknown moments by their sample estimates).[14]

The results from Table A1 look nearly identical to those in Table 4. The three additional columns are added, but the winner still continues to be a particular ROT bandwidth. This is to be expected as these distributions are well behaved and data driven methods often undersmooth. There are seven different distributions in Tables A2 and A3. The SJ bandwidths dominate in two of them (Kurtotic and Separate Bimodal). LSCV bandwidths dominate for larger samples in another two scenarios (Claw and Asymmetric Claw). If there are any surprising results, it is for the Kurtotic distribution. Here we expected one of the ROT bandwidths to have picked up the fat tails better than a plug-in bandwidth. The Separated Bimodal is understandable as none of the ROT bandwidths account for bimodality (especially such separated modes). Similarly, the Claw and Asymmetric Claw are highly variable densities with multiple modes and LSCV bandwidths tend to undersmooth and hence mimic the data better.

Overall, we see that our ROT bandwidths often outperform data driven bandwidth procedures. Typically the data driven methods show improvements when the underlying density is multimodal. It may be useful to consider extending our bandwidths to account for densities that are not single peaked.

[13]Tables A1-A3 are available in the Online Appendix.

[14]In all the simulations we ran, there was only one case, for a single sample size whereby the GC bandwidth performed best.

6.2. **Performance via Model Selection Criteria.** The results above are useful, but these are essentially oracle results because the underlying distribution is known. In practice, we will not know the true underlying distributions. Therefore, we consider a method for selection of reference (or data driven) bandwidth selection when the true underlying density is unknown. In this subsection, we look at the same sets of densities as before, but via the lens of Section 5.3. Tables 6 and 7 are analogus to Tables 4 and 5, respectively. Each value in the table represents the percentage of time the AIC criterion picks a given bandwidth for a given sample size for a given underlying density amongst the set of candidate bandwidths.[15]

6.2.1. *Known Parent Distributions.* Table 6 presents the percentage of time (over the 1,000 simulations) that the model selection criterion (AIC) picks a given bandwidth. These results are somewhat striking. Regardless of the underlying distribution, the model selection criterion picks the Normal ROT bandwidth. The Online Appendix shows this to be the case for each of the model selection criteria. The BIC criterion picks the Normal ROT nearly 100% of the time for $n \geq 100$.

6.2.2. *Non-standard Distributions.* For the (Marron & Wand 1992) distributions, Table 7 shows a very different picture from what we saw with the standard distributions. AIC also picks the Normal ROT for the Skewed distributon. The Kurtotic, Separated Bimodal, Claw (for $n \geq 400$) and Asymmetric Claw (for $n \geq 200$) all point to the Asymmetric Laplace ROT bandwidth. The Bimodal (for $n \geq 200$) and Asymmetric Bimodal (for $n \geq 400$) point to the Logistic ROT bandwidth. The Online Appendix shows nearly identical qualitative results for AIC_c, GCV and RICE model selection criteria. BIC again picks Asymmetric Laplace for the Separated Bimodal distribution, but typically the Normal ROT for the other distributions (the exception being for $n = 800$ with the Kurtotic distribution).

[15]The full set of simulations for all model selection criteria are available in the Online Appendix. Most of the model selection criteria produce very similar results.

TABLE 6. Percentage of time AIC picked the ROT procedure across 1,000 Monte Carlo simulations.

Gaussian	Normal	Logistic	Student's-t	Laplace	Asym Laplace	winner
n=50	0.930	0.034	0.010	0.007	0.019	Normal
n=100	0.974	0.014	0.006	0.001	0.005	Normal
n=200	0.987	0.006	0.002	0.001	0.004	Normal
n=400	0.994	0.002	0.002	0.000	0.002	Normal
n=800	1.000	0.000	0.000	0.000	0.000	Normal
Logistic						
n=50	0.897	0.039	0.019	0.019	0.026	Normal
n=100	0.963	0.018	0.006	0.004	0.009	Normal
n=200	0.976	0.018	0.005	0.000	0.001	Normal
n=400	0.991	0.004	0.001	0.002	0.002	Normal
n=800	0.994	0.005	0.000	0.001	0.000	Normal
t_5						
n=50	0.851	0.045	0.044	0.015	0.045	Normal
n=100	0.896	0.056	0.027	0.013	0.008	Normal
n=200	0.935	0.036	0.016	0.004	0.009	Normal
n=400	0.964	0.023	0.008	0.004	0.001	Normal
n=800	0.984	0.013	0.002	0.001	0.000	Normal
Laplace						
n=50	0.775	0.087	0.048	0.036	0.054	Normal
n=100	0.794	0.094	0.046	0.027	0.039	Normal
n=200	0.825	0.103	0.033	0.020	0.019	Normal
n=400	0.850	0.101	0.033	0.009	0.007	Normal
n=800	0.846	0.120	0.028	0.004	0.002	Normal
Asym Laplace						
n=50	0.710	0.092	0.065	0.067	0.066	Normal
n=100	0.694	0.133	0.070	0.048	0.055	Normal
n=200	0.715	0.176	0.054	0.036	0.019	Normal
n=400	0.675	0.226	0.065	0.030	0.004	Normal
n=800	0.665	0.254	0.068	0.012	0.001	Normal

