Statistics Surveys
Vol. 0 (0000)
ISSN: 1935-7516
DOI: 10.1214/154957804100000000

A Review of Uncertainty Quantification for Density Estimation

Shaun McDonald[*]

Department of Statistics & Actuarial Science
Simon Fraser University
Room SC K10545
8888 University Drive
Burnaby, B.C.
Canada V5A 1S6
e-mail: shaun_mcdonald@sfu.ca

and

David Campbell[†]

School of Mathematics and Statistics
4302 Herzberg Laboratories
Carleton University
1125 Colonel By Drive
Ottawa, ON, K1S 5B6
e-mail: davecampbell@math.carleton.ca

Abstract: It is often useful to conduct inference for probability densities by constructing "plausible" sets in which the unknown density of given data may lie. Examples of such sets include pointwise intervals, simultaneous bands, or balls in a function space, and they may be frequentist or Bayesian in interpretation. For almost any density estimator, there are multiple approaches to inference available in the literature. Here we review such literature, providing a thorough overview of existing methods for density uncertainty quantification. The literature considered here comprises a spectrum from theoretical to practical ideas, and for some methods there is little commonality between these two extremes. After detailing some of the key concepts of nonparametric inference – the different types of "plausible" sets, and their interpretation and behaviour – we list the most prominent density estimators and the corresponding uncertainty quantification methods for each.

Keywords and phrases: Nonparametric inference, Density estimation.

Contents

[*]Supported by an NSERC Alexander Graham Bell Canada Graduate Scholarship.
[†]Supported by an NSERC Discovery Grant (RGPIN-2019-05115).

1. Introduction

Density estimation is one of the seminal examples of nonparametric statistical modelling. There are a litany of methods spread across decades of literature, from more "classical" approaches [159] to the most advanced modern techniques [23]. Estimation, however, is only one piece of the puzzle: as in any statistical

problem, it is desirable to also conduct *inference*, providing some quantification of uncertainty in addition to single estimates. Broadly speaking, uncertainty is quantified using sets of "plausible" values - for example, confidence intervals for frequentist methods and credible intervals for Bayesian ones. Although not as abundant as other areas in nonparametric statistics, there is a sizeable body of literature on uncertainty quantification (UQ) for density estimation, ranging from rigorously theoretical to extremely practical.

The following sections provide more detail on various types of "uncertainty sets", then outline several density estimation methods and review available literature dealing with UQ for each one. Although some combinations of estimation and inference ideas are not represented in the literature (in particular, a substantial gap exists between theoretical and practical UQ developments in many cases), in principle, one could *always* obtain some kind of uncertainty bounds on a density estimate, either by bootstrapping a frequentist method or taking quantiles of MCMC output for a Bayesian one. Whether or not such bounds have suitable coverage properties or otherwise perform adequately is another question for which the answers are not always known. Despite some of these limitations, this paper presents a comprehensive review of the work done thus far in unknown density UQ, and suggests promising areas to extend the research or "fill in the gaps".

2. Overview and notation

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a set of samples from some unknown "true density" f_0. The majority of discussion here will assume i.i.d. samples, but other data structures will also be considered as warranted. One structure that is common enough to justify mentioning here is the case of "noisy" observations $Y_i = X_i + Z_i$, where the errors Z_i have known distribution and the true X-values are unknown. Estimating the density of $\boldsymbol{X}$ in this case is called *deconvolution density estimation*. In the present context, $\hat{f}$ will denote a *specific* "point" estimate (in the sense that it is a *single element of a function space*) of f_0, such as a MLE or posterior mean; while f will typically be used to discuss classes or function spaces of estimators in more generality.

As mentioned in the introduction, UQ arises by considering "uncertainty sets". Such sets are random through their dependence on $\boldsymbol{X}$, but for brevity the notation here does not reflect this. As f_0 is a function, there are several ways to define uncertainty sets, each with different implications and advantages. Perhaps the most obvious examples are *pointwise intervals* $C_x = [L(x), U(x)]$, defined separately for each point x in the domain of f_0. A common special case is when the intervals are symmetric about an estimator $\hat{f}$: $C_x = \left[\hat{f}(x) - \epsilon_x, \hat{f}(x) + \epsilon_x\right]$. The goal with pointwise intervals is to achieve (possibly only approximately or asymptotically) $\mathbb{P}\left(f(x) \in C_x\right) \geq 1 - \alpha$ for all $x \in \mathrm{Dom}\left(f_0\right)$, where $1 - \alpha$ is the usual predetermined level. The meaning of the generic placeholders $\mathbb{P}$ and f depends on whether the inference is frequentist or Bayesian.

Pointwise intervals tend to be easy to implement, and also have nice theoretical properties for some (but not all) density estimation techniques. However, they are fundamentally limited in their ability to make "global" uncertainty statements. For a given level $1-\alpha$, even if $\mathbb{P}\left(f(x) \in C_x\right) \geq 1-\alpha \ \forall x$, the stronger and perhaps more meaningful statement $\mathbb{P}\left(f(x) \in C_x \ \forall x\right) \geq 1-\alpha$ cannot necessarily be deduced. However, in some cases the "simultaneous" statement *does* hold, in which case the set $C = \{(x,y) : x \in \mathrm{Dom}\left(f_0\right), y \in C_x\}$ is called a *confidence* or *credible band*. Like pointwise intervals, bands are often centered at a specific estimator, although this need not be the case. For instance, Hall and Titterington [89] proposed to construct frequentist bands for univariate densities based on simultaneous multinomial confidence intervals for the probability masses within consecutive subintervals of the domain. Classical approximation results allowed them to construct such intervals without a specific density estimator, and they constructed the density bands by interpolation, with modifications depending on f_0 being once or twice differentiable. Such bands are not smooth, but were shown to have suitable coverage properties and optimal asymptotic widths without making further assumptions or restrictions. Hengartner and Stark [94] devised conservative confidence bands for shape-restricted densities (either monotonic or having $\leq k$ modes for known k, possibly relative to some weight function), also obtained without an estimator. To derive their bands, they started with a confidence region for the cdf F_0 comprised of distributions with densities having the same shape restriction as f_0, and subsequently showed how to reduce the determination of the band for f_0 to a finite-dimensional linear program while conservatively preserving coverage probability.

