Submitted to Statistical Science

Studentization vs. variance stabilization: a simple way out of an old dilemma

Dimitris N. Politis

Abstract. Assume $\hat{\theta}_n$ is a statistic used to estimate a parameter θ on the basis of data $X_1, \ldots, X_n$. Further assume that $\hat{\theta}_n$ is consistent and asymptotically normal, with asymptotic variance given by $\sigma^2(\theta)$. Even if the functional form of $\sigma^2(\cdot)$ is known, its dependence on the unknown parameter θ creates a dilemma as regards the construction of a confidence interval for θ. Should the interval be based on the normal quantiles with estimated variance, i.e., studentization, or shall we transform the statistic $\hat{\theta}_n$ to $Y_n = g(\hat{\theta}_n)$ such that the asymptotic variance of Y_n does not depend on θ, i.e., variance stabilization? We show how this dilemma can be bypassed by a straightforward construction that applies rather generally, and just hinges on solving simple algebraic equations. We illustrate the new approach on a host of numerical examples, including two examples in nonparametric function estimation. In the latter, a different sort of dilemma arises: employing under-smoothing vs. an explicit bias correction. This paper is dedicated to the memory of Dr. Dimitrios Gatzouras (1962-2020).

Key words and phrases: Bias correction, confidence intervals, Edgeworth expansion, finite-sample coverage, probability density estimation, under-smoothing.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be a set of observed data governed by a probability law P. Suppose θ is a parameter of interest, i.e., a feature of P, and $\hat{\theta}_n$ is a statistic used to estimate θ on the basis of $X_1, \ldots, X_n$. Further assume that $\hat{\theta}_n$ is consistent and asymptotically normal at rate τ_n; here, τ_n is some sequence diverging to ∞ as $n \to \infty$.

The prototypical example is when $X_1, \ldots, X_n$ are independent, identically distributed (i.i.d.) with mean θ and finite variance σ^2. Let $\hat{\theta}_n = \bar{X}$ where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ is the sample mean; then, the Central Limit Theorem for i.i.d. random variables (r.v.) implies

$$(1) \qquad \sqrt{n}(\hat{\theta}_n - \theta) \stackrel{\mathcal{L}}{\Longrightarrow} N\left(0, \sigma^2\right) \quad \text{as} \quad n \to \infty$$

in which case the rate τ_n is tantamount to $\sqrt{n}$.

Nevertheless, the variance of the limiting normal distribution will sometimes depend on the unknown parameter θ; this is particularly common if the support of the data is bounded below and/or above. In that case, eq. (1) has to be modified to:

$$(2) \qquad \tau_n(\hat{\theta}_n - \theta) \stackrel{\mathcal{L}}{\Longrightarrow} N\left(0, \sigma^2(\theta)\right) \quad \text{as} \quad n \to \infty$$

where $\sigma(\cdot)$ is a continuous function taking only positive values.

EXAMPLE 1.1. **[Poisson]** Suppose $X_1, \ldots, X_n$ are i.i.d. with a Poisson distribution of mean θ. Since the variance of a Poisson r.v. equals its mean, it follows that $\sigma^2(\theta) = \theta$. Hence, eq. (2) holds true with $\hat{\theta}_n = \bar{X}$ and $\tau_n = \sqrt{n}$.

Coming back to eq. (2), note that its Left-Hand-Side (LHS) is not a *pivot*, since its large-sample distribution is not free of parameters. To practically use eq. (2) in order to construct a large-sample $(1-\alpha)$ 100% confidence interval for θ — without resort to *bootstrap* — one of two ways has been typically adopted:

- **Studentization (ST):** Since $\hat{\theta}_n$ is consistent for θ and $\sigma(\cdot)$ is continuous, eq. (2) implies

$$(3) \qquad \tau_n \frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)} \stackrel{\mathcal{L}}{\Longrightarrow} N(0,1) \quad \text{as} \quad n \to \infty$$

leading to the (asymptotic) $(1-\alpha)$ 100% confidence statement:[1]

Dimitris N. Politis is Distinguished Professor at the Department of Mathematics and the Halicioglu Data Science Institute , University of California, San Diego, La Jolla, CA 92093-0112, USA (e-mail: dpolitis@ucsd.edu).

[1]The treatment in this paper applies equally to one-sided confidence bounds. In what follows, we focus on (symmetric) confidence

(4)
$$\left|\tau_n \frac{\hat\theta_n - \theta}{\sigma(\hat\theta_n)}\right| \le z_{1-\frac{\alpha}{2}},$$
$$\text{i.e., } \theta \in [\hat\theta_n - z_{1-\frac{\alpha}{2}} \frac{\sigma(\hat\theta_n)}{\tau_n},\ \hat\theta_n + z_{1-\frac{\alpha}{2}} \frac{\sigma(\hat\theta_n)}{\tau_n}].$$

As usual, $z_\beta = \Phi^{-1}(\beta)$ where $\Phi(\cdot)$ denotes the standard normal distribution.

REMARK 1.1. **On 'studentization'.** Note that we refer to the above construction as 'studentization' because the division by an estimated variance makes the LHS of (3) be a 'studentized' — as opposed to standardized — quantity.

- **Variance Stabilization (VS):** Let $Y_n = g(\hat\theta_n)$ where $g(\cdot)$ is a smooth (at least continuously differentiable) monotone function; denote also $Y = g(\theta)$. Assuming $g'(\theta) \neq 0$, eq. (2) together with the delta-method imply:

$$\tau_n(Y_n - Y) \overset{\mathcal{L}}{\Longrightarrow} N\left(0, [g'(\theta)]^2\sigma^2(\theta)\right) \quad \text{as } n \to \infty. \tag{5}$$

If we can choose $g(\cdot)$ such that $g'(x) = c/\sigma(x)$ for some constant c, then $g(\cdot)$ is called a variance stabilizing transformation, as it implies

$$\tau_n(Y_n - Y) \overset{\mathcal{L}}{\Longrightarrow} N\left(0, c^2\right) \quad \text{as } n \to \infty \tag{6}$$

which, in turn, yields the (asymptotic) $(1-\alpha)$ 100% confidence statement:

$$\left|\tau_n \frac{g(\hat\theta_n) - g(\theta)}{|c|}\right| \le z_{1-\frac{\alpha}{2}}$$

i.e.,

$$g(\hat\theta_n) - z_{1-\frac{\alpha}{2}} \frac{|c|}{\tau_n} \le g(\theta) \le g(\hat\theta_n) + z_{1-\frac{\alpha}{2}} \frac{|c|}{\tau_n}$$

leading to the (asymptotic) $(1-\alpha)$ 100% confidence interval

$$\theta \in [g^{-1}\left(g(\hat\theta_n) - z_{1-\frac{\alpha}{2}} \frac{|c|}{\tau_n}\right), \quad g^{-1}\left(g(\hat\theta_n) + z_{1-\frac{\alpha}{2}} \frac{|c|}{\tau_n}\right)] \tag{7}$$

To simplify notation, here and **throughout the paper**, we will denote $c_\alpha = z_{1-\frac{\alpha}{2}}/\tau_n$.

intervals in order to fix ideas, and also because they are the most popular in practice.

Example 1.1 [Poisson, continued] *In the Poisson mean case, the studentized confidence interval (4) reads*

$$\theta \in \left[\hat\theta_n - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat\theta_n}{n}},\ \hat\theta_n + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat\theta_n}{n}}\right].$$

where $\hat\theta_n = \bar X$. Since $\tau_n = \sqrt{n}$ here, the above can be written more compactly as:

$$\theta = \hat\theta_n \pm c_\alpha\sqrt{\hat\theta_n}.$$

Furthermore, we can achieve variance stabilization (with $c = 1/2$) by using the function $g(x) = \sqrt{x}$. Consequently, the variance stabilized confidence interval (7) reads

$$\theta \in \left[\left(\sqrt{\hat\theta_n} - z_{1-\frac{\alpha}{2}}\frac{0.5}{\sqrt{n}}\right)^2,\ \left(\sqrt{\hat\theta_n} + z_{1-\frac{\alpha}{2}}\frac{0.5}{\sqrt{n}}\right)^2\right]$$

that can be compactly written as:

$$\theta = \hat\theta_n + \frac{c_\alpha^2}{4} \pm c_\alpha\sqrt{\hat\theta_n}.$$

Although both confidence intervals (4) and (7) have asymptotic coverage probability $1-\alpha$, they may suffer from finite-sample inaccuracies. For example, the studentized statistic at the LHS of (3) may be quite non-normal for small samples; recall the special case of Student's t distribution that obtains when $\hat\theta_n = \bar X$, and the X_i are exactly Normal. Regarding variance stabilization, the main issue is bias. To elaborate, in the case of the sample mean $\hat\theta_n = \bar X$, it follows that $E\hat\theta_n = \theta$, but $Eg(\hat\theta_n) \neq Eg(\theta)$. Even though $Eg(\hat\theta_n) - Eg(\theta)$ is typically of order $o(1/\sqrt{n})$, this bias can still be problematic in moderate samples; see Ch. 4 of DasGupta (2008).

In the next section, we present a simple approach, termed the Confidence Region (CR) method, that yields confidence intervals devoid from the abovementioned deficiencies; in particular, no transformation or studentization is needed. Section 3 discusses the possible preliminary use of a normalizing transformation, while Section 4 compares the proposed methods via a numerical simulation. Applications to two problems in nonparametric function estimation are discussed in Sections 5 and 6; here, the question on whether to employ studentization or variance stabilization is overshadowed by a new dilemma: whether to employ undersmoothing or an explicit bias-correction in the construction of confidence intervals. Appendix A elaborates on nonparametric estimation of higher-order derivatives via flat-top kernels that are useful in order to estimate the aforementioned bias.

We conclude the present section by giving two more important examples. Recall the general definition $c_\alpha = z_{1-\frac{\alpha}{2}}/\tau_n$.

EXAMPLE 1.2. **[Gamma]** Suppose $X_1, \ldots, X_n$ are i.i.d. with a Gamma density $\Gamma(\beta)^{-1}\,\theta^{-\beta} x^{\beta-1}\exp(-x/\theta)$ with $\beta > 0$ assumed known for simplicity. E.g., if $\beta = 1$, then X_i has an Exponential density with mean θ. In the general Gamma case, $EX_i = \beta\theta$, so letting $\hat{\theta}_n = \beta^{-1}\bar{X}$ we can verify that eq. (2) holds true with $\sigma^2(\theta) = \theta^2$, and $\tau_n = \sqrt{n\beta}$.

Consequently, a studentized $(1-\alpha)$ 100% confidence interval for θ is:

$$\theta = \hat{\theta}_n \pm c_\alpha \hat{\theta}_n.$$

In addition, it is easy to see that the natural logarithm $g(x) = \log x$ achieves variance stabilization, leading to eq. (6) with $c^2 = \beta^{-1}$. The $(1-\alpha)$ 100% VS confidence interval (7) reads

$$\theta \in \left[\exp\left(\log\hat{\theta}_n - c_\alpha\right),\ \exp\left(\log\hat{\theta}_n + c_\alpha\right)\right].$$

REMARK 1.2. **On τ_n.** When the asymptotic variance of $\hat{\theta}_n$ involves a multiplicative constant, as in the above Gamma example, we will absorb it in τ_n; this is in order for $\sigma^2(\theta)$ to have a simple form to work with, and compare with other similar examples.

EXAMPLE 1.3. **[Binomial]** Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli (θ), so that $\hat{\theta}_n = \bar{X}$; recall, that in this case $n\hat{\theta}_n$ has a Binomial (n, θ) distribution. As well known, eq. (2) holds true with $\sigma^2(\theta) = \theta(1-\theta)$, and $\tau_n = \sqrt{n}$ and the studentized $(1-\alpha)$ 100% confidence interval reads

$$\theta = \hat{\theta}_n \pm c_\alpha\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}. \tag{8}$$

In the binomial case, the function $g(x) = \arcsin\sqrt{x}$ achieves variance stabilization, leading to eq. (6) with $c^2 = 1/4$. The $(1-\alpha)$ 100% VS confidence interval (7) reads

$$\theta \in \left[\sin\left(\arcsin\hat{\theta}_n - c_\alpha/2\right),\ \sin\left(\arcsin\hat{\theta}_n + c_\alpha/2\right)\right].$$

2. A SIMPLE WAY OUT OF THE DILEMMA

There is abundant literature on the dichotomy between variance stabilization and studentization. For example, dividing by $\sigma(\hat{\theta}_n)$ may significantly alter the quantiles of the distribution of $\hat{\theta}_n$; hence, it can be said that the ST interval (4) amounts to "looking up the wrong tables"; see e.g. Hall (1988).