6.2.3. *Versus Data-Driven Bandwidths.* In practice practicioners have access to data driven methods as well. We created analogous tables to those in Tables A1–A3, but for the model selection criteria. For example, Table B11 in the Online Appendix, gives the percentage of time the AIC criterion picked a given bandwidth procedure over the 1,000 simulations for the standard distributions by including both ROT and data driven methods (GC, LSCV and SJ). For an underlying Gaussian distribution, AIC picked LSCV for each sample size.

TABLE 7. Percentage of time AIC picked the ROT procedure across 1,000 Monte Carlo simulations - Marron and Wand (1992).

Skewed	Normal	Logistic	Student's-t	Laplace	Asym Laplace	winner
n=50	0.866	0.059	0.024	0.018	0.033	Normal
n=100	0.919	0.046	0.018	0.006	0.011	Normal
n=200	0.952	0.031	0.009	0.004	0.004	Normal
n=400	0.968	0.027	0.004	0.001	0.000	Normal
n=800	0.990	0.009	0.001	0.000	0.000	Normal
Kurtotic						
n=50	0.373	0.066	0.063	0.061	0.437	aLap
n=100	0.133	0.084	0.075	0.092	0.616	aLap
n=200	0.006	0.014	0.017	0.084	0.879	aLap
n=400	0.000	0.000	0.000	0.081	0.919	aLap
n=800	0.000	0.000	0.000	0.121	0.879	aLap
Bimodal						
n=50	0.582	0.197	0.106	0.045	0.070	Normal
n=100	0.481	0.315	0.126	0.033	0.045	Normal
n=200	0.357	0.417	0.186	0.020	0.020	Log
n=400	0.213	0.539	0.196	0.030	0.022	Log
n=800	0.114	0.644	0.221	0.010	0.011	Log
Sep Bimodal						
n=50	0.000	0.000	0.001	0.072	0.927	aLap
n=100	0.000	0.000	0.000	0.117	0.883	aLap
n=200	0.000	0.000	0.000	0.154	0.846	aLap
n=400	0.000	0.000	0.000	0.195	0.805	aLap
n=800	0.000	0.000	0.000	0.301	0.699	aLap
Asym Bimodal						
n=50	0.567	0.172	0.099	0.059	0.103	Normal
n=100	0.488	0.225	0.140	0.054	0.093	Normal
n=200	0.335	0.325	0.197	0.078	0.065	Normal
n=400	0.172	0.402	0.279	0.100	0.047	Log
n=800	0.070	0.430	0.369	0.092	0.039	Log
Claw						
n=50	0.768	0.076	0.039	0.026	0.091	Normal
n=100	0.812	0.052	0.027	0.020	0.089	Normal
n=200	0.622	0.062	0.023	0.027	0.266	Normal
n=400	0.150	0.004	0.005	0.110	0.731	aLap
n=800	0.000	0.000	0.000	0.174	0.826	aLap
Asym Claw						
n=50	0.762	0.077	0.031	0.025	0.105	Normal
n=100	0.591	0.080	0.044	0.026	0.259	Normal
n=200	0.167	0.072	0.068	0.064	0.629	aLap
n=400	0.009	0.018	0.023	0.033	0.917	aLap
n=800	0.000	0.000	0.000	0.000	1.000	aLap

It also picked LSCV for $n = 50$ for an underlying Logistic or t_5 distribution. For every other case, the Normal ROT produced the lowest value for AIC. The results were nearly identical for the the remaining selection criteria.