If C_x has the same width for all x in a band, then it is uniform and is therefore a L^∞-ball in a suitable function space $\mathcal{F}$. Thus, a uniform band is a special case of a more general idea: using a ball C in some pseudo-metric space of functions $(\mathcal{F}, d)$ as an uncertainty set. Analogously to pointwise intervals and bands, here the goal is to have $\mathbb{P}\left(f \in C\right) \geq 1-\alpha$. For choices of d such as the Hellinger or L^2 distances, such sets arise in nonparametric literature due to their satisfying theoretical properties. However, their practicality is somewhat limited: an L^2-ball of functions, for instance, does not provide error bounds that can be easily visualized or understood, short of simply plotting a large number of functions from the ball alongside $\hat{f}$. For example, Szabó, van der Vaart and van Zanten [198] visualized L^2-balls from an empirical Bayesian model for nonparametric regression by sampling functions from the posterior and plotting the $100\left(1-\alpha\right)\%$ of draws closest in the L^2 sense to the posterior mean. In a discussion of this paper, Low and Ma [133] suggested using this procedure to generate bands for the regression function whose boundaries are simply the pointwise maxima and minima of these closest posterior draws. Their simulations showed that the bands thus obtained performed quite well with respect to the framework of Cai, Low and Ma [20]. Beyond the aforementioned examples and those in Section 4.4.3, discussion of these "uncertainty balls" is limited, although Chapter 6 of Csörgo and Révész [35] contains theorems on the asymptotic distributions of the L^2-errors of several "classical" frequentist estimators (KDE's, histograms, and

certain orthonormal basis expansions). These results *could* be relevant towards the construction of confidence balls, but this seems not to have been done in practice, likely due to their limited visual utility. On the other hand, if d is the L^∞ distance, one recovers the meaningful and easily-visualized UQ given by bands, at the expense of nice theory in some cases. As before, for *any* pseudometric a common special case arises by taking the associated sets to be centered at some estimator:

$$C(\epsilon) = \left\{ f \in \mathcal{F} : d\left(f, \hat{f}\right) < \epsilon \right\}. \tag{1}$$

In frequentist inference, uncertainty quantification relies on *confidence sets* of any of the forms described above, typically obtained in practice using asymptotic arguments and/or bootstrapping. Confidence sets are designed in view of the "ground truth" $\boldsymbol{X} \sim f_0$: letting $\mathbb{P}_0$ denote the probability law associated with f_0, the goal is to achieve *coverage probability* $\mathbb{P}_0\left(C_x \ni f_0(x)\right) \geq 1-\alpha \ \forall x$ in the pointwise case, or $\mathbb{P}_0\left(C \ni f_0\right) \geq 1-\alpha$ for bands or function balls. The Bayesian approach employs *credible* sets instead: using $\Pi\left(\cdot \mid \boldsymbol{X}\right)$ as generic notation for the posterior over a space of densities f, the sets of interest are either pointwise intervals such that $\Pi\left(f(x) \in C_x \mid \boldsymbol{X}\right) \geq 1-\alpha \ \forall x$, or bands/balls such that $\Pi\left(f \in C \mid \boldsymbol{X}\right) \geq 1-\alpha$. To facilitate validation and comparison, it is possible to view Bayesian methods through a frequentist lens by acknowledging the existence of the "ground truth" f_0, in which case the posterior $\Pi\left(\cdot \mid \boldsymbol{X}\right)$ is considered as a random measure due to its dependence on $\boldsymbol{X} \sim \mathbb{P}_0$. This leads to a similar interpretation of credible sets as functions of the data. It is then natural to ask if they achieve coverage in the aforementioned frequentist sense. Put another way, can credible sets also serve as valid confidence sets? The difficulty of answering this question for nonparametric Bayesian methods is well-known and an active area of research; discussion of coverage therefore tends to be easier in the frequentist paradigm.

Naturally, the best possible inference produces small sets with high coverage probability. To this end, the concepts of *honesty* and *adaptivity* are relevant. Consider a confidence set C_n, where the subscript n is added to emphasize limiting behaviour with respect to sample size. The remainder of this section ignores the distinction between pointwise intervals, bands, and balls.

In the context of density estimation, C_n is *honest* at level $1-\alpha$ if

$$\liminf_n \inf_{f_0 \in \mathcal{F}} \mathbb{P}_0\left(C_n \ni f_0\right) \geq 1-\alpha, \tag{2}$$

where $\mathcal{F}$ is once again a suitable function space of interest [95]. In words, an honest confidence set asymptotically achieves the desired coverage level *uniformly* over all possible "ground truths". Honesty is crucial for practical finite-sample inference: without it, it is possible in some cases for the infimum of coverage probability over $\mathcal{F}$ to be zero for *any* n [123].

The precise definitions and presentations underpinning the notion of *adaptivity* vary throughout the nonparametric literature [19, 69, 95, for instance]. The present discussion will focus as narrowly as possible on material relevant to

density UQ. Suppose $\mathcal{F} = \cup_{s\in\mathcal{S}}\mathcal{F}_s$ for some ordered index set $\mathcal{S}$, where for $s > t$ it holds that $\mathcal{F}_s \subseteq \mathcal{F}_t$ and the elements of $\mathcal{F}_s$ are smoother than those of $\mathcal{F}_t \setminus \mathcal{F}_s$. Typically each subset $\mathcal{F}_s$ is, say, a ball in a suitable Besov space of regularity s, with an associated minimax-optimal contraction rate $r_n(s)$ decreasing in both n and s [70]. Following Hoffmann and Nickl [95], call C_n *adaptive* if there exists $L > 0$ such that, for all $s \in \mathcal{S}$ and for all n large enough,

$$\sup_{f_0\in\mathcal{F}_s} \mathbb{E}_0 |C_n| \leq L r_n(s), \tag{3}$$

where the expectation is with respect to $\mathbb{P}_0$, and $|C_n|$ is the diameter of C_n with respect to the metric by which it is defined (typically L^2 or L^∞ in this context). Naturally, less uncertainty is expected in the estimation of smoother functions. Adaptive confidence sets take advantage of this fact: they are optimal in the sense that, for every level of smoothness under consideration, their "maximum" expected size contracts at the optimal rate. This is especially useful since the actual smoothness of the true density is likely to be unknown, and it does not have to be specified for adaptive C_n. Unfortunately, adaptivity is an elusive goal which cannot be achieved without caveats, especially if honesty is also desired. As it pertains to density estimation, one of the earliest results to this effect came from Low [132], who considered pointwise inference for f_0 with uniformly bounded k^{th} derivatives. They showed that, over this space, an honest confidence interval could achieve the worst-case contraction rate for *any* f_0, regardless of its true smoothness. Confidence sets in L^∞ are particularly tricky: full adaptivity over finitely many smoothness levels can only be achieved by swapping the lim inf and inf in (2) (i.e. considering "dishonest" bands) [70, 95], but Bull [15] showed that even with this modification it is still impossible to adapt over a continuous range of smoothness levels in the white noise model. Dümbgen [41] defined density confidence bands using a test statistic depending on the cdf values at order statistics and showed some adaptivity results based on local smoothness, but they are only valid over sets of shape-restricted (e.g. unimodal or monotonic) densities. Such difficulties are pervasive for all types of confidence sets: to achieve honesty and adaptivity together, it is necessary to assume additional restrictions on the smoothness classes under consideration or the functions therein. The theory shows that L^2 confidence sets are less restrictive in this regard than confidence bands, but neither are without their difficulties. Section 8.3 of Giné and Nickl's textbook [70] is an excellent and comprehensive discussion of these ideas, and the references in their notes provide further details. The authors explored adaptation theory for the white noise model, but noted that it can be made to apply to density estimation.