By contrast, variance stabilization may introduce bias that also influences the confidence intervals. To see why, consider the case where $\hat{\theta}_n$ is the sample mean $\bar{X}$ of i.i.d. data. Here $E\bar{X} = \theta$ but it is apparent that $Eg(\bar{X}) \neq g(\theta)$ in general. Although the delta method shows that $\sqrt{n}\ E\left(g(\bar{X}) - g(\theta)\right) = O(1/\sqrt{n})$, this $O(1/\sqrt{n})$ term is of the same order of magnitude as the statistical error in approximating the true $1-\alpha/2$ quantile of $\sqrt{n}\,E\left(g(\bar{X}) - g(\theta)\right)$ by the normal $z_{1-\alpha/2}$.

It is well known that variance stabilization is preferable over studentization in the Poisson and Binomial examples; see Ch. 4 of DasGupta (2008) and the references therein. In the Gamma example the situation is not so clear; our Section 4 attempts to shed some light. Nevertheless, it is a false premise that a practitioner must choose one of these two options; there is a third option that is more straightforward.

To elaborate, note that in the context of eq. (2) we can simply write

$$\tau_n \frac{\hat{\theta}_n - \theta}{\sigma(\theta)} \overset{\mathcal{L}}{\Longrightarrow} N(0,1) \quad \text{as} \quad n \to \infty \tag{9}$$

leading to the (asymptotic) $(1-\alpha)$ 100% *confidence region*

$$\{\text{all } \theta : \left|\tau_n \frac{\hat{\theta}_n - \theta}{\sigma(\theta)}\right| \leq z_{1-\frac{\alpha}{2}}\}. \tag{10}$$

The key observation is that when $\sigma(x)$ is of simple enough functional form, e.g. when $\sigma^2(x)$ is a second order polynomial in x, the confidence region (10) may be turned into a confidence *interval* using simple algebraic manipulations.

We work out some important examples below; in what follows, recall that $c_\alpha = \tau_n^{-1} z_{1-\frac{\alpha}{2}}$, and assume n is large enough so that $c_\alpha < 1$.

1. **Case $\sigma(x) = \sqrt{x}$.** Squaring both sides of the inequality in (10) and solving for θ leads to the following (asymptotic) $(1-\alpha)$ 100% confidence interval

$$\theta = \hat{\theta}_n + \frac{c_\alpha^2}{2} \pm c_\alpha\sqrt{\hat{\theta}_n + c_\alpha^2/4}\,. \tag{11}$$

   The above is applicable to the Poisson Example 1.1 using $\tau_n = \sqrt{n}$; it can be compared to the intervals obtained via studentization and variance stabilization.

2. **Case $\sigma(x) = x$.** In this case, (10) is equivalent to the (asymptotic) $(1-\alpha)$ 100% confidence interval

$$\theta \in \left[\frac{\hat{\theta}_n}{1+c_\alpha}, \frac{\hat{\theta}_n}{1-c_\alpha}\right]. \tag{12}$$

   The above is applicable to the Gamma Example 1.2 using $\tau_n = \sqrt{n\beta}$; it can be compared to the intervals obtained via studentization and variance stabilization.

3. **Case $\sigma(x) = \sqrt{x(1-x)}$.** Here (10) is equivalent to the (asymptotic) $(1-\alpha)$ 100% confidence interval

$$\theta = \frac{2\hat{\theta}_n + c_\alpha^2 \pm c_\alpha\sqrt{4\hat{\theta}_n(1-\hat{\theta}_n) + c_\alpha^2}}{2(1+c_\alpha^2)} \tag{13}$$

which is applicable to the Binomial Example 1.3 using $\tau_n = \sqrt{n}$; it can be compared to the intervals obtained via studentization and variance stabilization.

We will call the above method of constructing confidence intervals, the **Confidence Region (CR)** method, to distinguish it from the intervals obtained via either variance stabilization or studentization. Note that if the functional form of $\sigma(x)$ is more complicated, it may still be possible to reduce the confidence region (10) to a confidence interval (or a union of intervals) by solving an equation such as $c_\alpha \sigma(\theta) + \theta = \hat{\theta}_n$ *numerically* for θ, and then costructing the relevant inequalities.

REMARK 2.1. **On the Poisson.** When applied to the Poisson Example 1.1, the confidence interval (11) was first proposed by Crow and Gardner (1959), and subsequently refined by Casella and Robert (1989). The construction of Crow and Gardner (1959) gives a further motivation for the the general confidence region (10): it is the confidence region obtained by inverting an α–level two-sided test of a point null hypothesis on θ, where the test is based on the asymptotic distribution (9).

REMARK 2.2. **On the Gamma.** A special case of the Gamma Example 1.2 is the Exponential distribution. E.g., Section 6 describes an application to spectral density estimation that falls under this setup. The confidence interval (12) was employed in the spectral estimation application by Politis et al. (1992) who also considered a relevant resampling scheme to improve upon it; see Remark 3.2 for more details on the bootstrap.

REMARK 2.3. **On the Binomial.** When applied to the Binomial Example 1.3, the confidence interval (13) was first proposed by Wilson (1927); it is one of the preferred intervals for a binomial proportion as discussed in the comprehensive review of Brown et al. (2001) who warn against using the 'Wald' interval, i.e., the studentized interval (8). Furthermore, it is well known that the binomial CLT —and its associated confidence intervals— can be aided by a continuity correction; alternatively, the *split-sample* method of Decrouez and Hall (2014) could be used. To describe it, let the i.i.d. sample $X_1, \dots, X_n$ be split into two subsamples, say $X_1, \dots, X_m$ and $X_{m+1}, \dots, X_n$ with (sub)sample means $\bar{X}_1$ and $\bar{X}_2$ respectively. The split-sample estimator is $\tilde{\theta}_n = (\bar{X}_1 + \bar{X}_2)/2$. If $m \neq n-m$ but with $m/(n-m) \to 1$, then $\tilde{\theta}_n$ has the same asymptotic normal distribution as $\hat{\theta}_n = \bar{X}$ but devoid of the need for continuity correction. Nevertheless, since $\tilde{\theta}_n$ and $\hat{\theta}_n$ have the same asymptotic normal distribution, the interval (13) applies *verbatim* with $\tilde{\theta}_n$ instead $\hat{\theta}_n$; see Thulin (2014). Notably, the split-sample method can be applied to other lattice r.v.'s. Focusing on the Poisson Example 1.1, we could construct a split-sample estimator is $\tilde{\theta}_n$ as above; then, the CR interval (11) would apply *verbatim* with $\tilde{\theta}_n$ instead $\hat{\theta}_n$, and may yield improved accuracy.

REMARK 2.4. **On Bias-Correction.** In anticipation of the nonparametric examples in Sections 5 and 6, we now discuss the construction of *bias-corrected* confidence intervals. Suppose that instead of eq. (2) we have

$$\tau_n(\hat{\theta}_n - E\hat{\theta}_n) \overset{\mathcal{L}}{\Longrightarrow} N\left(0, \sigma^2(\theta)\right) \quad \text{as} \quad n \to \infty \tag{14}$$

with

$$\tau_n(E\hat{\theta}_n - \theta) = b(\delta) + o(1) \quad \text{as} \quad n \to \infty \tag{15}$$

for some continuous function $b(\cdot)$ capturing the asymptotic bias; here, δ is an unknown parameter that can be related to θ but it does not have to be. If $b(\delta) = 0$, then eq. (2) follows; but if $b(\delta) \neq 0$, then the following procedure can be used:

(a) Use eq. (14) with any of the abovementioned methods (ST, VS or CR) to construct an (asymptotic) $(1-\alpha)$ 100% confidence interval for $E\hat{\theta}_n$; denote this interval by $[\underline{C}, \overline{C}]$.

(b) Note that $\underline{C} \leq E\hat{\theta}_n \leq \overline{C}$ is equivalent to

$$\underline{C} - \frac{b(\delta)}{\tau_n} \leq E\hat{\theta}_n - \frac{b(\delta)}{\tau_n} \leq \overline{C} - \frac{b(\theta)}{\tau_n}.$$

Hence, an (asymptotic) $(1-\alpha)$ 100% confidence interval for θ is

$$\left[\underline{C} - \frac{b(\delta)}{\tau_n}, \ \overline{C} - \frac{b(\delta)}{\tau_n}\right]. \tag{16}$$

Since δ is unknown, the above can be thought to be an *oracle* statement.

(c) If $\hat{\delta}_n$ is a consistent estimator of δ, then a practically useful *Bias-Corrected* (BC) (asymptotic) $(1-\alpha)$ 100% confidence interval for θ is

$$\left[\underline{C} - \frac{b(\hat{\delta}_n)}{\tau_n}, \ \overline{C} - \frac{b(\hat{\delta}_n)}{\tau_n}\right]. \tag{17}$$

The BC interval (17) has asymptotically correct coverage level but it will not yield an improvement over the original interval $[\underline{C}, \overline{C}]$ unless $\hat{\delta}_n$ is accurate enough. For example, if $\hat{\delta}_n$ has a slower rate of convergence as compared to $\hat{\theta}_n$, then the uncorrected interval $[\underline{C}, \overline{C}]$ may be preferable. If $\hat{\delta}_n$ and $\hat{\theta}_n$ have the same rate of convergence, then the situation is not clear, and has to be examined given the particulars of the problem at hand. Nevertheless, if $\hat{\delta}_n$ has a *faster* rate of convergence than $\hat{\theta}_n$, then the BC interval (17) will be asymptotically equivalent to the *oracle* interval (16), and therefore preferable; this is the case in nonparametric function estimation when $\hat{\theta}_n$ is obtained via a 2nd order kernel, while $\hat{\delta}_n$ is obtained via a flat-top kernel—see Sections 5 and 6.

3. ONE STEP FURTHER: NORMALIZING TRANSFORMATIONS

The CR method for confidence intervals outlined in Section 2 is just based on the limiting distribution (2). If the Right-Hand-Side (RHS) of (2) is a good approximation to the distribution of the LHS for the problem at hand, then the CR confidence intervals would be quite accurate. For example, if the normality in (2) is exactly achieved, e.g. $\hat{\theta}_n$ is the sample mean of Normal r.v.'s, then the coverage of the CR confidence intervals would be exact; the coverage of the ST or VS confidence intervals will not be exact in such a case, since they both entail an adulteration of the distribution of the statistic in question.

On the other hand, if the RHS of (2) is *not* a good approximation to the distribution of the LHS for the problem (and sample size) at hand, then the coverage of the CR confidence intervals may suffer. This motivates the (potential) need for a *normalizing* (instead of variance stabilizing transformation). Nevertheless, the caveat still applies in that any transformation may introduce bias that — as skewness — is not captured in the Gaussian limit of (2).

By the Berry-Esseen theorem, the speed of convergence in (2) is often dictated by the skewness of $\hat{\theta}_n$ in the sample mean and related cases. Hence, a normalizing transformation may be constructed with the purpose of reducing skewness which is defined as $skew(\hat{\theta}_n) = \frac{E(\hat{\theta}_n-\theta)^3}{Var(\hat{\theta}_n)^{3/2}}$.

The last section discussed the dichotomy between the CR method and VS, i.e., using a variance stabilizing transform. However, there is the additional dilemma on whether to employ a variance stabilizing or a normalizing transformation—see the seminal paper of Box and Cox (1964). In view of the fact that Section 2 put forth the CR method as an alternative to VS, we may now propose a general procedure: first apply a normalizing transformation and afterwards employ the CR method whenever, of course, the latter is feasible.

Let CRaN denote the above general procedure, i.e., applying the CR method after Normalization. The CRaN proposal is then elaborated upon as follows:

1. Find a smooth, one-to-one function $h(\cdot)$, such that $skew(\hat{\zeta}_n) = o(skew(\hat{\theta}_n))$ where $\hat{\zeta}_n = h(\hat{\theta}_n)$, i.e., reduce the skewness by an order of magnitude.
2. Provided the bias of $\hat{\zeta}_n$ is negligible, apply the CR method of Section 2 to $\hat{\zeta}_n$, and construct a $(1-\alpha)$ 100% confidence interval for $\zeta = h(\theta)$, say $\zeta \in [\underline{C}, \overline{C}]$.
3. Finally, invert the function h to construct a $(1-\alpha)$ 100% confidence interval for θ. For example, if h is monotone increasing, then the confidence interval is $\theta \in [h^{-1}(\underline{C}), h^{-1}(\overline{C})]$; if h is monotone decreasing, then the confidence interval is $\theta \in [h^{-1}(\overline{C}), h^{-1}(\underline{C})]$.