For the (Marron & Wand 1992) data, there is substantial heterogeneity. For example, Table B16 in the Online Appendix gives the AIC criterion for the same set of (Marron & Wand 1992) distributions. The Skewed distribution again suggests a Normal ROT. The Kurtotic distribution leans towards an Asymmetric Laplace bandwidth while the Bimodal distribution (and the Asymmetric Bimodal distribution for $n \geq 200$) leans towards the Logistic bandwidth. The Separated Bimodal distribution suggests that the SJ bandwidth would be appropriate, while the Claw and Asymmetric Claw distributions suggest LSCV for smaller sample sizes and SJ for larger sample sizes. Similar results hold for AIC_c, GCV and RICE. The BIC criterion (Table B20) often suggests the GC bandwidth (the only table where that bandwidth shows prominence).

6.3. **What did we learn from these simulations.** When attempting to minimize AMISE, using the correct bandwidth selector for the underlying density generally appears to be acceptable. However, in practice, we do not know the true underlying density. When we move to model selection criteria, if the distribution is relatively well behaved, the Normal ROT appears to do well. However, with data that is less well behaved, which we often obtain in practice, different methods may work in different scenarios. It therefore seems prudent to use a model selection criteria in practice to help determine which bandwidth is best suited for a particular dataset.

7. Illustrations

In this section, we apply both existing and proposed bandwidth selectors to two separate datasets. Our first example is the well studied annual Buffalo snowfall data (Thaler 1974). This example is interesting as there is a debate in the litearature (Parzen 1979) on the number of modes in the density (1 vs 3). It is well known that the number of modes here is

tied to the bandwidth (Scott 1992). In our second example, we look at daily stock returns. This type of data is particularly relevant as it is well known that daily stock prices are often characterized by skewness (Mills 1995) and have heavy tails (Fama 1965).

We note here that it is arguable that the first dataset is i.i.d., whereas the second is clearly not. Our assumptions call for i.i.d. data, but we are curious to see how they perform with time dependency.

7.1. **Yearly snowfall.** The annual Buffalo snowfall data is well studied in density estimation (Parzen 1979). There is a debate in the literature as to whether the density is unimodal or trimodal and that conclusion is directly tied to the bandwidth. Here we propose to use existing and our alternative bandwidth selection criteria to estimate this density. We further use our model selection criteria to attempt reach a conclusion on the number of modes in the density.

Yearly snowfall values (to the nearest tenth of an inch) for Buffalo were recorded from 1910 to 1972. For this dataset, we have an IQR of 27.3750, a skewness of 0.0366 and an excess kurtosis of -0.5942 (i.e., the density is platykurtic). Figure 1 presents the kernel density estimate for our sample of 63 observations using the ubiquitous Normal ROT, along with each of our newly derived rules-of-thumb bandwidths as well as data-driven methods like least-squares cross-validation (LSCV), and the direct plug-in method (SJ).[16] A close look shows that the Normal and GC bandwidths produce a unimodal estimated density. The remaining bandwidths either produce a multimodal density or a density with shoulders. This figure shows why there is a debate in the literature.

Given that we do not have a prior on the number of modes, we turn to our model selection criteria. Table 8 gives the values for the model selection criteria for each bandwidth for each criterion. Regardless of the criterion, the GC bandwidth is selected and the normal rule-of-thumb bandwidth is a close second. As we mentioned above, each of these bandwidths led

[16]See Henderson & Parmeter (2015, Chapter 2).

FIGURE 1. Kernel density estimates for daily returns of Buffalo annual snowfall.

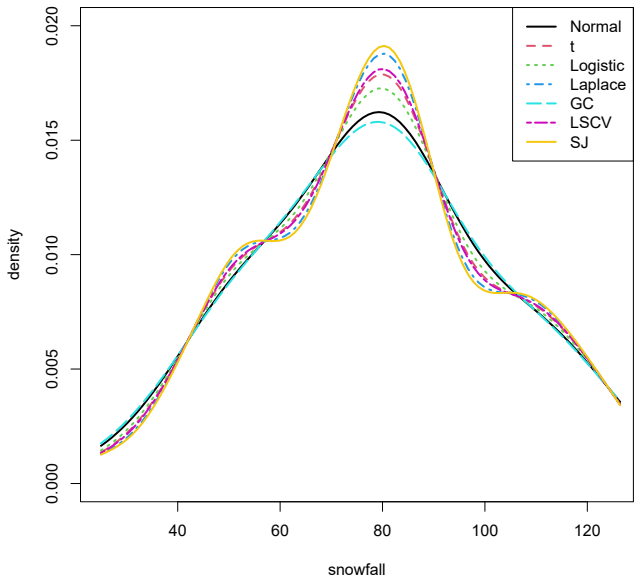


TABLE 8. Value of model selection criteria for each bandwidth (Buffalo snowfall and GameStop data).