Adaptivity and honesty are central to the theory of nonparametric inference, but to many practitioners they may ultimately be less important than the aforementioned visual aspect of UQ. Figure 1 shows how to graphically represent the uncertainty associated with a density estimate by plotting multiple estimators and corresponding UQ methods, all based on the same simulated dataset. The figure includes both frequentist and Bayesian inference methods, and demonstrates the differences between pointwise intervals and simultaneous bands (in

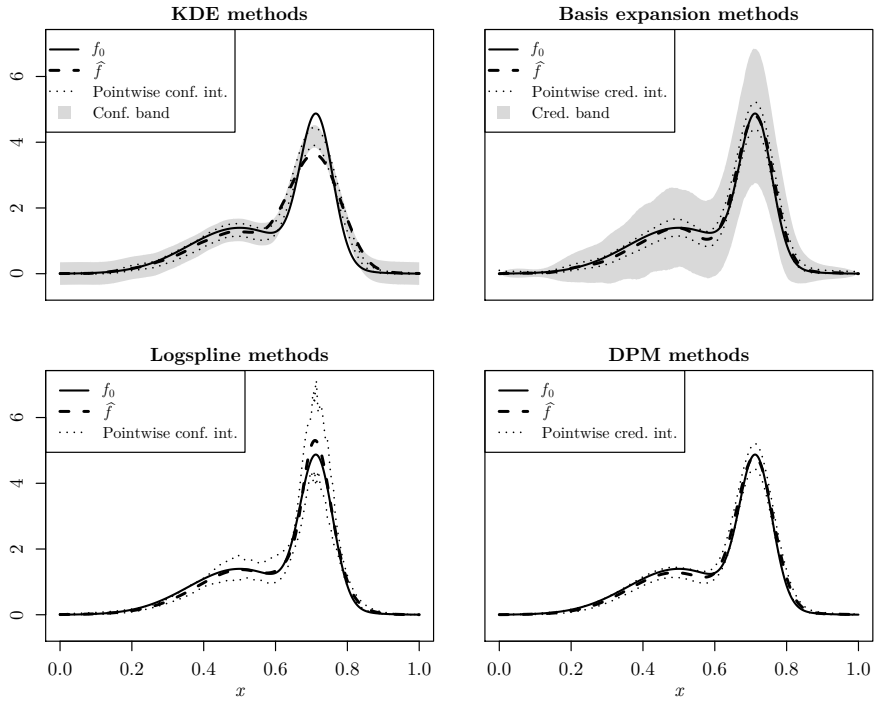

FIG 1. *Different combinations of density estimation and UQ methods applied to the same sample.*

particular, the latter are wider than the former, as one would expect to be necessary for this stronger type of inference). The methods shown in Figure 1 are among the many described in the following sections, each of which explores UQ in terms of the concepts described above. The figure itself is discussed in more detail in Section 9 as well as the supplementary material [141].

3. Kernel density estimators

KDE's are one of the most used and well-studied density estimation methods, at least in the frequentist literature. They are ubiquitous enough that their properties are arguably "common knowledge", receiving extensive documentation in textbooks, undergraduate course material, and review papers unto themselves [e.g. 28, whose review informs much of the discussion in this section]. Recall that a kernel density estimate for a density on $\mathbb{R}^d$ is of the form

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{4}$$

where K is some (typically symmetric) kernel function and h is a bandwidth which controls smoothing, or bias/variance tradeoff. Asymptotic theory for estimation is typically based on h decaying to zero in some "big-O" relationship with n that optimizes MSE, or integrated MSE. Practical methods for obtaining h include cross-validation, plug-in methods, rules of thumb, and bootstrapping [102]. Note that, as the estimator is little more than a sample mean, it is equivalent to a conditional expectation with respect to the random measure F_n, the empirical distribution function of $\boldsymbol{X}$.

3.1. Pointwise inference

For pointwise inference, it is well-known that KDE's are asymptotically normal: with $\hat{f}$ as defined in (4), for all $x \in \mathbb{R}^d$ it holds that

$$\sqrt{\frac{nh^d}{f_0(x) \int K^2(t)\mathrm{d}t}} \left(\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right) \xrightarrow{d} \mathcal{N}(0, 1). \tag{5}$$

Furthermore, the distributions for a finite collection of points are asymptotically independent [18]. Using this fact, it follows that the endpoints for pointwise confidence intervals should be roughly of the form $\hat{f}(x) \pm z_{1-\alpha/2}\sigma_x$, where σ_x^2 is the variance which is asymptotically equal to $\frac{f_0(x) \int K^2(t)\mathrm{d}t}{nh^d}$. In practice, intervals can be computed by estimating σ_x: either by using one of its asymptotically-equivalent formulae ("plugging in" $\hat{f}$ in place of f_0 [28, 87] or replacing expectations with sample averages [96, 57]) or bootstrapping [28]. Many papers also replace the standard normal quantiles with those of bootstrap t-statistics [e.g. 96, 84]. Such studentized or "percentile-t" confidence intervals seem to be the

most commonly-discussed in the literature, but any method of bootstrap confidence interval construction should be valid - for instance, Chen [28] discussed a bootstrap interval based entirely on the percentiles of absolute deviations. Hall and Kang [87] showed that re-calculating the bandwidth for each bootstrap sample does not provide worthwhile improvements to the accuracy of inference[1], so computational difficulty is avoided by using the same bandwidth across all replications. In the univariate case with a compactly-supported kernel, Chen [27] considered the construction of confidence intervals based on *empirical likelihood*, a nonparametric analogue to the standard methods of profile log-likelihood ratios. The theory is similar to the parametric case: viewing $\hat{f}(x)$ as a sample mean of random variables $K\left(\frac{x-X_i}{h}\right)/h$, Chen derived a limiting chi-squared distribution for $\ell\left(\mathbb{E}\left[\hat{f}(x)\right]\right)$, where ℓ is the profile empirical log-likelihood ratio. This allows for pointwise intervals of the form $\{y : \ell(y) \leq c_{1-\alpha}\}$, where $c_{1-\alpha}$ is the $1-\alpha$ quantile of the χ^2_1 distribution. Chen showed that such intervals have asymptotic performance comparable to the percentile-t bootstrap and can outperform it in simulations, especially with Bartlett correction.

Note that everything discussed thus far is based on (5), which is centered about $\mathbb{E}\left[\hat{f}\right]$ instead of f_0. This poses a problem for inference when choosing an "optimal" bandwidth minimizing (integrated) MSE or some proxy. The aforementioned intervals provide asymptotically-correct coverage for the expectation of $\hat{f}(x)$; in order for this to hold for $f_0(x)$ instead, the quantities in the numerator of (5) must be interchanged. This is only possible if the ratio of bias and asymptotic standard deviation

$$\sqrt{\frac{nh^d}{f_0(x)\int K^2(t)\mathrm{d}t}}\left(f_0(x) - \mathbb{E}\left[\hat{f}(x)\right]\right)$$

goes to zero. However, the optimal asymptotic error rate is achieved when h is set proportional to some power of n such that the squared bias and variance decay at equal rates [28, 159]. Thus, with an "optimal" bandwidth, the ratio above tends to a nonzero constant so that confidence intervals do *not* have the correct coverage properties[2]. There are two main ways to handle this. The first is *undersmoothing*, where a lower-than-optimal bandwidth is selected. In the univariate case, Horowitz [96] suggested taking h proportional to a higher power of n than usual, thereby allowing the squared bias to decay faster than the variance; while Hall [84] multiplied a rule-of-thumb bandwidth by a constant $c \in (0, 1)$. Chen [27] used a version of the former when a confidence interval at only one point is desired: first obtaining kernel estimates f_0 and its second derivative at the point with approximations of the (local) MSE-optimal bandwidths, then using

[1] In simulations, they found that recalculating the bandwidths can provide higher coverage, but at the expense of more conservative intervals. They showed that it doesn't asymptotically make a difference for compactly-supported kernels, but used the Gaussian kernel in simulations since its tails are light enough that it is "almost compact".