To give flesh to the above ideas, let us focus momentarily on the case where $\hat{\theta}_n = \bar{X}$, with $X_1, \ldots, X_n$ i.i.d. from a distribution with mean θ and central moments $\mu_k = E(X_1-\theta)^k$. If the function $h(\cdot)$ is sufficiently smooth, i.e., admitting a Taylor expansion such as

$$h(x) = h(\theta) + h'(\theta)(x-\theta) + \frac{1}{2}h''(\theta)(x-\theta)^2 + \cdots$$

to some appropriate order, then eq. (2) implies

$$\sqrt{n}(\hat{\zeta}_n - \zeta) \overset{\mathcal{L}}{\Longrightarrow} N(0, \mu_2[h'(\theta)]^2) \text{ as } n \to \infty; \tag{18}$$

however, for the above to be useful, the asymptotic variance must be expressed as a function of ζ. Furthermore,

$$\begin{aligned} E\hat{\zeta}_n &= h(\theta) + \frac{\mu_2[h''(\theta)]}{2n} + O(\frac{1}{n^2}) \\ \text{and } Var(\hat{\zeta}_n) &= \frac{\mu_2[h'(\theta)]^2}{n} + O(\frac{1}{n^2}) \end{aligned} \tag{19}$$

implying that the bias of $\hat{\zeta}_n = h(\bar{X})$ is of small enough order (since we are comparing $Var(\hat{\zeta}_n)$ with $(E\hat{\zeta}_n - h(\theta))^2$); this underscores the importance of applying the transformation on a statistic that is exactly unbiased, whenever possible.

In addition, Example 6.1 in Ch. 14 of Shorack (2000) yields[2]

$$\begin{aligned} & E\left(\hat{\zeta}_n - E\hat{\zeta}_n\right)^3 \\ &= \frac{\mu_3[h'(\theta)]^3 + 3\mu_2^2[h'(\theta)]^2 h''(\theta)}{n^2} + o(\frac{1}{n^2}). \end{aligned} \tag{20}$$

Hence, the defining property of a normalizing transformation is

$$\mu_3 h'(\theta) + 3\mu_2^2 h''(\theta) = 0$$

which is a first order ordinary differential equation for $h'(x)$.

It is easy to check that for our three main Examples 1.1 (Poisson), 1.2 (Gamma), and 1.3 (Binomial), the normalizing transformations are $h(x) = x^{2/3}$, $h(x) = x^{1/3}$, and $h(x) = \int_0^x [s(1-s)]^{-1/3} ds$ respectively; these are all different from their respective variance stabilizing transformations.

Notably, we recognize $\theta' = EX_1 = \beta\theta$ as the underlying parameter for the Gamma distribution (instead of θ). Therefore, as demonstrated in example 1.2, we need to divide by β after calculating the quantiles. In the special case of the Exponential distribution, however, $\beta = 1$ and the two parameters θ', θ coincide.

[2]Note, however, a typo in eq. (6) on p. 396 of Shorack (2000); the correct expression for the third moment appears in his Example 6.3 on p. 397.

Example 1.1 (Poisson, continued) *The normalizing transformation is $h(x) = x^{2/3}$. Then, eq. (18) reads $\sqrt{n}(\widehat{\zeta}_n - \zeta) \overset{\mathcal{L}}{\Longrightarrow} N(0, \frac{4}{9}\theta^{1/3})$. Recall that $\zeta = h(\theta)$, i.e., $\zeta = \theta^{2/3}$; hence, $\theta^{1/3} = \zeta^{1/2}$, implying*

$$\sqrt{n}(\hat{\zeta}_n - \zeta) \overset{\mathcal{L}}{\Longrightarrow} N(0, \frac{4}{9}\,\zeta^{1/2}) \ \text{ as } \ n \to \infty. \tag{21}$$

To apply the CR method, we need to solve the relation $\left|\frac{\hat{\zeta}_n - \zeta}{\zeta^{1/4}}\right| \le \frac{2}{3}c_\alpha$ for ζ. Let $C = [\frac{2}{3}c_\alpha]^4$, $a = \hat{\zeta}_n$, and $x = \zeta$. Then, the CR relation is equivalent to

$$[x-a]^4 \le Cx \ \text{ i.e., } \ [x-a]^4 - Cx \le 0 \tag{22}$$

that can be solved numerically for $x > 0$ to yield the desired CRaN interval with approximate 95% confidence level.

Example 1.2 (Gamma, continued) *The normalizing transformation is $h(x) = x^{1/3}$. Then, eq. (18) reads $\sqrt{n}(\hat{\zeta}_n - \zeta) \overset{\mathcal{L}}{\Longrightarrow} N(0, \frac{1}{9}\beta^{-1/3}\ \theta^{2/3})$. Recall that $\zeta = (\beta\theta)^{1/3}$, i.e., $\theta = \frac{\zeta^3}{\beta}$; hence,*

$$\sqrt{n}(\hat{\zeta}_n - \zeta) \overset{\mathcal{L}}{\Longrightarrow} N(0, \frac{\zeta^2}{9\beta}) \ \text{ as } \ n \to \infty. \tag{23}$$

To apply the CR method, we need to solve the relation $\left|\frac{\hat{\zeta}_n - \zeta}{\beta^{-1/2}\,\zeta}\right| \le \frac{1}{3}c_\alpha$ for ζ which is one of the cases explicitly addressed in Section 2. Thus, a $(1-\alpha)$ 100% confidence interval for ζ is $\zeta \in [\underline{C}, \overline{C}]$ where

$$\underline{C} = \frac{\hat{\zeta}_n}{1 + \beta^{-1/2}c_\alpha/3} \ \text{ and } \ \overline{C} = \frac{\hat{\zeta}_n}{1 - \beta^{-1/2}c_\alpha/3}.$$

The CRaN interval with approximate 95% confidence level for θ is $\theta \in [\frac{\underline{C}^3}{\beta}, \frac{\overline{C}^3}{\beta}]$.

Example 1.3 (Binomial, continued) *The normalizing transformation is $h(x) = \int_0^x [s(1-s)]^{-1/3}ds$. Then, eq. (18) reads $\sqrt{n}(\hat{\zeta}_n - \zeta) \overset{\mathcal{L}}{\Longrightarrow} N(0, [\theta(1-\theta)]^{1/3})$. To implement it, we need to solve*

$$-c_\alpha \le \frac{\int_\theta^{\hat{\theta}_n} [s(1-s)]^{-1/3}ds}{[\theta(1-\theta)]^{1/6}} \le c_\alpha.$$

As suggested by a reviewer, the integral $\int_0^x [s(1-s)]^{-1/3}ds$ can be expressed by $\Gamma(4/3)\Gamma(2/3)^{-2}$ multiplied by the cumulative distribution of Beta (2/3, 2/3) evaluated at x, while the endpoints can be found by the R *function* `uniroot()`*. There is an easier alternative to refine the Binomial CR interval, namely the split-sample method—see Remark 2.3; hence, we will not pursue the CRaN method for the Binomial in what follows.*

REMARK 3.1. **Normalization after studentization or variance stabilization.** What has been described so far is the standard approach towards a normalizing transformation for the statistic $\hat{\theta}_n$. Noting that studentization can exacerbate non-normality in finite samples, an alternative line of work has focused on devising normalizing transformations for $R_{ST} = \tau_n(\hat{\theta}_n - \theta)/\sigma(\hat{\theta}_n)$, i.e., the studentized quantity at the LHS of (3). For example, Johnson (1978) developed a monotone transformation of R_{ST} that reduces skewness; his proposal was based on estimated coefficients of the Cornish-Fisher expansion of R_{ST}. Hall (1992c) went about a similar construction based on estimated Edgeworth expansion coefficients. A little earlier, DiCiccio and Tibshirani (1987) had designed a different transformation where variance stabilization of $\hat{\theta}_n$ was followed by a skewness-reducing step; their approach seems also to be connected to the Cornish-Fisher expansion. Interestingly, the aim of DiCiccio and Tibshirani (1987) as well as Hall (1992c) was to improve and refine the construction of *bootstrap* confidence intervals; the former focused on the BCa intervals (see Efron and Tibshirani, 1993), while the latter focused on the studentized bootstrap—see the following remark for more details.

REMARK 3.2. **On the Bootstrap.** The bootstrap of Efron (1979) was a game-changer in modern statistics. For starters, it provided the original way out of the dilemma of studentization vs, variance stabilization as it was able to capture the whole distribution of θ_n, i.e., both its shape and its scale. By contrast, the paper at hand proposes the CR method as a simple—not computationally intensive—way out of the same old dilemma. Nevertheless, some discussion on the bootstrap is warranted; for conciseness, our discussion will exclude the aforementioned BCa intervals. To start, denote $R_n = \tau_n(\hat{\theta}_n - \theta)$ the LHS of (1), $R_{VS} = \tau_n(g(\hat{\theta}_n) - g(\theta))$ the LHS of (5), and $R_{CR} = \tau_n(\hat{\theta}_n - \theta)/\sigma(\theta)$ the LHS of (9). The naive bootstrap can produce multiple pseudo-copies of R_n, and thus approximate R_n's distribution by the empirical distribution of the pseudo-copies. However, it was soon realized that more accurate confidence intervals can be obtained by resampling an asymptotically pivotal quantity; see Hall (1988, 1992a) and the references therein. Note that R_{ST}, R_{VS} and R_{CR} are all asymptotically pivotal. Hence, resampling *any* of these three quantities will give confidence intervals that are 2nd order correct, i.e., more accurate than the normal approximation, and is therefore preferable to any of the intervals discussed so far (including the naive bootstrap). The new question is: if the practitioner is up for the computer simulation required to perform the bootstrap, which of the three quantities should they resample, R_{ST}, R_{VS} or R_{CR}? The theoretical analysis of Lahiri (1997) showed that there is no general preference between resampling R_{ST} vs. R_{VS}; in some problems R_{ST} is preferable—in others it is R_{VS}. However, Lahiri (1997) did not consider resampling R_{CR}, so this comparison is still open.

4. NUMERICAL COMPARISONS

In this section, we compare the aforementioned methods via simulation using the running examples, i.e., data from Poisson, Gamma or Binomial distributions; in the Gamma case, we assume that $\beta = 1$, in which case Gamma reduces to the Exponential distribution with mean θ.

As mentioned in the last section, we will not pursue the CRaN method in the Binomial case; instead, we include in the simulation the CR method applied to the split-sample estimator as described in Remark 2.3.

With regards to the Poisson example, the length of a confidence interval in CRaN method is nominal, i.e., the largest root of (22) minus the smallest root of (22); this shortcut is needed for the simulation, and is indicated as CRaN method* in Tables 1 and 2. The roots of the polynomial $[x-a]^4 - Cx \leq 0$ are derived through the R function `polyroot()`.

Tables 1–6 give the empirical coverage (CVR) and average length (LEN) of 95% confidence intervals based on $N = 10,000$ repetitions of each experiment. To discuss the results: for the Poisson case all four methods seem to do well regardless of sample size. In the Exponential case, the ST method underperforms for $n = 50$ but all four methods do well when $n = 100$. It is in the Binomial example that we see some major discrepancies. The worst performing overall seems to be the ST method; however, the VS method also appears very erratic with huge overcoverage for small θ, and undercoverage for $\theta = 1/2$. By contrast, the CR method does very well, in particular when used in tandem with the split-sample technique.