	Normal	Logistic	Student's-t	Laplace	Asym Laplace	GC	lscv	sj	Overall winner	ROT winner
Buffalo snowfall										
AIC	4.681	4.689	4.695	4.708	6.564	4.678	4.698	4.713	GC	Normal
AIC_c	5.732	5.746	5.756	5.775	7.601	5.728	5.761	5.784	GC	Normal
GCV	4.685	4.695	4.703	4.718	6.565	4.682	4.707	4.724	GC	Normal
RICE	4.690	4.702	4.712	4.730	6.565	4.687	4.716	4.737	GC	Normal
BIC	4.819	4.852	4.874	4.912	6.599	4.807	4.883	4.927	GC	Normal
GameStop										
AIC	3.677	3.658	3.650	3.642	3.654	3.856	3.634	3.635	lscv	Laplace
AIC_c	4.690	4.671	4.664	4.657	4.699	6.322	4.656	4.657	lscv	Laplace
GCV	3.679	3.660	3.652	3.645	3.670	4.371	3.640	3.641	lscv	Laplace
RICE	3.680	3.662	3.655	3.648	3.689	NaN	3.646	3.647	lscv	Laplace
BIC	3.825	3.821	3.821	3.827	4.082	5.822	3.894	3.901	Log	Log

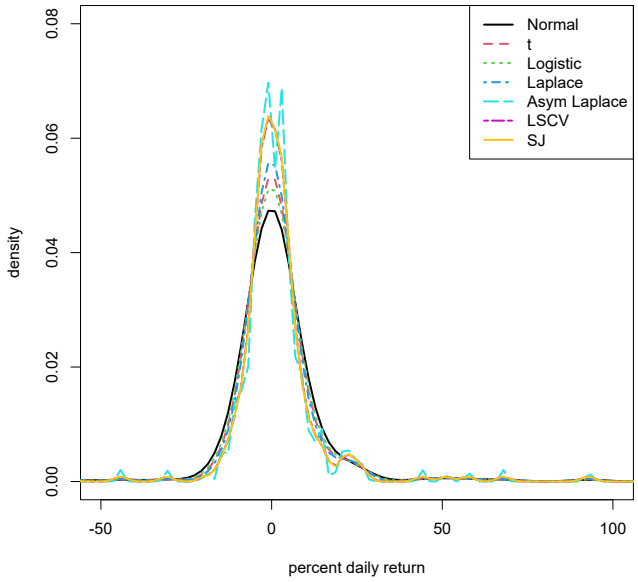
to a unimodal density. If we believe these results, we side with authors who believe that the density of annual snowfall data is unimodal (Parzen 1979).

7.2. **Daily stock returns.** Here we look at daily returns of GameStop stock over the period February 4, 2020 to February 2, 2021. This particular stock is both statistically (for the above aforementioned reasons) and economically interesting. In January, 2021, GameStop became headline news as a Reddit community (r/wallstreetbets) decided to work together in order (many via purchases on the Robinhood app) to increase the stock price of GameStop (GME). At the time, it was known that GameStop was one of the most shorted publicly traded firms. Short sellers borrow and sell a stock when its price is high, betting that it will fall. The activity of the Reddit community resulted in a temporary squeeze on these short sellers and some firms lost tens of billions of dollars in a very brief time period (e.g., S3 Partners). Over this period of time, the daily return hit a maximum change of 135% and a minimum daily return of roughly -60%. As expected, the bulk of returns of the distribution are near zero, but there is a lot of activity in the tails. It seems unlikely that a Normal approximation is appropriate for this particular dataset.

The data fit our setting, we have an IQR of 7.6598, a skewness of 3.8645 and an excess kurtosis of 27.4361. Figure 2 presents the kernel density estimate for our sample of 252 observations using most of the same bandwidths presented in Figure 1. We see several immediate features. First, Student-t, Asymmetric Laplace and SJ suggest a pronounced peakedness that is completely missing in the Normal ROT setup. Second, Logistic, Laplace and LSCV offer a compromise between the three aforementioned methods and the Normal ROT, picking up some peakedness, but not to the extent of Student-t/Asymmetric Laplace/SJ. Third, the three bandwidths that produced the highest peak seem to suggest the possibility of a second mode around 25.