[2] This is one example of inference being at odds with the goal of optimal estimation. This will become a familiar refrain in theoretical ideas discussed throughout this paper.

these to estimate the bias. Chen suggested simply using the optimal bandwidth in confidence interval construction when the estimated relative bias is small, or an estimate of a coverage-optimal undersmoothing bandwidth when it is large. Because a smaller h means higher variance, confidence intervals based on undersmoothing may be wider than one would prefer [28]. The second method is therefore to estimate the bias term with $\hat{b}$ and replace $\hat{f}$ with the bias-corrected estimator $\hat{f} - \hat{b}$. Assuming a kernel of order[3] r is used, the bias depends on the r^{th}-order derivatives of f_0, assuming these are bounded and continuous [26]. These derivatives can also be estimated with kernel methods, but require higher bandwidths for optimality than the density estimator itself; for this reason, traditional bias correction uses an *oversmoothed* KDE to obtain $\hat{b}$ [84, 28]. Hall [84] showed through both asymptotics and simulations that undersmoothing with a higher-order kernel results in percentile-t bootstrap confidence intervals with smaller coverage errors than those based on such "oversmoothing" bias corrections. However, Calonico, Cattaneo and Farrell [21] developed a "robust" bias correction, in which the variance estimate used in confidence interval construction is modified to account for the correction, and showed that it can perform as well as undersmoothing-based intervals, with more robustness to bandwidth selection. Notably, their results hold when using the MSE-optimal bandwidth and second-order kernels for both $\hat{f}$ and the bias correction, which they noted to be convenient automatic choices. While lower error rates and narrower intervals are desirable, it should be noted that the bias-corrected centers $\hat{f} - \hat{b}$ are not necessarily nonnegative. Also note that the aforementioned results are based on a kernel with compact support.

Hall and Horowitz [86] devised another novel bootstrap approach. Starting with the original KDE $\hat{f}$, they repeatedly drew "bootstrap" samples *from* $\hat{f}$ (some papers call this the *smoothed bootstrap*, e.g. [148]) and used these to create Gaussian plug-in intervals at each point x in the domain, with some nominal confidence level $1 - \beta(x)$. They set each $\beta(x)$ to ensure that the actual coverage (as estimated with bootstrap replicates) achieved the desired level $1 - \alpha$. Letting $\hat{\beta}_\delta$ be the ξ-quantile, for some low ξ, of the $\beta(x)$-values over a fine grid of x's with edge width δ, they took $\hat{\beta}$ as the limit of $\hat{\beta}_\delta$ as $\delta \to 0$ and finally used standard normal quantiles $z_{1-\hat{\beta}/2}$ to construct pointwise plug-in intervals centered at $\hat{f}$. Their theory and simulation studies focused on nonparametric regression, and in this case they showed asymptotic pointwise coverage of at least $1 - \alpha$ at roughly $(1 - \xi)100\%$ of points in the domain. However, they suggested that all results would translate to KDE's as well.

Similar ideas to those discussed above extend to situations besides a single i.i.d. sample $\boldsymbol{X}$. Louani [130] derived theoretical results for the case of randomly right-censored data: when there exists another sample $\boldsymbol{Y}$ of size n, and only $Z_i = \min\{X_i, Y_i\}$ and $\mathbb{1}(X_i \leq Y_i)$ are observed. They considered a modified KDE defined by integrating with respect to a Kaplan-Meier estimate of the cdf, rather than the usual edf F_n. Relaxing some of the conditions required

[3] The *order* r of a kernel K is the smallest positive integer such that the r^{th} moment of K is nonzero.

for previous similar results [142, 127] (in particular, assuming only one bounded continuous derivative of f_0), Louani showed pointwise asymptotic normality for this estimator when using a kernel of compact support. The asymptotic standard deviation is similar to the left-side factor in (5), but with an extra factor of $\sqrt{1-G(x)}$, where G is the cdf of $\boldsymbol{Y}$. Another theoretical extension came from Giné and Mason [68], who considered kernel-based U-statistic estimators for the densities of functions $g(X_1, \ldots, X_m)$ with $m > 1$. Analogously to other results described here, they derived central limit theorems for such estimators, noting the bias can be eliminated if the bandwidth decays appropriately in the special case where g is additive in its arguments (see also [185] for related results). Schick and Wefelmeyer [186] studied the case when the data is a linear process: $X_i = \sum_{s=0}^{\infty} a_s \epsilon_{i-s}$ for zero-mean ϵ_s and absolutely convergent $\{a_s\}$. The asymptotic mean in their limiting normal distribution is the convolution of the true density with the kernel. For a more practically-oriented extension, Wang and Wertelecki [214] considered data observed with rounding errors. They proposed a multi-step process to estimate the density of $\boldsymbol{X}$: first deriving a rough, convolution-based estimate for the cdf of the non-rounded data; then using this to generate a sample from the estimated distribution of the rounding errors; and finally subtracting the simulated errors from the rounded data and constructing a KDE from the resulting quantities. Because the procedure involves simulated sampling, it naturally lends itself to bootstrap-style uncertainty quantification, which the authors showed in the form of pointwise confidence intervals for real data.

Further results exist for deconvolution density estimation. In this case, it is common to use a specialized *kernel deconvolution density estimator*, replacing the "standard" kernel in (4) with a *deconvolution kernel*: the Fourier transform of the ratio between the characteristic functions of some kernel function and the known error distribution. Fan [49] provided asymptotic normality results for such estimators in two cases: *ordinary smooth* deconvolution (where the tails of the error characteristic function decay at a polynomial rate) and *supersmooth* deconvolution (where they decay exponentially). In addition to the usual corollary of bias removal with undersmoothing, Fan also showed that the asymptotic variance (which depends on the true unknown density of the noisy data in the ordinary smooth case, and does not have a general expression in the supersmooth case) in the pivotal quantity could be replaced by a sample-dependent term: either the sum of squared deconvolution kernel values, or their sample variance (only the former was considered for the supersmooth case). Zhang [217] showed similar results for a similar estimator. Fan and Liu [50] later relaxed the conditions assumed in [49] for the ordinary smooth case, allowing the asymptotic normality results to apply to a wider variety of commonly-used error distributions. van Es and Uh [205] showed for a subset of supersmooth error densities that, under certain conditions on the kernel, the asymptotic variance of the estimator does not depend on the data or the true density. They noted that this allows in this case for the construction of pointwise confidence intervals without data-dependent standardization, although they do not address the issue of bias. Further asymptotic normality results with known variances are given in van Es

and Uh [204] and Uh [201] for somewhat more general kernels and subsets of supersmooth error densities. Masry [139] generalized the classical results of Fan and Zhang to inference on the joint density of stationary process data based on observations with i.i.d. additive noise. They showed asymptotic normality for various types of mixing with both ordinary smooth and supersmooth error distributions, but only considered undersmoothing-based bias removal and sample-based standardization for the former. For both i.i.d. and strongly-mixing data, Zu [219] proved asymptotic normality of the estimator when the noise is logarithmic chi-squared, a case not covered by the assumptions in the previous literature. The asymptotic variance in this case depends once again on the true density of the noisy data; Zu suggested that it could be consistently estimated by a classical KDE to facilitate construction of (biased) confidence intervals.