REMARK 4.1. Suppose $X_1, X_2, ..., X_n$ have an exponential distribution with mean θ. Then the (ST) method implies $\sqrt{n}\frac{\widehat{\theta}_n - \theta}{\widehat{\theta}} \xrightarrow{\mathcal{L}} N(0,1)$. Meanwhile, the (VS) method implies $\sqrt{n}(\log(\widehat{\theta}_n) - \log(\theta)) \xrightarrow{\mathcal{L}} N(0,1)$. The theoretical guarantee for (CR) is $\sqrt{n}\frac{\widehat{\theta} - \theta}{\theta} \xrightarrow{\mathcal{L}} N(0,1)$ and the theoretical guarantee for (CRaN) comes from $3\sqrt{n}\frac{\widehat{\theta}^{1/3} - \theta^{1/3}}{\theta^{1/3}} \xrightarrow{\mathcal{L}} N(0,1)$. Asymptotically all of these four results hold true; but in a finite sample situation, if the distribution of one of these statistics (i.e., $\sqrt{n}\frac{\widehat{\theta}_n - \theta}{\widehat{\theta}}$, $\sqrt{n}(\log(\widehat{\theta}_n) - \log(\theta))$, $\sqrt{n}\frac{\widehat{\theta} - \theta}{\theta}$ and $3\sqrt{n}\frac{\widehat{\theta}^{1/3} - \theta^{1/3}}{\theta^{1/3}}$) is closer to $N(0,1)$, then the associated confidence interval should have better coverage probability. As an illustration, we plot the cumulative distribution of those statistics in Figure 1. Statistics based on (CR) and (CRaN) methods have distributions that are close to the standard normal. On the other hand, the (ST) method has a distribution that significantly deviates from the standard normal as compared to the other methods; see also Remark 3.1.

5. APPLICATION: PROBABILITY DENSITY ESTIMATION

Let $X_1, \cdots, X_n$ be i.i.d. with probability density $f(\cdot)$ which is unknown (but assumed smooth). The kernel smoothed estimator of $f(x)$ at some particular point x that lies inside the support of $f(\cdot)$ is

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) \tag{24}$$

where the kernel $K(\cdot)$ is assumed (for simplicity) to be nonnegative, integrating to one, and being square-integrable, i.e.,

$$K(s) \geq 0, \ \int K(s)ds = 1, \quad \text{and} \ \int K(s)^2 ds < \infty. \tag{25}$$

A kernel $K(\cdot)$ satisfying (25) is called a *second-order* kernel. The bandwidth h is a function of n but will not be explicitly denoted as such.

Assume $f(\cdot)$ is (at least) twice continuously differentiable, and that $h \to 0$ but $hn \to \infty$ as $n \to \infty$. In addition, suppose

$$\int sK(s)ds = 0 \text{ and } \int s^2 K(s)ds < \infty \tag{26}$$

Then, a Taylor expansion yields

$$E\hat{f}(x) = f(x) + h^2 \frac{f''(x)}{2}\int s^2 K(s)ds + o(h^2) \tag{27}$$

and

$$Var f(x) = \frac{1}{nh} f(x)\int K(s)^2 ds + o(\frac{1}{nh}). \tag{28}$$

One can try to choose the bandwidth h with the goal of minimizing the Mean Squared Error (MSE) of $\hat{f}(x)$. Simple calculus shows that MSE-optimal estimation occurs with $h = C_f \, n^{-1/5}$ where

$$C_f = \left(\frac{f(x)\int K(s)^2 ds}{\left[f''(x)\int s^2 K(s)ds\right]^2} \right)^{1/5}. \tag{29}$$

Under standard conditions — see e.g. Ch. 32 of DasGupta (2008) — we further have

$$\tau_n(\hat{f}(x) - E\hat{f}(x)) \xrightarrow{\mathcal{L}} N(0, f(x)) \ \text{ as } \ n \to \infty \tag{30}$$

where $\tau_n = \sqrt{hn}\, \left[\int K(s)^2 ds\right]^{-1/2}$.

TABLE 1
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods with data from a **Poisson** *distribution with mean θ; sample size $n = 50$.*

$\theta =$	0.5		1		2		4	
	CVR	LEN	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	95.28%	0.3902	95.08%	0.5534	94.58%	0.7832	94.58%	1.1078
VS METHOD	94.46%	0.3902	94.46%	0.5534	94.80%	0.7832	94.81%	1.1078
CR METHOD	94.82%	0.3978	94.50%	0.5588	94.65%	0.7870	94.94%	1.1105
CRaN METHOD*	95.89%	0.3916	95.33%	0.5544	94.80%	0.7839	95.31%	1.1083

TABLE 2
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods with data from a **Poisson** *distribution with mean θ; sample size $n = 200$.*

$\theta =$	0.5		1		2		4	
	CVR	LEN	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	94.78%	0.1958	95.12%	0.2771	95.11%	0.3918	94.89%	0.5543
VS METHOD	95.01%	0.1958	95.35%	0.2771	95.21%	0.3918	94.84%	0.5543
CR METHOD	94.97%	0.1968	95.39%	0.2778	95.26%	0.3923	94.82%	0.5546
CRaN METHOD*	95.01%	0.1960	95.91%	0.2772	94.96%	0.3919	94.84%	0.5543

TABLE 3
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods with data from an **Exponential** *distribution with mean θ; sample size $n = 50$.*

$\theta =$	0.25		0.5		1		2	
	CVR	LEN	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	93.84%	0.1385	93.80%	0.2772	93.97%	0.5545	93.83%	1.1069
VS METHOD	94.68%	0.1403	94.46%	0.2807	94.78%	0.5616	94.80%	1.1211
CR METHOD	95.02%	0.1501	94.73%	0.3003	95.13%	0.6006	95.28%	1.1991
CRaN METHOD	94.85%	0.1425	94.70%	0.2852	94.86%	0.5705	95.07%	1.1390

TABLE 4
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods with data from an **Exponential** *distribution with mean θ; sample size $n = 200$.*

$\theta =$	0.25		0.5		1		2	
	CVR	LEN	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	94.65%	0.0692	94.74%	0.1387	94.30%	0.2772	94.48%	0.5541
VS METHOD	94.87%	0.0694	94.97%	0.1392	94.41%	0.2781	94.39%	0.5559
CR METHOD	94.76%	0.0706	95.07%	0.1414	94.52%	0.2826	94.52%	0.5650
CRaN METHOD	94.83%	0.0697	95.04%	0.1397	94.41%	0.2792	94.47%	0.5581

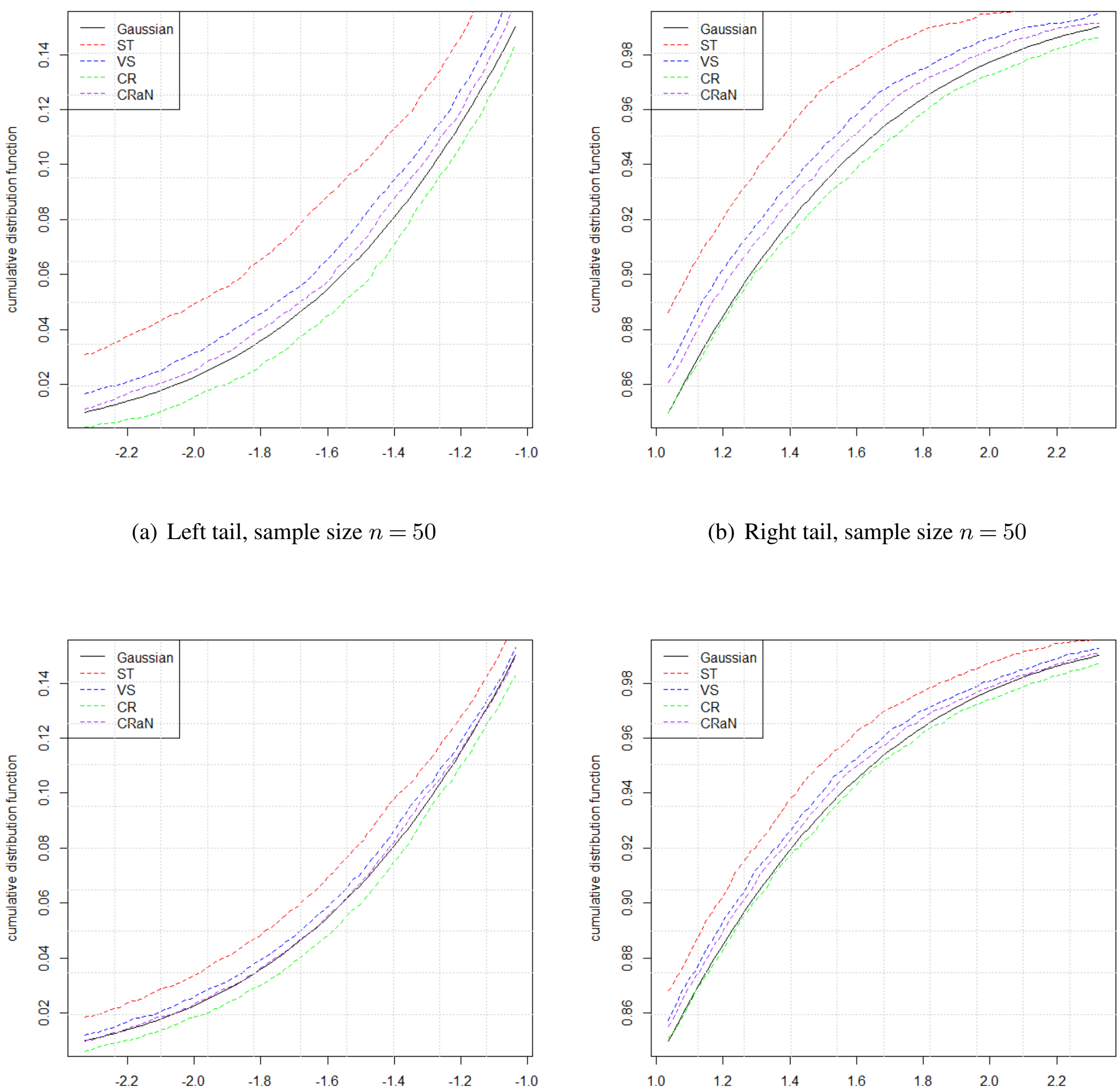
TABLE 5
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods with data from a **Bernoulli** *distribution with mean θ; sample size $n = 50$. For the split-sample estimator, the choice $m = 23$ was used.*

$\theta =$	0.07		0.13		0.25		0.5	
	CVR	LEN	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	86.58%	0.1341	89.20%	0.1821	93.93%	0.2368	93.67%	0.2744
VS METHOD	99.98%	0.2754	99.56%	0.2736	97.74%	0.2669	91.04%	0.2382
CR METHOD	97.87%	0.1454	96.56%	0.1840	95.27%	0.2313	93.67%	0.2646
CR with Split-Sample	96.48%	0.1454	96.14%	0.1840	95.30%	0.2313	95.25%	0.2646

TABLE 6

Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods with data from a **Bernoulli** *distribution with mean θ; sample size $n = 200$. For the split-sample estimator, the choice $m = 97$ was used.*

$\theta =$	0.07		0.13		0.25		0.5	
	CVR	LEN	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	93.19%	0.0699	92.89%	0.0927	93.99%	0.1196	94.45%	0.1382
VS METHOD	99.97%	0.1381	99.61%	0.1373	97.39%	0.1340	90.87%	0.1198
CR METHOD	96.38%	0.0712	95.14%	0.0929	96.18%	0.1189	94.45%	0.1369
CR with Split-Sample	95.54%	0.0712	95.13%	0.0929	95.14%	0.1189	94.78%	0.1369



(a) Left tail, sample size $n = 50$

(b) Right tail, sample size $n = 50$

(c) Left tail, sample size $n = 200$

(d) Right tail, sample size $n = 200$

FIG 1. *Cumulative distribution functions of different quantities obtained from data following an Exponential distribution. The notation is ST: $\sqrt{n}\frac{\widehat{\theta}_n-\theta}{\widehat{\theta}}$, VS: $\sqrt{n}(\log(\widehat{\theta}_n) - \log(\theta))$, CR: $\sqrt{n}\frac{\widehat{\theta}-\theta}{\theta}$ and CRaN: $3\sqrt{n}\frac{\widehat{\theta}^{1/3}-\theta^{1/3}}{\theta^{1/3}}$. The black line plots the cumulative distribution function of a Gaussian $N(0,1)$ random variable. The distribution of the ST quantity significantly deviates from a Gaussian distribution compared to the others.*

5.1 Estimating the MSE-optimal bandwidth using flat-top kernels

To use the MSE-optimal bandwidth, we need to estimate $f(x)$ and $f''(x)$ and plug them in eq. (31); let $\tilde{f}(x)$ be an estimate of $f(x)$ for the purpose of plugging in eq. (31), and let $\tilde{f}''(x)$ be its 2nd derivative at point x. Then, we can let $\tilde{h} = \tilde{C}_f\ n^{-1/5}$ where

$$(31)\qquad \tilde{C}_f = \left(\frac{\tilde{f}(x) \int K(s)^2 ds}{\left[\tilde{f}''(x) \int s^2 K(s) ds\right]^2} \right)^{1/5}$$

provided the denominator does not vanish. In order for the MSE of $\hat{f}(x)$ using bandwidth $\tilde{h}$ to be asymptotically the same as the MSE of $\hat{f}(x)$ using the oracle $h = C_f\ n^{-1/5}$, the estimator $\tilde{f}(x)$ must have a faster rate of convergence than $\hat{f}(x)$; this can be accomplished by basing $\tilde{f}(x)$ on a *higher-order* kernel[3] but this would then require having a way to choose the bandwidth of $\tilde{f}(x)$.