Given that we do not know the underlying distribution, we switch our attention to the model selection criteria for bandwidth selection. Table 8 gives the values for the model

FIGURE 2. Kernel density estimates for daily returns of GameStop stock.



selection criteria for each bandwidth for each criterion. For AIC, AIC_c, GCV and RICE, the Laplace bandwidth produces the smallest value amongst ROT bandwidths. Amongst all bandwidths, LSCV produces the smallest value across these four selection criteria. As for BIC, the Logisitc bandwidth produces the smallest value amongst all bandwidths. We note that visually the Laplace, Logistic and LSCV bandwidths produce very similar features in Figure 2. Each of these methods appears appropriate here as the others tend to have spurious modes. The literature suggests that the density for a single asset should be unimodal (Schmitt & Westerhoff 2017).

8. CONCLUSION

In this study, we have enlarged the menu of available ROT bandwidths by considering alternative reference densities that reflect excess kurtosis and/or skewness. A technical

benefit of such ROTs is that they impart simple intuition on the behavior of the ROT bandwidth. When the skewness increases, the ROT bandwidth is monotonically decreasing in the skewness (for a given variance). This feature stems from the impact that the IQR has on the overall density. The IQR carries information about the slope of the density graph which translates to the magnitude of the 2^{nd} derivative of the density, the main feature of the density which determines the size of the AMISE optimaL bandwidth.

Given the series of reference bandwidths, we also proposed a novel density (model) selection approach based on common model selection criterion. This model selection approach offers a new way to think about bandwidth selection amonst a set of reference densities, not unsimilar to the work of Hansen (2005) for kernel order selection.

Simulations indicated that they outperform the Normal ROT that is in widespread use, especially with data whose distribution has a slim body or exhibits skewness. They also often outperform data-driven and plug-in methods that appear to continue to be plagued by unfulfilled asymptotic promises when they face finite data samples. We also demonstrated that once skewness or excess kurtosis becomes prevalent in the data, that the common Normal ROT is less likely to be deemed the winner for model selection. That honor fell to the Asymmetric Laplace ROT.

We applied our bandwidth procedures to two separate empirical datasets. For each dataset, different bandwidths told a different story. Our model selection criteria picked bandwidths which aligned with results in the literature.

References

Abadir, K. M. & Lawford, S. (2004), 'Optimal asymmetric kernels', *Economics Letters* **83**, 61–68.

Birnbaum, Z. W. (1948), 'On Random Variables with Comparable Peakedness', *The Annals of Mathematical Statistics* **19**(1), 76–81.

Boguth, O., Duchin, R. & Simutin, M. (2021), 'Dissecting conglomerate valuations', *Journal of Finance* . Forthcoming, http://dx.doi.org/10.2139/ssrn.2693847.

Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y. & Kulveit, J. (2021), 'Inferring the effectiveness of government interventions against COVID-19', *Science* **371**(6531).

Dharmani, B. (2015), 'The Gram-Charlier A series based extended rule-of-thumb for bandwidth selection in univariate and multivariate kernel density estimations', *arXiv e-prints* .
URL: *https://arxiv.org/abs/1504.00781*

Epanechnikov, V. A. (1969), 'Nonparametric estimation of a multidimensional probability density', *Theory of Probability and its Application* **14**, 153–158.

Fama, E. F. (1965), 'The behavior of stock-market prices', *Journal of Business* **38**(1), 34–105.

Guerre, E., Perrigne, I. & Vuong, Q. (2000), 'Optimal nonparametric estimation of first-price auctions', *Econometrica* **68**, 525–574.

Hall, P., Minnotte, M. C. & Zhang, C. (2004), 'Bump hunting with non-Gaussian kernels', *The Annals of Statistics* **32**(5), 2124 – 2141.

Hansen, B. E. (2005), 'Exact mean integrated squared error of higher order kernel estimators', *Econometric Theory* **21**, 1031–1057.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
URL: *http://www.jstatsoft.org/v27/i05/*

Henderson, D. J. & Parmeter, C. F. (2012), 'Normal reference bandwidths for the general order, multivariate kernel density derivative estimator', *Statistics and Probability Letters* **82**(8), 2198–2205.

Henderson, D. J. & Parmeter, C. F. (2015), *Applied Nonparametric Econometrics*, Cambridge University Press.