Returning once more to the case of an i.i.d. sample, a final extension is the *adaptive kernel density estimator* implemented in Stata by Van Kerm [206]. This method starts with a "pilot" density estimate of fixed bandwidth; its values at the sample points are used to assign individual bandwidths to each of the kernels in (4), which can also be given individual weights. These variable bandwidths reduce variance in regions where data is sparse, and bias in regions where it is dense. As with the normal KDE, it is an easy matter to get a plug-in estimate of standard error; this is how Van Kerm's software implements simple pointwise inference.

3.2. Simultaneous inference

Moving beyond the pointwise case, consider simultaneous UQ on the entire support or some subset thereof. The aforementioned undersmoothing and/or bias-correction principles still apply, and the rest of this section will largely take the application of such principles for granted. Bickel and Rosenblatt [9] provided perhaps the first results to this effect for univariate KDE's, showing that, under suitable technical conditions,

$$\mathbb{P}\left[A_n\left(\sqrt{\frac{nh}{\int K^2(t)\mathrm{d}t}}\sup_x\left|\frac{\hat{f}(x)-\mathbb{E}\hat{f}(x)}{\sqrt{f_0(x)}}\right|-d_n\right)<z\right]\to e^{-2e^{-z}} \tag{6}$$

for suitable sequences A_n and d_n (the latter being a function of the former), with the supremum taken over a compact interval (say, $[0,1]$ w.l.o.g.) on which f_0 is bounded away from 0. They further showed that with moderate undersmoothing, it is possible to replace $\mathbb{E}\hat{f}$ and $\sqrt{f_0}$ in (6) by f_0 and $\sqrt{\hat{f}}$, respectively, thereby justifying variable-width confidence bands $\hat{f}(x)\pm\sqrt{\frac{\hat{f}(x)\int K^2(t)\mathrm{d}t}{nh}}\left(\frac{z}{A_n}+d_n\right)$ for $x\in[0,1]$, where z is such that $e^{-2e^{-z}}=1-\alpha$. Note that using a differently-scaled A_n, the factor of 2 in the exponent of the limiting distribution can be eliminated, thereby turning it into the c.d.f. of a *standard* Gumbel random variable [e.g. 69, who derived such a result for an undersmoothed data-driven bandwidth choice and plug-in estimator for $\sqrt{f_0}$; see Section 4.4 for further details]. In

either case, the limiting probability law is of the extreme value or "double exponential" form. Rosenblatt [177] expanded upon Bickel and Rosenblatt's results, slightly relaxing the conditions under which (6) holds in the univariate case and generalizing to the multivariate case. However, their multivariate results required rather strong restrictions on the bandwidth, differentiability of f_0, and moments of K. Rio [171] gave another rather technical result on the limiting distributions of suprema over closed subsets of $(0,1)^d$ for d-dimensional densities. Additional generalizations of the Bickel-Rosenblatt results in the univariate case were provided by Giné, Koltchinskii and Sakhanenko [66], who gave conditions for results similar to (6) to hold with a different weight function Ψ replacing the factor $\left(\sqrt{f_0}\right)^{-1}$ or the supremum taken over a data-dependent set. The same authors provided further theory to this end in a companion paper, in which they considered suprema over the whole real line [67]. Sakhanenko [183] further modified and extended these results to multivariate densities. Using moderate deviations principles, Mokkadem and Pelletier [143] showed that it is actually possible to construct Bickel/Rosenblatt-style confidence bands with asymptotic coverage level equal to 1 by using separate bandwidths for the KDE's in the mean and variance estimates (i.e. in the quantities used to define, respectively, the centre and margins of the bands). Further technical refinements allowed them to achieve this with narrower bands, at the expense of a slower convergence rate. While remarkable, these results have not been applied in practice in literature to date.

A drawback of using these asymptotics in practice is that convergence to the extreme value distribution is known to be very slow [e.g. 83]. Thus, it may be advisable to use bootstrapping for confidence bands. In what follows, let f^* denote a KDE based on a bootstrap resample of $\boldsymbol{X}$. Hall [85] considered bands (over compact intervals) of the type

$$\left\{(x,y) : 0 \leq x \leq 1, \hat{L} \leq \frac{\hat{f}(x) - y}{\sqrt{y}} \leq \hat{U}\right\}$$

$$\text{where } \mathbb{P}\left(\hat{L} \leq \inf_x \frac{f^*(x) - \hat{f}(x)}{\sqrt{\hat{f}(x)}} \leq \sup_x \frac{f^*(x) - \hat{f}(x)}{\sqrt{\hat{f}(x)}} \leq \hat{U} \mid \boldsymbol{X}\right) = 1 - \alpha.$$

This band is based on the "studentized" quantity $\left(\hat{f} - \mathbb{E}\hat{f}\right) / \sqrt{\mathbb{E}\hat{f}}$, but differs from others by not using any kind of estimator for the denominator. Hall found in simulations that this interval had better coverage for $\mathbb{E}\hat{f}$ than a bias-corrected translation had for f_0, presumably due to inaccuracy in the bias correction. Hall and Owen [88] recommended the bootstrap to construct simultaneous confidence bands on $[0,1]$ with profile empirical likelihood methods. Recalling the notation for empirical likelihood in Section 3.1, they found an extreme value limiting result for $\sup_x \sqrt{\ell\left(\mathbb{E}\left[\hat{f}(x)\right]\right)}$ similar to (6), but recommended using percentile bootstrap methods to find a suitable bound $\hat{c}$ for a band of the form $\{f : \ell(f(x)) \leq \hat{c} \ \forall x \in [0,1]\}$. The technicalities and variations they considered

are too cumbersome to discuss further here; see [88] for the full details. They found their intervals to be disappointingly wide when applied to real data, but suspected that this was due to the inherent variability of the density estimation itself. Neumann [148] gave quite general theoretical results for uniform-width percentile bootstrap bands of the form $\hat{f} \pm t^*_\alpha$, where t^*_α is the bootstrap quantile of $\sup_x |f^* - \mathbb{E}f^*|$. Their results are valid for multivariate densities, suprema over all of $\mathbb{R}^d$, and weakly-dependent data. Neumann used compactly-supported kernels and the smoothed bootstrap: generating $\boldsymbol{X}^*$ from a possibly-different KDE based on the original sample, rather than from the empirical distribution. In a recent paper, Cheng and Chen [29] used the debiased estimator of Calonico, Cattaneo and Farrell [21] to derive asymptotically correct bands, via the bootstrap, of either uniform width (using quantiles of $\sup_x \left|f^* - \hat{f}\right|$) or variable width (using quantiles of $\sup_x \left|\left(f^* - \hat{f}\right)/\sigma^*\right|$ multiplied by $\hat{\sigma}(\cdot)$), where the bootstrap density and associated variance estimates were all computed based on the bias-correction approach. Their results extend to the multivariate case and assume a compactly-supported f_0. Their simulation study showed that their bands achieved better coverage and narrower width than those based on the standard KDE, although some undercoverage still occurred for small samples without undersmoothing.