It turns out that there is a class of higher-order (actually, infinite-order) kernels, the so-called *flat-top* kernels, that (a) achieve the fastest rate of convergence in a given smoothness class; see e.g. Politis (2001); and (b) it is straightforward to choose their bandwidth using a graphical tool; see Politis (2003), and the R package `iosmooth()`. Hence, one may optimally construct $\hat{f}(x)$ using $\tilde{h} = \tilde{C}_f\ n^{-1/5}$, with $\tilde{C}_f$ obtained as detailed in Politis (2003), e.g., based on a trapezoidal or other choice of flat-top kernel.

5.2 Confidence intervals via 'undersmoothing'

Eq. (30) can be used to construct confidence intervals for the center of the asymptotic normal distribution which is $E\hat{f}(x)$. To turn these into intervals for $f(x)$ there have been two general approaches in the literature: 'undersmoothing' vs. explicit bias correction.

Under an 'undersmoothed' choice of bandwidth $h = o(n^{-1/5})$, it follows that $\tau_n(E\hat{f}(x) - f(x)) \to 0$, and we can write

$$(32)\qquad \tau_n(\hat{f}(x) - f(x)) \stackrel{\mathcal{L}}{\Longrightarrow} N(0, f(x)) \ \text{ as } \ n \to \infty$$

Note that this falls under the framework of Case 1 in Section 2. Hence, an asymptotic $(1-\alpha)$ 100% confidence interval based on the CR method is given by eq. (11) with $\hat{\theta}_n = \hat{f}(x)$, and $\theta = f(x)$. In addition, the ST interval (4) and the VS interval (7) can be constructed as well; the variance stabilizing transformation here is $g(x) = \sqrt{x}$, as in the Poisson example.

Undersmoothing has the advantage of simplicity: we just ignore the bias. Nevertheless, although the bias of $\hat{f}(x)$ is asymptotically negligible here, it may present problems in finite samples. In addition, there is no recommendation on how we should choose h since the requirement $h = o(n^{-1/5})$ is rather vague.

5.3 Optimal confidence intervals via bias correction with flat-top kernels

Hall (1992b) compared 'undersmoothing' to explicit bias correction for confidence intervals in this setting, and concluded that 'undersmoothing' is preferable. However, to perform the bias correction, Hall (1992b) estimated $f''(x)$ using a second-order kernel with a possibly different bandwidth. In retrospect, it is not hard to see is why problems, both theoretical and practical, arose in his construction. We now show how to construct confidence intervals based on the MSE-optimal bandwidth and an explicit bias correction; the key is to use *flat-top kernels* (with their own bandwidth) in order to estimate the proportionaliy constant in the bias expansion just as in Section 5.1.

Note that eq. (30) brings us to the set-up of Remark 2.4 with $\theta = f(x)$ and $\hat{\theta}_n = \hat{f}(x)$ using a bandwidth h of optimal order, i.e., proportional to $n^{-1/5}$. Then,

$$(33)\qquad \tau_n(E\hat{f}(x) - f(x)) \to \tau_n h^2 \frac{f''(x)}{2} \int s^2 K(s) ds \equiv b(\delta)$$

where the above serves as the definition of $b(\delta)$, and $\delta = f''(x)$. Now let $\hat{\delta}_n = \tilde{f}''(x)$ where $\tilde{f}$ is a flat-top estimator of f with bandwidth chosen as detailed in Politis (2003a). Then, the procedure outlined in Remark 2.4 can be followed *verbatim* leading to bias-corrected confidence intervals for θ using any of the three methods: ST, VS or CR; the latter would follow the framework of Case 1 in Section 2. Most importantly, the data-based optimal bandwidth $\tilde{h} = \tilde{C}_f\ n^{-1/5}$ can be used throughout this construction, both for $\hat{f}(x)$ and for τ_n. Note that $\tilde{h}$ can be obtained via section 3.2 of Politis (2003a); and the optimally tuned flat-top estimator of $f''(\cdot)$ needed to estimate the bias can be obtained via eq. (17) of Politis (2003a); see our Appendix A for details.

EXAMPLE 5.1 (Normal density estimation). We generate i.i.d. standard normal random variables $X_1, ..., X_n$, then use the kernel estimator (24) to estimate the density at $x = 0.5$. We adopt the three methods (e.g., studentization, variance stabilization and the confidence region method) to construct confidence intervals. The kernel K is chosen as the standard normal density, i.e., $K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$, so $\int K(s) ds = 1$ and $\int K^2(s) ds = \frac{1}{2\sqrt{\pi}}$. We use section 3.2 of Politis (2003) to construct the optimal bandwidth $\widetilde{h}$.

Our simulation considers all three situations:

(1) 'undersmoothed', in which case the bandwidth should

---

[3] To estimate $f''(x)$ accurately, we would need to additionally assume that $f(\cdot)$ is (at least) four times continuously differentiable, with 4th derivative satisfying a Lipschitz condition.

TABLE 7

Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods. The data are generated by normal random variables with mean 0 *and variance* 1*. We estimate the density at point* $x = 0.5$*. The sample size is* $n = 200$ *and the number of repetitions is* 1000*. In this and the latter tables, 'Without BC' is short for 'Without bias-correction'*

	Undersmoothed		Bias-corrected		Without BC	
Data type	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	94.8%	0.299	96.5%	0.130	92.3%	0.130
VS METHOD	95.8%	0.299	97.1%	0.130	93.6%	0.130
CR METHOD	96.5%	0.306	97.0%	0.131	95.1%	0.131

have order $o(\widetilde{h})$–to practically illustrate that, we use $\widetilde{h}' = \widetilde{h}/5$ in (24).
(2) 'Bias-corrected', i.e., we use the optimal bandwidth $\widetilde{h}$ in (24) and apply the bias-correction technique in section 5.3 to construct confidence intervals.
(3) 'Without bias-correction', in which case we use $\widetilde{h}$ as in (24) but do not apply bias-correction techniques in creating confidence intervals.
The results are demonstrated in table 7.

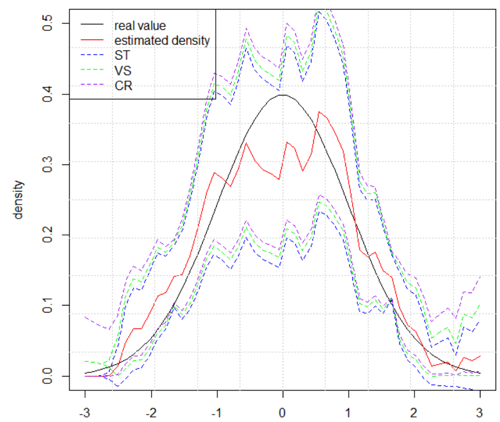
A visual representation of these processes is provided in figure 2 that plots the true density $f(x)$, the different confidence intervals of level 95%, as well as the estimator on which the confidence intervals are based, i.e., the center of the intervals. The confidence intervals are pointwise, meaning a 95% confidence interval was constructed at each of a finite number of x points. Figure 2 is constructed from just one of the realizations of the data process; its purpose is to illustrate the issues at hand.

The large width of the undersmoothed intervals is apparent but also their unusual/unsmooth shape as a function of x. Table 7 confirms that the uncorrected intervals tend to undercover. Both the undersmoothed and the bias-corrected intervals achieve good coverage but the latter have half the (average) width, so they are preferable.

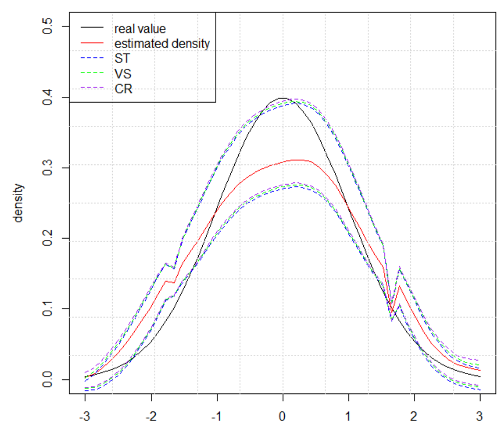
EXAMPLE 5.2 (χ^2 density estimation). In this example, we generate i.i.d. random variables $X_1, ..., X_n$ with X_1 having a χ^2 distribution with 5 degrees of freedom. The result is demonstrated in figure 3 and table 8. As before the width of the undersmoothed confidence intervals is too large compared to the bias-corrected confidence intervals while both construction yield good coverage. As in example 5.1, the confidence intervals without bias-correction tend to have undercoverage issues.

EXAMPLE 5.3 (Mixed normal density estimation). In this example, we generate i.i.d. random variables $X_1, ..., X_n$ from a mixed normal density given by

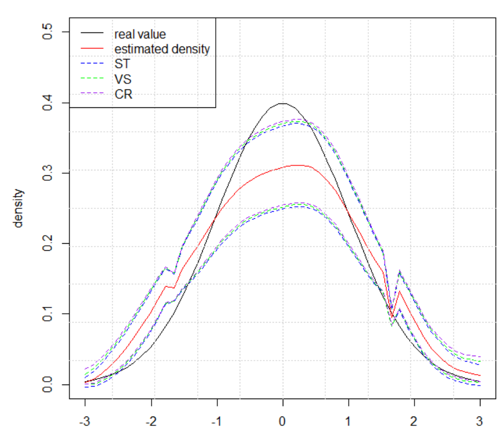
$$h(x) = 0.3f_1(x) + 0.3f_2(x) + 0.4f_3(x); \tag{34}$$



(a) Undersmooth

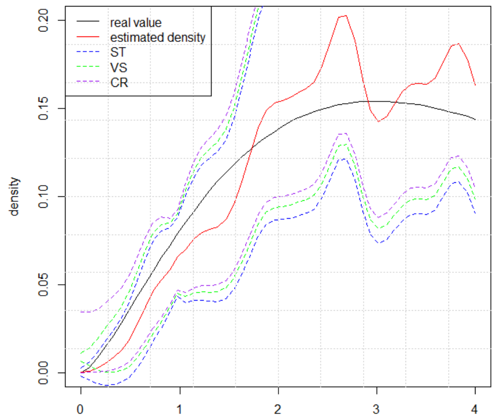


(b) Bias-corrected

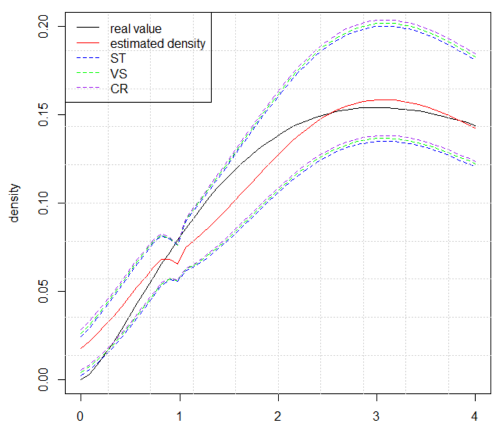


(c) Without bias-correction

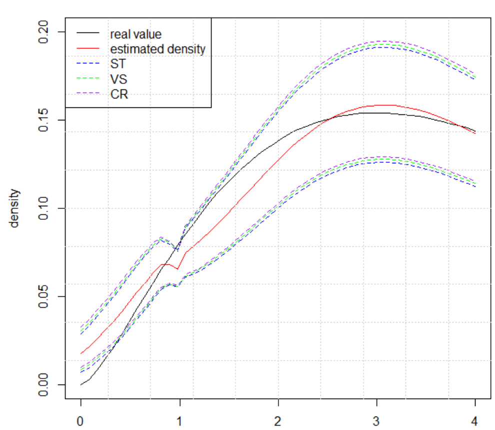
FIG 2. 95% *point-wise confidence intervals for the kernel density estimator. The setting coincides with table 7.*



(a) Undersmooth



(b) Bias-corrected



(c) Without bias-correction

FIG 3. 95% *point-wise confidence intervals for the kernel density estimator. The setting coincides with table 8.*

TABLE 8

Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods. The data is generated by χ^2 distribution with 5 degrees of freedom. We estimate the density at $x = 3.0$. The sample size is $n = 200$ and the number of repetitions is 1000.