Henderson, D. J., Parmeter, C. F. & Russell, R. R. (2008), 'Modes, Weighted Modes, and Calibrated Modes: Evidence of Clustering Using Modality Tests', *Journal of Applied Econometrics* **23**(5), 607–638.

Jeong, D., Im, J. & Kim, Y. M. (2021), 'Cosine-based variable bandwidth selection for nonparametric spectral density estimation under long-range dependence', *Journal of Statistical Computation and Simulation* .

Känzig, D. R. (2021), 'The macroeconomic effects of oil supply news: Evidence from OPEC announcements', *American Economic Review* **111**(4), 1092–1125.

Katsiampa, P. (2019), 'Volatility co-movement between Bitcoin and Ether', *Finance Research Letters* **30**, 221–227.

Le Quéré, C., Peters, G. P., Friedlingstein, P., Andrew, R. M., Canadell, J. G., Davis, S. J., Jackson, R. B. & Jones, M. W. (2021), 'Fossil CO2 emissions in the post-COVID-19 era', *Nature Climate Change* **11**, 197–199.

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Loader, C. R. (1999), 'Bandwidth selection: Classical or plug-in?', *Annals of Statistics* **27**(2), 415–438.

Marron, J. S. & Wand, M. P. (1992), 'Exact mean integrated squared error', *Annals of Statistics* **20**(2), 712–736.

McCloud, N. & Parmeter, C. F. (2020), 'Determining the number of effective parameters in kernel density estimation', *Computational Statistics & Data Analysis* **143**, 106843.

McCloud, N. & Parmeter, C. F. (2021), 'Calculating degrees of freedom in multivariate local polynomial regression', *Journal of Statistical Planning and Inference* **210**, 141–160.

Mills, T. C. (1995), 'Modelling skewness and kurtosis in the london stock exchange ft-se index return distributions', *The Statistician* **44**(3), 323–332.

Minnotte, M. C. (2010), 'Mode Testing via Higher-Order Density Estimation', *Computational Statistics* **25**, 391–407.

Muller, H.-G. (1984), 'Smooth optimum kernel estimators of densities, regression curves and modes', *Annals of Statistics* **12**(2), 766–774.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *Annals of Mathematical Statistics* **33**, 1065–1076.

Parzen, E. (1979), 'Nonparametric statistical data modeling', *Journal of the American Statistical Association* **74**, 105–121.

Poiraud-Casanova, S. & Thomas-Agnan, C. (2000), 'About monotone regression quantiles', *Statistics & probability letters* **48**(1), 101–104.

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

Rudy, S. H., Kutz, J. N. & Brunton, S. L. (2019), 'Deep learning of dynamics and signal-noise decomposition with time-stepping constraints', *Journal of Computational Physics* **396**, 483–506.

Schmitt, N. & Westerhoff, F. (2017), 'On the bimodality of the distribution of the S& P 500's distortion: Empirical evidence and theoretical explanations', *Journal of Economic Dynamics & Control* **80**, 34–53.

Scott, D. W. (1992), *Multivariate density estimation: Theory, practice, and visualization*, John Wiley and Sons, New York.

Sheather, S. J. & Jones, M. C. (1991), 'A reliable data-based bandwidth selection method for kernel density estimation', *Journal of the Royal Statistical Society, Series B* **53**, 683–690.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall, New York.

Sun, H., Yang, X. & Gao, H. (2019), 'A spatially constrained shifted asymmetric Laplace mixture model for the grayscale image segmentation', *Neurocomputing* **331**, 50–57.

Terrell, G. R. (1990), 'The maximal smoothing principle in density estimation', *Journal of the American Statistical Association* **85**(410), 470–477.

Thaler, H. (1974), Nonparametric probability density estimation and the empirical characteristic function. Ph.D. thesis, State University of New York at Buffalo.

Tiwari, A. K., Raheem, I. D. & Kang, S. H. (2019), 'Time-varying dynamic conditional correlation between stock and cryptocurrency markets using the copula-ADCC-EGARCH model', *Physica A: Statistical Mechanics and its Applications* **535**, 122295.

Wand, M. P. & Jones, M. C. (1994), *Kernel Smoothing*, Chapman & Hall/CRC, Boca Raton.

Westfall, P. H. (2014), 'Kurtosis as peakedness, 1905-2014. R.I.P.', *American Statistician* **68**, 191–195.