Yeh [215] used the bootstrap to create a rather novel type of confidence band. Generating a large number of KDE's from bootstrap samples of $\boldsymbol{X}$, they retained the $100\,(1-\alpha)\,\%$ of them with the largest *curve depth* (a way of ranking functions "from the centre out" based on some distance from a central curve, in this case the KDE $\hat{f}$). Simulation studies showed that such bands had reasonable performance compared to the asymptotic methods discussed in this section.

As is the case for pointwise inference, for simultaneous bands there are analogous results for deconvolution KDE's, described by Bissantz et al. [10]. A necessary assumption for these results is that the characteristic function of the error density decays as $t^{-\beta}$ for large $|t|$ and some known constant $\beta > 0$. Their asymptotic results for bands over a compact interval are nearly equivalent to those derived from (6), although in the asymptotic standard deviation they divided by an extra factor of h^β and replaced $\hat{f}$ with $\hat{g}$, where the latter is an estimate of the density g of the observed Y-values (a standard KDE suffices). However, they noted slightly better coverage probability (especially in terms of robustness to model misspecification) can be achieved with percentile bootstrap confidence bands of variable width based on the quantiles of $\left[f^*(x) - \hat{f}(x)\right]/\hat{g}(x)$, where f^* is a deconvolution estimator from a bootstrap sample of the observed noisy data.

3.3. Miscellaneous

Aside from some technical considerations in the previous section, not much consideration is given to the support of the true density. Indeed, the issue of

KDE boundary bias for f_0 of restricted support is well-known and several mitigating strategies exist [e.g. 101, and discussion therein], but this is rarely discussed in the context of uncertainty quantification. One exception is given by Bouezmarni and Rombouts [11], who considered the *gamma kernel estimator* for time series data on $[0, \infty)$. The gamma kernel has a shape parameter varying with x and leads to an estimator (asymptotically) free of boundary bias. The authors showed pointwise asymptotic normality analogously to the results discussed above, based on the behaviour of the gamma scale parameter which acts as a bandwidth. In practice, it can be selected by cross-validation; the authors did so and constructed confidence intervals for real data based on their asymptotic results.

This section concludes by discussing a paper on large-sample Bayesian methods by Lo [126]. The key observation for this discussion is to recall that one can view the KDE (4) as a (conditional) expectation with respect to F_n. Lo's ideas are based on replacing F_n in this expectation with a different random distribution F conditional on $\boldsymbol{X}$. One such example is the empirical distribution of a bootstrap sample; this is equivalent to a probability measure with atoms at the sample values and weights randomly selected from $\{1/n, 2/n, \ldots, 1\}$. Lo also considered the *Bayesian bootstrap*, where the weights on the atoms are drawn from a uniform Dirichlet distribution [180]. This is equivalent to a draw from the posterior when $\boldsymbol{X} \sim F$, and F is given an improper Dirichlet Process prior with zero base measure. Finally, Lo generalized this to allow for a non-zero base measure in the DP prior. They showed an asymptotic result analogous to (6) for all three aforementioned KDE variants, where the limit holds for f_0-almost all $\boldsymbol{X}$. This allowed them to use extreme value asymptotics to derive appropriate *Bayesian* bands for f centred at the usual KDE $\hat{f}$. In practice one may prefer not to do this, given the substantial developments in Bayesian computation since the time of Lo's paper.

4. Adaptive basis expansion methods

This section considers estimates for f_0 of the form

$$f(x) = \sum_{j=1}^{K} b_j B_{j,K}(x), \tag{7}$$

where the $B_{j,K}$'s are a suitable set of fixed nonnegative "basis functions" for a given K. The simplest choice is taking them to be indicator functions on disjoint subsets of the support, in which case f is simply a histogram. Other options include Bernstein polynomials [e.g. 208, 160], B-splines [e.g. 190], and wavelets [e.g. 69]. The coefficient vector $\boldsymbol{b} \in \mathbb{R}^K$ is constrained such that f is a valid density. The remainder of this section will use $\hat{\boldsymbol{b}}$ to denote the coefficients associated with a specific estimator $\hat{f}$.

The dimensionality K is of particular interest, serving as a smoothing parameter that controls the bias-variance tradeoff of the estimator. The basis functions

corresponding to higher K-values are typically "narrower", allowing for more intricate shape detail to be captured in estimates. For instance, taking a high K-value for the histogram corresponds to using a larger number of narrower bins. Conversely, a value that is too high will result in a high-variance estimator that is unacceptably noisy. In general, higher K-values are required for larger samples to capture the true density.

One can choose K in a data-driven way. Many theoretical results for this approach rely on K increasing with n, usually appealing to some "big-O" conditions on its growth [e.g. 3, 69, 200]. In practice, a value could be chosen by cross-validation [120], changepoint methods [81], or appealing to known asymptotic theory [3, who derived nice properties for a method with $K = o(n/\log n)$ and then simply used $K = n/\log n$ in a simulation study]. In the theoretical Bayesian context, Rousseau and Szabó [179] considered *maximum marginal likelihood* (MML) estimates for K, marginalizing over a prior for $\boldsymbol{b}$. Further discussion of such ideas is beyond the scope of this paper; see van de Wiel, Te Beest and Münch [202] for details on practical implementation of MML.

In the Bayesian literature, methods involving a data-driven choice of a single K-value are often called *empirical* [179]. All frequentist methods discussed in this section are of this type. On the Bayesian side, such methods contrast with *hierarchical* ones, which use a suitable discrete prior on K and allow it to "vary" [179]. In general terms, the K-values obtained with any approach will reflect what is necessary to capture the true shape of f_0. It is in this respect that these estimators are said to be "adaptive".

4.1. Histograms

Perhaps the simplest of density estimators, a histogram (sometimes referred to as an *empirical density* [169]) is piecewise constant over some division of the support into disjoint subsets, or "bins". In the most general form with countably many bins $\{A_j\}$, it can be written as

$$f(x) = \sum_j c_j \mathbb{1}_{A_j}(x), \tag{8}$$

with the constants c_j chosen to ensure that the estimator is a valid density. The regularity of a histogram is controlled by adjusting the sizes of the bins. For instance, a very common form for the univariate case is

$$\hat{f}(x) = \frac{K}{n} \sum_j n_j \mathbb{1}_{\left[\frac{j-1}{K}, \frac{j}{K}\right)}(x), \tag{9}$$

where n_j is the number of sample values in the interval $\left[\frac{j-1}{K}, \frac{j}{K}\right)$ and K provides the needed regularity control. Assuming a bounded support, say $[0, 1]$, the sum in (9) is over $j = 1, \ldots, K$ and is equivalent to (7) using a basis of indicator functions: $B_{j,K} = K\mathbb{1}_{\left[\frac{j-1}{K}, \frac{j}{K}\right)}$.