	Undersmooth		Bias-corrected		Without BC	
Data type	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	94.9%	0.140	96.2%	0.061	88.8%	0.061
VS METHOD	95.7%	0.140	96.7%	0.061	90.5%	0.061
CR METHOD	96.2%	0.144	96.6%	0.061	92.6%	0.061

here $f_1(x)$ is a normal density with mean -2 and variance 1, $f_2(x)$ is a normal density with mean 1 and variance 4, $f_3(x)$ is a normal density with mean 2 and variance 1.

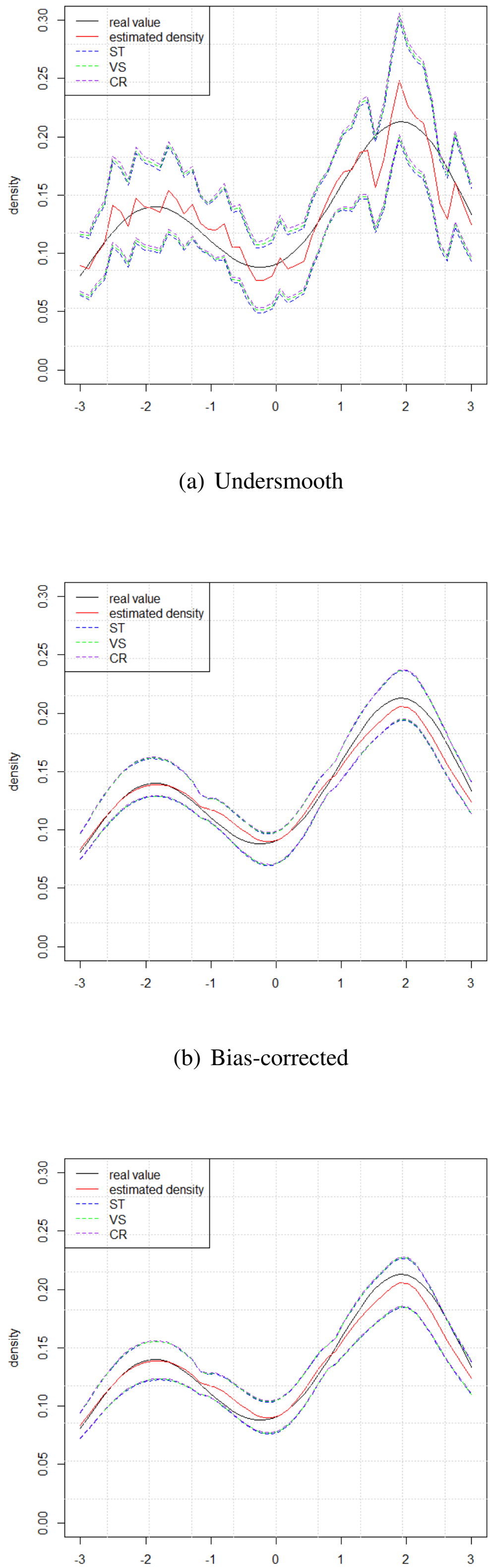
The mixed normal example is meant to be more challenging than the previous two. To make it even more challenging, we estimate the density at $x = 2.0$ which is the point of maximum curvature, and therefore maximum (absolute) bias. The results are shown in figure 4 and table 9. Bias correction gives an improvement here as well but does not fully capture the large bias—resulting in undercoverage. Undersmoothing works as before; it gives intervals with coverage close to 95% at the expense of intervals that are more than double the width. This example was chosen to show that there are no panaceas in difficult problems such as nonparametric function estimation; the practitioner must take into account the particulars of the dataset at hand, and make informed choices of all quantities involved, first and foremost the bandwidth(s).

TABLE 9

Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals according to different methods. The data is generated by mixed normal distribution described in example 5.3. We estimate the density at $x = 2.0$. The sample size is $n = 500$ and the number of repetitions is 1000.

	Undersmoothed		Bias-corrected		Without BC	
Data type	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	94.2%	0.141	87.2%	0.061	81.7%	0.061
VS METHOD	94.8%	0.141	88.2%	0.061	83.2%	0.061
CR METHOD	95.5%	0.143	88.5%	0.061	84.4%	0.061

As a summary, it appears that the three methods ST, VS, and CR are roughly comparable in the density estimation simulation experiments. Here, the question on whether to employ ST or VS (or CR) appears to be overshadowed by a new dilemma: whether to employ undersmoothing or an explicit bias-correction in the confidence interval construction.



(a) Undersmooth

(b) Bias-corrected

(c) Without bias-correction

FIG 4. 95% *point-wise confidence intervals for the kernel density estimator. The setting coincides with table 9.*

Notably, the undersmoothed confidence intervals achieve the desired coverage probability but have large width and unusual functional shape—see Figures 2 to 4. In addition, the undersmoothed intervals are not centered well (since they are centered at expectation and not at the true value). On the other hand, the bias-corrected confidence intervals are centered well but may yield undercoverage in some "difficult" examples such as the mixed normal at the point of its highest curvature. Importantly, their width is significantly smaller—less than half—compared to the undersmoothed intervals.

The comparison between the bias-corrected confidence intervals and the confidence intervals without bias-correction shows that the bias of eq. (33) is not negligible if the practitioner wants to adopt the optimal bandwidth, e.g., the bandwidth chosen via section 3.2 of Politis (2003); when employing the optimal rate for the bandwidth, the bias-correction procedure is definitely needed in order to obtain a consistent confidence interval.

All in all, it seems that the bias corrected intervals (using optimal smoothing) may be preferable to the undersmoothed ones. This goes against the recommendation of Hall (1992b) but note that our bias correction is achieved using accurate bias estimates derived from flat-top kernels with their own optimal bandwidth. By contrast, Hall (1992b) used bias estimates derived from a second order kernel with suboptimal bandwidth which is an outdated technology.

REMARK 5.1. **On Bias-Correction, continued.** Following up the discussion of Remark 2.4, note that even if the oracle interval (16) is fine-tuned to have exact coverage $1-\alpha$, the practical BC interval (17) will have lower coverage, i.e., it will undercover. For the BC interval to have correct coverage, its endpoints would need to be adjusted to capture the variance of the estimated shift $b(\hat{\delta}_n)$ that is used in place of the true shift $b(\delta)$. Recall from eq. (33) that $\delta = f''(x)$, and $\hat{\delta}_n = \tilde{f}''(x)$ where $\tilde{f}$ is a flat-top estimator of f.

Motivated by a suggestion of a reviewer, note that eq. (14) and (15) imply that

(35)
$$r_{CR,b} = \tau_n \frac{\hat{\theta}_n - \theta - b(\delta)}{\sigma(\theta)} \overset{\mathcal{L}}{\Longrightarrow} N(0,1) \quad \text{as} \quad n \to \infty.$$

Recall that we have assumed that $b(\cdot)$ is a continuous function, and $\hat{\delta}_n$ is consistent for $\hat{\delta}$. Hence, we also have

(36)
$$R_{CR,b} = \tau_n \frac{\hat{\theta}_n - \theta - b(\hat{\delta}_n)}{\sigma(\theta)} \overset{\mathcal{L}}{\Longrightarrow} N(0,1) \quad \text{as} \quad n \to \infty.$$

Recall that the BC interval (17) is built based on (36). Although they have the same asymptotic variance, $R_{CR,b}$ is expected to have bigger finite-sample variance compared to $r_{CR,b}$ due to the variability of the random term $b(\hat{\delta}_n)$. It is generally intractable to explicitly account for the variance of $b(\hat{\delta}_n)$ —as well as its covariance with $\hat{\theta}_n$— so that the BC confidence interval is adjusted appropriately. However, a bootstrap scheme could be devised to emulate the distribution of the quantity $R_{CR,b}$. The resulting bootstrap BC confidence interval is expected to have improved coverage compared to interval (17) because it will (partially) account for the variability of $b(\hat{\delta}_n)$. The details are nontrivial and could be the focus of a future work focusing on the bootstrap.

6. APPLICATION: SPECTRAL DENSITY ESTIMATION

Let $X_1, \cdots, X_n$ be a stretch of a strictly stationary time series with mean 0, autocovariance $\gamma(k) = Cov(X_0, X_k)$ that is absolutely summable, and spectral density

$$f(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) \exp(ik\omega)$$

Define the periodogram

$$I(\omega) = \frac{1}{2\pi} \sum_{k=-n}^{n} \hat{\gamma}(k) \exp(ik\omega)$$
$$\text{with } \hat{\gamma}(k) = \frac{1}{n} \sum_{t=1}^{n-|k|} X_t X_{t+|k|} \tag{37}$$

Note that $I(\omega)$ is approximately unbiased for $f(\omega)$ for $\omega \in (0, \pi)$ if $\sum_{k=0}^{\infty} k|\gamma(k)| < \infty$; however, it is not consistent. In fact, under standard conditions,

$$\frac{I(\omega)}{f(\omega)} \overset{\mathcal{L}}{\Longrightarrow} \exp(1) \ \text{ as } \ n \to \infty$$

for any fixed $\omega \in (0, \pi)$; see e.g. Proposition 10.3.2 in Brockwel and Davis (1991).

To create a consistent estimator of $f(\omega)$ for some fixed $\omega \in [0, \pi]$, we can taper the sample autocovariances before creating the Fourier series, i.e., let

$$I_{weight}(\omega) = \frac{1}{2\pi} \sum_{k=-n}^{n} \widehat{\gamma}(k) \times A(kh) \exp(ik\omega)$$

here $A(\cdot) : \mathbf{R} \to [0, \infty)$ is an even, Lipschitz continuous function with support $[-1, 1]$ and $A(0) = 1$. In the time series literature, $A(\cdot)$ is called a 'lag-window', see e.g., Politis and Romano (1995).

If we define $W(\omega) = \frac{1}{2\pi} \sum_{k=-n}^{n} A(kh) \exp(\mathrm{i} \times k\omega)$, then we have

$$I_{weight}(\omega) = \int_{-\pi}^{\pi} I(x) W(\omega - x) dx. \tag{38}$$

$W(\omega)$ is called the 'spectral window'. Since convolution is a smoothing operation, it follows that I_{weight} is a smoothed version of the periodogram. Assume $\lim_{x \to 0} x^{-2} \times (1 - A(x)) = c_A \neq 0$ exists, $\sum_{k \in \mathbf{Z}} k^2 |\gamma(k)| < \infty$, and $\frac{1}{h} = o(n^{1/3})$. Then,

$$\frac{1}{h^2} (\mathbf{E} I_{weight}(\omega) - f(\omega)) \to c_A f''(\omega); \tag{39}$$

see Shao and Wu (2007) who also show under standard conditions that

$$\tau_n \left(I_{weight}(\omega) - \mathbf{E} I_{weight}(\omega) \right) \overset{\mathcal{L}}{\Longrightarrow} N(0, f^2(\omega)) \tag{40}$$

where

$$\tau_n = \sqrt{\frac{nh}{(1 + \eta(\omega)) \times \int_{-1}^{1} A^2(x) dx}}$$

and

$$\eta(\omega) = \begin{cases} 1 \text{ if } \omega = k\pi, k \in \mathbf{Z} \\ 0 \text{ otherwise.} \end{cases}$$

Notice that

$$\begin{aligned} &\tau_n \left(I_{weight}(\omega) - f(\omega) \right) \\ &= \tau_n \left(I_{weight}(\omega) - \mathbf{E} I_{weight}(\omega) \right) \\ &+ \tau_n \left(\mathbf{E} I_{weight}(\omega) - f(\omega) \right). \end{aligned} \tag{41}$$

if we adopt an 'under-smoothing' choice of bandwidth, i.e., $h = o(n^{-1/5})$, then the bias of $I_{weight}(\omega)$ is asymptotically negligible, and we can write

$$\tau_n \left(I_{weight}(\omega) - f(\omega) \right) \overset{\mathcal{L}}{\Longrightarrow} N(0, f^2(\omega)) \quad \text{as } nh \to \infty \tag{42}$$

According to (38)—and in particular its Riemann sum approximation over the grid of Fourier frequencies—, $I_{weight}(\omega)$ is a weighted average of periodogram ordinates $I(x)$ that are asymptotically independent and exponentially distributed; see e.g. Ch. 9 of McElroy and Politis (2020). Hence, it is hardly surprising that the limit law (42) falls under the framework of Case 2 in Section 2. Consequently, an asymptotic $(1 - \alpha)$ 100% confidence interval for $\theta = f(\omega)$ based on the CR method is given by eq. (12) with $\hat{\theta}_n = I_{weight}(\omega)$. In addition, the ST interval (4) and the VS interval (7) can be constructed as well; the variance stabilizing transformation here is $g(x) = \log x$, as in the Gamma example.

Note that the MSE-optimal bandwidth in constructing estimator $I_{weight}(\omega)$ has the order $n^{-1/5}$; see Politis (2003) for details. Hence, we could undersmooth, by using a bandwidth of order $o(n^{-1/5})$. If we wish to use a bandwidth that has order $n^{-1/5}$, then, according to (39), the bias would not be negligible; in this case, we would need to estimate the bias and construct the associated *bias-corrected* confidence intervals analogously to Section 5.3. To estimate the bias, this section adopts the estimator proposed in Politis (2003) based on a flat-top kernel estimate of $f''(\omega)$.

REMARK 6.1. As in Section 5.1, to estimate $f''(\cdot)$ accurately we would need to additionally assume that $f(\cdot)$ is (at least) four times continuously differentiable, with 4th derivative satisfying a Lipschitz condition.

In order to use the results in Politis (2003) and Shao and Wu (2007), the lag-window $A(\cdot)$ should be even, Lipschitz continuous, and yield a nonnegative spectral window (i.e., $W(\omega) \geq 0$ for all ω). These conditions are easily achievable, e.g., the Parzen kernel

$$A(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, \ |x| < 1/2 \\ 2 \times (1 - |x|)^3, \ 1/2 \leq |x| < 1 \\ 0, \ \text{otherwise} \end{cases}$$

can meet all requirements; see e.g. Ch. 10 of Brockwell and Davis (1991).

Note that eq. (40) brings us to the set-up of Remark 2.4 with $\theta = f(\omega)$ and $\hat{\theta}_n = I_{weight}(\omega)$ if we use a bandwidth h of optimal order, i.e., proportional to $n^{-1/5}$. Then,

$$\tau_n \left(\mathbf{E} I_{weight}(\omega) - f(\omega) \right) \to \tau_n h^2 c_A f''(\omega) \equiv b(\theta)$$

where the above serves as the definition of $b(\theta)$ in the spectral density case. Now let $\tilde{\theta}_n = \tilde{I}(\omega)$ where $\tilde{I}(\omega)$ is a flat-top estimator of f with bandwidth chosen as detailed in Politis (2003).

The procedure outlined in Remark 2.4 can now be followed verbatim leading to bias-corrected confidence intervals for θ using any of the three methods: ST, VS or CR; the latter would follow the framework of Case 2 in Section 2. Importantly, the data-based optimal bandwidth $\widetilde{h}$ can be used for I_{weight} (and for τ_n) throughout this construction. In practice, the bandwidth $\widetilde{h}$ can be obtained via section 2.2 of Politis (2003), using an optimally tuned flat-top estimator of $f''(\cdot)$; this procedure would be based on eq. (6) in Politis (2003) but see our Appendix A for a clarification.

EXAMPLE 6.1 (Moving average process). Suppose the data $X_1, ..., X_n$ come from a moving average process $X_i = \epsilon_i + 0.9\epsilon_{i-1} - 0.5\epsilon_{i-2} - 0.3\epsilon_{i-3}$ where ϵ_i are i.i.d. standard normal. In this case, the spectral density is given by

$$f(\omega) = \frac{1}{2\pi} |1 + 0.9e^{-\mathrm{i}\omega} - 0.5e^{-2\mathrm{i}\omega} - 0.3e^{-3\mathrm{i}\omega}|^2$$

We estimate the spectral density at point $\omega = \pi/3$ with a sample size of 400. The result is demonstrated in table 10. The bandwidth $\widetilde{h}$ for the 'bias correction method' is chosen using section 2.2 of Politis (2003); while the bandwidth for the 'undersmooth' method is chosen simply as $\widetilde{h}/2.0$.

A visual representation of these processes is provided in figure 5 that plots the spectral density $f(\omega)$ for $\omega \in [-\pi, \pi]$, the different confidence intervals of level 95%, as well as the estimator on which the confidence intervals are based, i.e., the center of the intervals. The confidence intervals are point-wise, meaning a 95% confidence interval was constructed at each of the Fourier frequencies. Figure 5 is constructed from just one of the 1000 realizations of the process; its purpose is to illustrate the issues at hand.

The large width of the undersmoothed intervals is apparent but also their unusual/unsmooth shape as a function of ω. For the particular realization involved, there is little pictorial difference between the bias corrected and the uncorrected processes although we know —both from theory as well as table 10– that the uncorrected confidence intervals will tend to undercover.

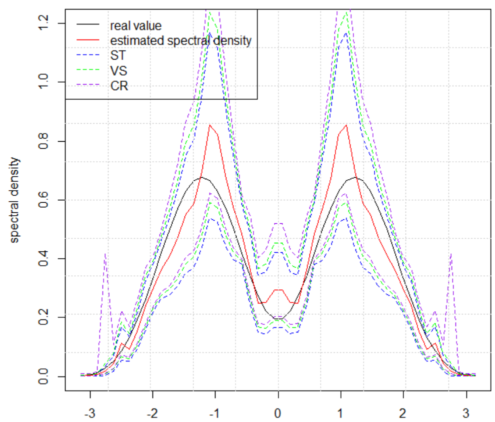
TABLE 10
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals for the spectral density estimator of the moving average process. The data generation mechanism coincides with example 6.1. We estimate the spectral density at $\omega = \pi/3$. The sample size is $n = 400$ and the number of repetitions is 1000.

	Undersmooth		Bias-corrected		Without BC	
Data type	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	89.2%	0.465	95.3%	0.316	80.4%	0.316
VS METHOD	93.4%	0.476	94.2%	0.319	83.2%	0.319
CR METHOD	95.3%	0.541	93.0%	0.339	86.5%	0.339

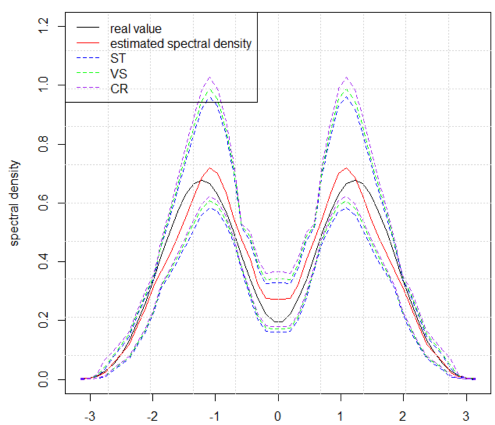
EXAMPLE 6.2 (Autoregressive process). Suppose $X_1, ..., X_n$ satisfy an autoregressive process $X_i = 0.7X_{i-1} + \epsilon_i$ where ϵ_i are i.i.d. standard normal. In this example, the true spectral density is $f(\omega) = \frac{1}{2\pi} \times \frac{1}{1.49 - 1.4\cos(\omega)}$. We estimate the spectral density at point $\omega = \pi/3$. The sample size is 400 and the bandwidth $\widetilde{h}$ for 'bias correction method' is chosen via section 2.2 of Politis (2003); the bandwidth for 'undersmooth method' is chosen as $\widetilde{h}/2.0$. The result is demonstrated in figure 6 and table 11.

TABLE 11
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals for the spectral density estimator of the autoregressive process. The data generation mechanism coincides with example 6.2. We estimate the spectral density at $\omega = \pi/3$. The sample size is $n = 400$ and the number of repetitions is 1000.

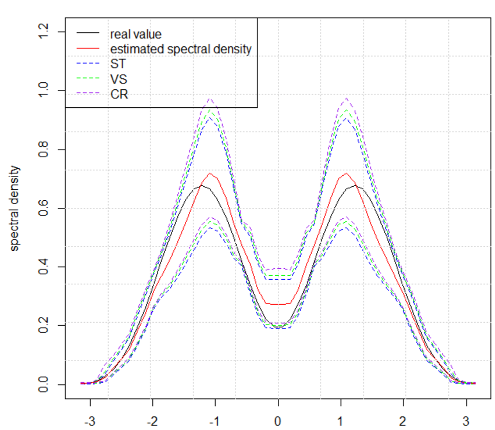
	Undersmooth		Bias-corrected		Without BC	
Data type	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	93.0%	0.467	93.8%	0.343	90.3%	0.343
VS METHOD	94.3%	0.490	95.0%	0.351	88.4%	0.351
CR METHOD	93.1%	0.680	95.1%	0.403	86.4%	0.403



(a) Undersmooth

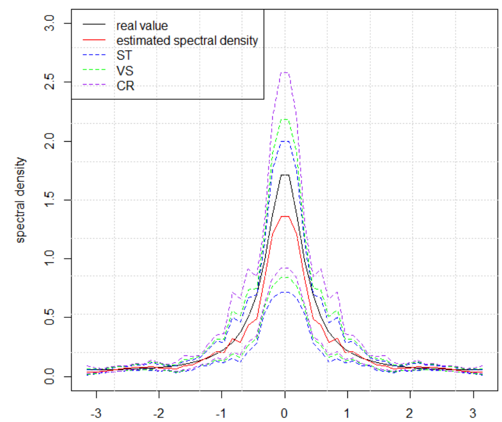


(b) Bias correction

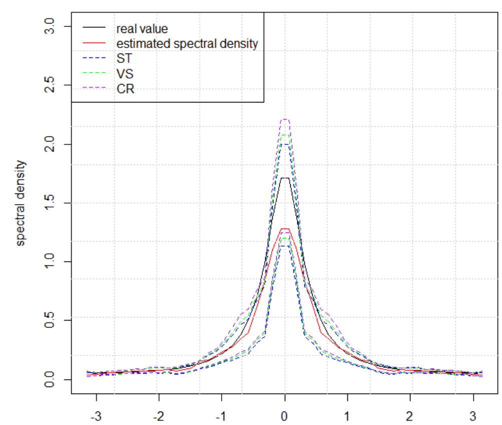


(c) Without bias correction

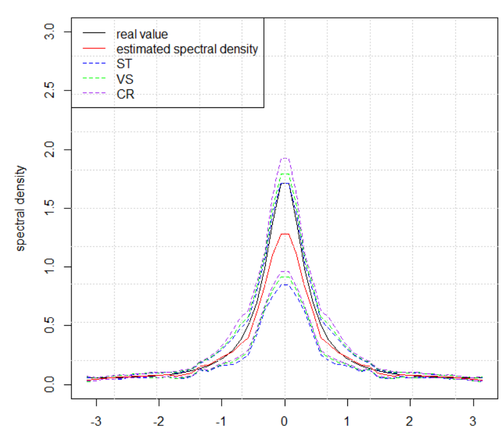
FIG 5. 95% *point-wise confidence intervals for the kernel spectral density estimator. The setting coincides with example 6.1.*



(a) Undersmooth



(b) Bias correction



(c) Without bias correction

FIG 6. 95% *point-wise confidence intervals for the kernel spectral density estimator. The setting coincides with example 6.2.*

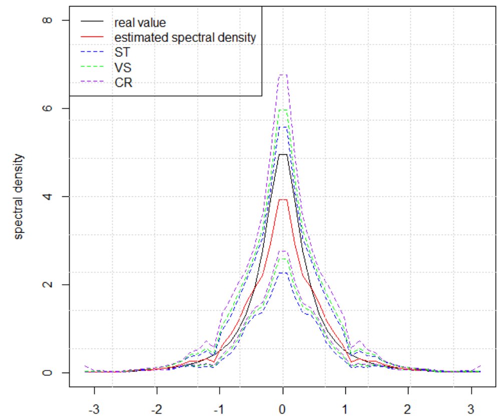
EXAMPLE 6.3 (ARMA model). Suppose $X_1, ..., X_n$ satisfy an ARMA process $X_i = 0.7X_{i-1} + \epsilon_i + 0.7\epsilon_{i-1}$ where ϵ_i are i.i.d. standard normal. In this case, the spectral density is given by $f(\omega) = \frac{1}{2\pi} \times \frac{1.49+1.4\cos(\omega)}{1.49-1.4\cos(\omega)}$. We estimate the spectral density at point $\omega = \pi/3$ with a sample size of 400. The result is demonstrated in table 12 and figure 7. The bandwidth $\widetilde{h}$ for 'bias correction method' is chosen using section 2.2 of Politis (2003), while the bandwidth for the 'undersmooth' method is again chosen as $\widetilde{h}/2.0$.