Temporarily ignoring the notion of empirical or hierarchical approaches to K, suppose for now that it is fixed at some arbitrary value irrespective of everything else. Then the histogram simply becomes a problem of multinomial inference: the coefficient b_j in (7) is an estimate of the probability that $X \sim f_0$ falls in the j^{th} bin, say p_j. In this respect, the "traditional" histogram, where $\hat{b}_j = n_j/n$ as in (9), is a MLE. Here the object of inferential interest is not necessarily f_0, but rather the so-called *theoretical histogram* $\overline{f}$ [194], a piecewise-constant density equal to Kp_j in the j^{th} bin. With this view, (piecewise-constant) pointwise intervals arise by considering the single binomial proportion p_j, and simultaneous bands by considering the vector of multinomial probabilities $\boldsymbol{p} = (p_1, \ldots, p_K)$. Frequentist and Bayesian methods for both are well-studied; see Vermeesch [207] for some practical applications to histograms.

4.1.1. Simultaneous frequentist inference

To discuss inference for f_0 itself, it is necessary to return to the adaptive paradigm. Much of the frequentist literature for histogram density UQ is theoretical and predates developments such as the bootstrap, relying on extreme value asymptotics similar to those for KDE's. One of the first such papers is by Smirnov [194]. They derived a limiting result much like (6) for the normalized quantity

$$\sqrt{nK^{-1}} \sup_x \frac{\left|\hat{f}(x) - f_0(x)\right|}{\sqrt{\overline{f}(x)}}, \tag{10}$$

where the supremum is over a compact interval on which f_0 is bounded away from 0 and has total mass less than 1. Smirnov claimed that it was not possible to replace $\overline{f}$ in the denominator with f_0 due to the systematic difference between them dominating the error. However, they stated that it *is* possible to do so by replacing the histogram with a *frequency polygon* (a linear interpolation between the histogram values at the sample points) and imposing some extra conditions on the relationship between K and n. Although Smirnov did not provide proofs for these results, they will be shown later to be a special case of proven results for wavelets [69]. For f_0 supported on a compact interval, Révész [169] was able to prove a somewhat modified extreme value limit for the distribution of a quantity similar to (10), except $\hat{f}$ can be either the traditional histogram or a frequency polygon (with a slightly different interpolation scheme than that considered by Smirnov), f_0 replaces $\overline{f}$ in the denominator, and the supremum is taken over an interval converging to the whole support of f_0. Révész also derived a similar result with the absolute value removed from the supremum. Further results to this effect were given by Freedman and Diaconis [58]. For everywhere-positive densities with a unique maximum, they considered a quantity similar to (10) without the absolute value (i.e. considering only the maximum *positive* deviation, although they claimed their proofs can be adapted

to the maximum absolute deviation), the supremum taken over the whole real line, and the factor of $\overline{f}(x)$ in the denominator replaced by the maximal value of f_0 (a fixed constant). Their limiting results are quite similar to those of Révész.

The three papers just discussed allow for (using a moderate amount of algebra) the construction of asymptotically correct simultaneous confidence bands for univariate densities satisfying suitable technical conditions, provided K increases at a suitably fast rate with respect to n (this roughly corresponds to the notion of undersmoothing discussed in Section 3). However, these papers did not concern themselves with the practicality of these ideas applied to actual data. It seems reasonable to suspect that slow convergence could be an issue which could be rectified with bootstrap methods, as was the case with KDE's.

4.1.2. Pointwise frequentist inference

Consider now the issue of frequentist pointwise intervals. For the "traditional" histogram (of the form (9)), Laloë and Servien [115] showed conditions on K and $f_0(x)$ for the quantity

$$\sqrt{nK^{-1}}\frac{\hat{f}(x) - f_0(x)}{\sqrt{f_0(x)}}, \tag{11}$$

to have a limiting distribution, which they proved to be a standard Gaussian when it does exist. Their proof applied the Lindberg-Feller Central Limit Theorem to the histogram values (recall that these are scaled binomial random variables for each K). Their conditions were more general (but also more technical) than those in many other papers: in particular, they did not even require f_0 to be continuous. Some literature provides results for non-traditional histogram variants with non-uniform bin spacing. For univariate densities, Kim and Van Ryzin [106] showed pointwise asymptotic normality for a histogram with randomly-spaced bins. They required bin spacings to meet certain conditions for their results to hold; one valid option is to fix the number of observations in each bin and determine their widths by the spacings of the sample's order statistics. The same authors showed analogous results for an extension to the bivariate case [107]. Another variant for the univariate case is the *maximum entropy histogram estimator* (MEHE), which works by dividing the real line into $K \leq n$ subintervals (with the first and K^{th} respectively extending to $-\infty$ and $+\infty$ and $\hat{f}$ having some suitable tail behaviour there) and choosing the spacing of their boundaries to maximize entropy subject to preservation of sample means and mass in the subintervals. Rodriguez and Van Ryzin [175] considered this estimator and a "symmetrized" variant and showed pointwise asymptotic normality of the quantity (11) for both. Their conditions on continuity and growth of the number of subintervals were slightly different for the symmetrized version, and the limiting law does not concentrate around zero as it does for the regular MEHE, presumably necessitating bias correction. Stadtmüller [195] considered asymptotics for yet another variant of the form (9), first considered by Gawronski and Stadtmüller [59], in which the indicator functions in the summands

are replaced by the values of *lattice distributions* to yield a smoothed estimate. They gave a few suitable examples: replacing $\mathbb{1}_{\left[\frac{j}{K},\frac{j+1}{K}\right)}(x)$ by $\mathbb{P}(Y=j)$ with, say, $Y \sim \text{Bin}(K,x)$ for densities supported on $[0,1]$, or with $Y \sim \text{Poi}(Kx)$ for those supported on $[0,\infty)$. Note that the lattice distributions do *not* necessarily constitute probability distributions with respect to x. Thus, density estimators of this type may not integrate to 1, although some examples presented in [59, 195] certainly will. For these estimators, Stadtmüller [195] showed pointwise asymptotic normality of the quantity

$$\left(\frac{4\pi\sigma^2(x)n^2}{Kf_0^2(x)}\right)^{1/4}\left(\hat{f}(x)-\mathbb{E}\hat{f}(x)\right),$$

where σ^2 depends on the lattice distributions. Additionally, they showed extreme value limiting results [somewhat similar in form to those in 9, as usual] for the supremum of this quantity (as well as the supremum of its absolute value) over compact intervals, under some regularity conditions on f_0 and the lattice distributions. They also noted that it is possible, as usual, to replace $\mathbb{E}\hat{f}$ by f_0 (thereby achieving correct asymptotic coverage for confidence intervals or bands) by undersmoothing - in this case, increasing K at a higher-than-optimal rate with respect to n.

4.1.3. A Bayesian approach

Recently, Rousseau and Szabó [179] discussed theory for Bayesian UQ of histogram estimators, assuming univariate f_0 supported on a compact interval. For this, return to the form (7), with the basis functions equal to indicators for equally-spaced bins. Rousseau and Szabó considered credible sets of the form (1), where d is the *Hellinger distance* and $\hat{f}$ is a suitable centering point such as the posterior mean. They showed that, under some regularity conditions on f_0 (it must be bounded away from zero and sufficiently smooth, and satisfy a "general polished tail assumption" defined by the authors and briefly described below) and the prior (a suitable K-dimensional Dirichlet for $\boldsymbol{b} \mid K$, and others omitted here for brevity), posterior credible sets of this type have arbitrarily high asymptotic frequentist coverage if their diameter is increased by an appropriate factor. In mathematical terms, for any $\epsilon \in (0,1)$, there exists $L_\epsilon > 0$ such that

$$\begin{aligned}&\liminf_{n\to\infty}\mathbb{P}_0\left(C\left(L_\epsilon\sqrt{\log n}r_\alpha\right)\ni f_0\right)\geq 1-\epsilon, \\ &\text{where } \Pi\left(f\in C\left(r_\alpha\right)\mid \boldsymbol{X}\right)=1-\alpha.\end{aligned} \tag{12}$$

In fact, they showed the stronger honesty result that this limit inferior holds *uniformly* over a certain class of functions, and that the uninflated credible sets are also almost adaptive over this class (save for a logarithmic factor in the diameter contraction rate). The densities comprising this class are those in an arbitrary union of Hölder balls of equal radius and regularities in $(1/2,1]$.