TABLE 12
Empirical coverage (CVR) and average length (LEN) of 95% confidence intervals for the spectral density estimator of the autoregressive process. The data generation mechanism coincides with example 6.3. We estimate the spectral density at $\omega = \pi/3$. The sample size is $n = 400$ and the number of repetitions is 1000.

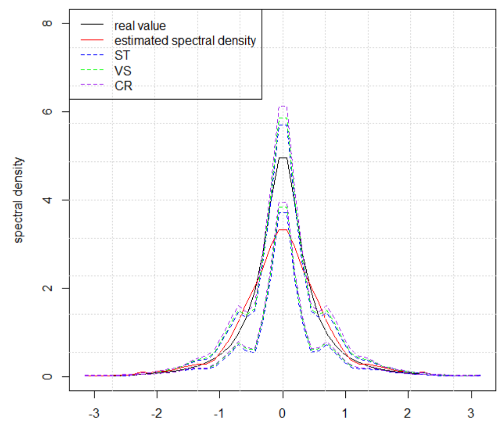
	Undersmooth		Bias-corrected		Without BC	
Data type	CVR	LEN	CVR	LEN	CVR	LEN
ST METHOD	93.0%	0.198	93.3%	0.146	88.7%	0.146
VS METHOD	93.4%	0.206	95.3%	0.149	88.2%	0.149
CR METHOD	94.4%	0.271	95.2%	0.168	86.4%	0.168

The numerical experiments portray a similar situation as in section 5.3, i.e., the undersmoothed confidence intervals have coverage probability close to nominal but large width and undesirable functional shape. In addition, the undersmoothed intervals are not centered well (since they are centered at expectation and not at the true value). By contrast, the bias-corrected confidence intervals are centered well, they have smaller width and acceptable coverage—albeit with a tendency towards undercoverage. The situation is analogous to the probability density case; it looks like optimal smoothing with bias correction is preferable to undersmoothing as regards confidence interval construction. As always, optimal smoothing without bias correction is never recommendable.

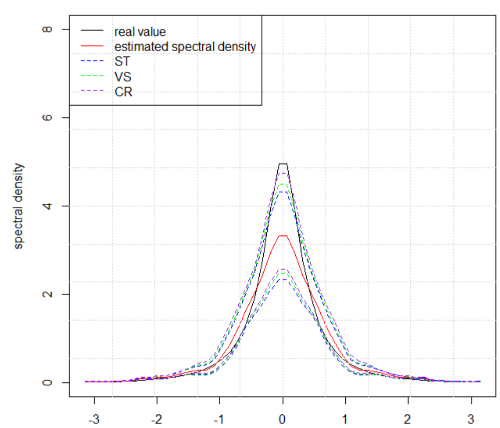
Finally, a similar discussion as in Remark 5.1 applies in the spectral density example as well, although in the time series setup the resampling schemes are more intricate; see e.g. Lahiri (2003), Politis (2003b), McElroy and Politis (2020), or Kreiss and Paparoditis (2023).



(a) Undersmooth



(b) Bias correction



(c) Without bias correction

FIG 7. 95% *point-wise confidence intervals for the kernel spectral density estimator. The setting coincides with example 6.3.*

Acknowledgment. This paper is dedicated to the memory of Dr. Dimitrios Gatzouras, a brilliant mathematician and wonderful friend. His untimely passing in the Fall of 2020 has left a rift that is hard to fill personally but also professionally, as he had always been so generous with his time in trying to help others with their work, including an early version of the paper at hand. Sincere thanks are also due to Yunyi Zhang and Jiang Wang for carrying out the numerical work in Sections 5 and 6 respectively. Special acknowledgement is due to Yunyi Zhang who compiled new R functions to compute the required flat-top estimates of 2nd derivatives; these new functions are now included in the R package `iosmooth()`. Many thanks are also due to the Editor, Associate Editor, and two reviewers for several constructive comments. This research was partially supported by NSF grant DMS 19-14556.

REFERENCES

[1] Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with Discussion), *J. Royal Statistical Society, Ser. B*., vol. 26, no. 2, pp. 211-252.

[2] Brown, L.D., Cai, T.T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion, *Statistical Science*, Vol. 16, No. 2, pp. 101-133.

[3] Brockwell, P.J. and Davis, R.A. (1991). *Time series: theory and methods*, 2nd edition, Springer, New York.

[4] Casella, G. and Robert, C. (1989). Refining Poisson confidence intervals. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, vol. 17, no. 1, pp. 45âĂŞ57.

[5] Crow, E.L. and Gardner, R.S. (1959). Confidence intervals for the expectation of a Poisson variable. *Biometrika*, 46(3/4), pp. 441âĂŞ453.

[6] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer, New York.

[7] Decrouez, G. and Hall, P. (2014). Split sample methods for constructing confidence intervals for binomial and Poisson parameters *J. Royal Statistical Society, Ser. B*., vol. 76, no. 5, pp. 949-975.

[8] DiCiccio, T. and Tibshirani, R. (1987). Bootstrap confidence intervals and bootstrap approximations. *Journal of the American Statistical Association*, 82(397), pp. 163âĂŞ170.

[9] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, vol. 7, pp. 1âĂŞ26.

[10] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[11] Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Ann. Statist.*, vol. 16, pp. 927-953.

[12] Hall, P. (1992a). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

[13] Hall, P. (1992b). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density, *Annals of Statistics*, vol. 20, pp. 675-694.

[14] Hall, P. (1992c). On the removal of skewness by transformation. *Journal of the Royal Statistical Society: Series B*, 54(1), pp. 221âĂŞ228.

[15] Johnson, N.J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, 73(363), pp. 536âĂŞ544.

[16] Kreiss, J.-P. and Paparoditis, E. (2023). *Bootstrap for Time Series: Theory and Methods.* Springer-Verlag, Heidelberg.

[17] Lahiri, S.N. (1997). Variance stabilizing transformations, studentization and the bootstrap, *J. Stat. Plann. Infer.* vol. 61, 105-123.

[18] Lahiri, S.N. (2003). *Resampling methods for dependent data,* Springer, New York.

[19] McElroy, T.S. and Politis, D.N. (2020). *Time Series: A First Course with Bootstrap Starter*, Chapman and Hall/CRC Press, Boca Raton.

[20] Politis, D.N. (2001). On nonparametric function estimation with infinite-order flat-top kernels, in *Probability and Statistical Models with applications*, Ch. Charalambides et al. (Eds.), Chapman and Hall/CRC, Boca Raton, pp. 469-483.

[21] Politis, D.N. (2003a). Adaptive bandwidth choice, *J. Nonparam. Statist*., vol. 15, no. 4-5, pp. 517-533.

[22] Politis, D.N. (2003b). The impact of bootstrap methods on time series analysis, *Statistical Science*, vol. 18, no. 2, pp. 219-230.

[23] Politis, D.N. and Romano, J.P. (1995). Bias-Corrected Nonparametric Spectral Estimation', *J. Time Ser. Anal.,* vol. 16, No. 1, pp. 67-104.

[24] Rosenblatt, M. (1991). *Stochastic Curve Estimation*, NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 3, Institute of Mathematical Statistics, Hayward, CA.

[25] Shao, X. and Wu, W. B. (2007). Asymptotic spectral theory for nonlinear time series, *Annals of Statistics*, vol. 35, pp. 1773-1801.

[26] Shorack, G.R. (2000). *Probability for Statisticians*, Springer, New York.

[27] Thulin, M. (2014). On split sample and randomized confidence intervals for binomial proportions, *Statist. Probab. Lett.*, Vol. 92, pp. 65-71.

[28] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.*, Vol. 22, pp. 209-212.

APPENDIX A: NONPARAMETRIC ESTIMATION OF HIGHER-ORDER DERIVATIVES

Flat-top kernel estimation of higher-order derivatives of a spectral or probability density function was addressed by Politis (2003). Because of certain issues in the way eq. (6) and (17) of Politis (2003) were presented, we provide the corrected formulas below. Note that R functions to compute the flat-top estimates of 2nd derivatives have been included in the R package `iosmooth()`.

For simplicity, Politis (2003) focused on employing a flat-top lag-window of trapezoidal shape. However, in what follows $\lambda(x)$ will denote *any* chosen flat-top lag-window, not necessarily trapezoidal. To elaborate, a general flat-top lag-window $\lambda(x)$ is a symmetric, bounded, and square-integrable function on $\mathbf{R}$, satisfying $\lambda(x) = 1$ for $|x| \leq$ some positive constant c. This general class of flat-top lag-windows was described in Politis (2011), where an Empirical Rule for data-based selection of the flat-top bandwidth is also given in order to estimate a function and/or its derivatives. Note that there is no discrepancy between the Empirical Rules given in Politis (2003) and Politis (2011); it is just that the former is focused on a trapezoidal lag-window while the latter applies for a general flat-top.

In what follows, we will assume that p is a nonnegative even integer, and the goal is estimation of the p-th derivative of a spectral or probability density function. The case where p is odd can be addressed in a similar way but one has to work with complex-valued random variables.

.1 Probability density function

We revert to the notation of Section 5 where $f(\cdot)$ is a probability density function. Let $R(s) = \int_{-\infty}^{\infty} e^{-isx} f(x)dx$ which is estimated by $\hat{R}(s) = n^{-1}\sum_{t=1}^{n} e^{-isX_t}$; these are related to the characteristic function and its empirical version respectively but using a different sign convention in order to parallel the results of Politis (2003).

The Fourier inversion formula gives an expression of $f(x)$ based on $R(s)$, i.e., $f(x) = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{isx} R(s)ds$; in this case, the p–th derivative of $f(x)$—when it exists—can be written as $f^{(p)}(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}(is)^p e^{isx}R(s)ds$. Since p is assumed to be even, this reduces to $\pm(2\pi)^{-1}\int_{-\infty}^{\infty}|s|^p e^{isx}R(s)ds$ depending on whether p is a multiple of 4 or not. Hence, we may define the quantity $f_p(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}|s|^p e^{isx}R(s)ds$ and estimate it by $\tilde{f}_p(x) = (2\pi)^{-1}\int_{-\infty}^{\infty}\lambda(s/M)|s|^p e^{isx}\hat{R}(s)ds$. Then, Theorem 3.1 of Politis (2003) applies *verbatim*, showing that $\tilde{f}_p(x)$ is consistent for $f_p(x)$ with a fast rate of convergence.

.2 Spectral density estimation

We now revert to the notation of Section 6 where $f(\cdot)$ is the spectral density function defined as $f(\omega) = (2\pi)^{-1}\sum_{k=-\infty}^{\infty}\gamma(k)\exp(ik\omega)$. The p–th derivative of $f(x)$—when it exists—can be written as $f^{(p)}(\omega) = (2\pi)^{-1}\sum_{k=-\infty}^{\infty}(ik)^p\gamma(k)\exp(ik\omega)$. Since p is even, this reduces to $\pm(2\pi)^{-1}\sum_{k=-\infty}^{\infty}|k|^p\gamma(k)\exp(ik\omega)$ depending on whether p is a multiple of 4 or not. Hence, we may define the quantity $f_p(\omega) = (2\pi)^{-1}\sum_{k=-\infty}^{\infty}|k|^p\gamma(k)\exp(ik\omega)$ and estimate it by $\tilde{f}_p(\omega) = (2\pi)^{-1}\sum_{k=-\infty}^{\infty}\lambda(k/M)|k|^p\hat{\gamma}(k)\exp(ik\omega)$. Theorem 2.1 of Politis (2003) now applies *verbatim*, showing that $\tilde{f}_p(\omega)$ is consistent for $f_p(\omega)$ with a fast rate of convergence.