They must also satisfy the aforementioned general polished tail assumption, which essentially controls their high-resolution behaviour. Further ideas of this type will emerge in Section 4.4. Rousseau and Szabó's results hold for both the empirical and hierarchical approaches to K. For the latter case, a geometric or Poisson prior for K satisfies the relevant conditions. The authors noted that the "blow-up factor" of $\sqrt{\log n}$ is unfortunate, but they believe it is necessary to prevent coverage from decaying to zero in certain cases. Although it is quite pleasant to have such theoretical guarantees, it may be a challenge to put them towards a practical end due to the blow-up factor. Given an MCMC method to generate posterior simulations from this model, a credible set can be roughly visualized by plotting the $(1-\alpha)100\%$ of f draws closest in Hellinger distance to $\hat{f}$, but plotting draws from the blown-up set is another matter since we are not aware of any way to estimate L_ϵ.

Given the popularity of histograms, it is somewhat surprising that practical implementations and demonstrations of UQ for them appear so rare in the literature. For practitioners thorough enough to quantify errors in density estimation, it is perhaps reasonable to conclude that histograms have been superseded by KDE's and other methods that produce smooth estimates. Certainly, smoothness is advantageous for interpretation, especially when one wishes to account for uncertainty.

The following sections will contain a few more results which are applicable to histograms, arising as special cases of other methods.

4.2. Bernstein polynomials

One of the earliest non-histogram methods of the type (7) was proposed by Vitale [208] for densities supported on $[0,1]$. They took

$$B_{j,K} = \text{Beta}\left(j, K-j+1\right),$$
$$\hat{b}_j = \frac{\#\left\{X_i \in \left(\frac{j-1}{K}, \frac{j}{K}\right]\right\}}{n}.$$

The basis functions are Beta densities with integer parameters (equivalently, scaled *Bernstein polynomials*), and the coefficients are equal to the proportion of sample values in each interval $\left(\frac{j-1}{K}, \frac{j}{K}\right]$. In this respect, Vitale's estimator is essentially a smoothed histogram. In fact, aside from a different scaling factor it is almost the same as the lattice-smoothed histogram of Gawronski and Stadtmüller [59]; it therefore seems reasonable to suspect that one could derive confidence bands from similar asymptotic arguments as Stadtmüller [195]. An equivalent way to interpret this estimator is as a mixture of Beta densities.

Babu, Canty and Chaubey [3] provided pointwise asymptotic normality results for this estimator under mild conditions, from which one can presumably derive expressions for approximate pointwise confidence intervals (subject to the usual handling of bias and variance terms). At interior points $x \in (0,1)$, Vitale's estimator is quite similar to the KDE: optimal MSE behaviour occurs with K

such that the asymptotic orders of variance and squared bias match [208]. With this choice, confidence intervals - based on either plug-in or bootstrap methods - will not have the correct asymptotic coverage, concentrating around $\mathbb{E}\left[\hat{f}(x)\right]$ instead of $f_0(x)$ [3]. "Correct" intervals could be obtained by undersmoothing, choosing a higher-than-optimal K [asymptotic conditions in 3]. Alternatively, noting that the bias term is a known function of the first two derivatives of f_0, it may be reasonable to estimate a bias correction with plug-in methods, again in analogy with KDE's. Tenbusch [200] proved analogous results for Vitale-style estimates of bivariate densities defined on triangular or rectangular regions, with some generalizations for the latter provided by Babu and Chaubey [4]. As they are quite similar to the univariate case, they are not repeated here. The aforementioned pointwise results are valid for interior points x, but these estimators are known to have different asymptotic behaviour at the boundaries [208, 200] and so UQ may also work differently there.

There are methods besides Vitale's for estimating a density with Bernstein polynomials. For another frequentist method, take the coefficients $\hat{\boldsymbol{b}}$ to be MLE's. Guan [81] claimed pointwise asymptotic normality results for this approach, but it is not clear how to turn these results into appropriate pointwise intervals.

Theory for Bayesian estimates of this type typically depends on the idea of viewing the coefficients $\boldsymbol{b}$ as increments of some unknown c.d.f. F: $b_j = F\left(\frac{j}{K}\right) - F\left(\frac{j-1}{K}\right)$ (Vitale's estimator fits this framework for F equal to the e.d.f. of $\boldsymbol{X}$). To that end, Petrone [160, 161] considered a hierarchical Bayesian formulation with a discrete prior on K and a Dirichlet Process prior on F. For practical implementation, they devised an equivalent formulation making use of the aforementioned "mixture-of-betas" interpretation. They introduced a vector of latent variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \sim F$, which provide "mixture labels" for the samples conditional on K: $X_i \mid Y_i, K, F \sim \text{Beta}\left(\lceil KY_i \rceil, K - \lceil KY_i \rceil + 1\right)$. See Petrone [161] for more details on the properties of this construction, as well as a Gibbs sampling algorithm for posterior inference. In principle, this formulation gives everything needed to obtain, at the very least, pointwise credible intervals - indeed, Petrone did so in these papers. Note that practical implementations of this model require the truncation of the prior for K for computations to be possible. This has theoretical implications, but is not an issue in practice provided the maximum value for K is reasonably high. Petrone and Veronese [162] generalized these ideas for data not necessarily in $[0, 1]$; see Section 7.3.1 for elaboration on this.

Following the analogous KDE ideas in Lo [126], Ghosal [64] considered an alternative "posterior" based on a (generalized) Bayesian bootstrap approach, where it is assumed that $\boldsymbol{X} \sim F$ and F is a random distribution from a Dirichlet Process with base measure $\alpha(\cdot) + \sum \delta_{X_i}$. They conjectured pointwise asymptotic normality (concentrating around the Bernstein density with coefficients from $F = F_0$, rather than f_0 itself), but could not adapt the results in Lo [126] to prove this.

4.3. B-splines

B-splines are another option for the basis functions in (7). They are piecewise polynomials, characterized by a set of points in the domain called *knots* at which the values of the piecewise functions and a certain number of their derivatives must match. The number of basis functions K depends on both the number of knots and the polynomial degree chosen. Cubic splines are the most common choice, but there are others: for instance, a Bernstein basis of size K is a special case of B-splines of degree $K-1$ and no interior knots [45]. Using interior knots allows B-splines to be sharper-peaked in general than Bernstein polynomials, even at lower degrees. Literature about B-splines abounds; see Dias [40] for one of many introductions.

Although there will be plenty of discussion of splines in Sections 5.1 and 6, their use in estimators of the form (7) is limited in the literature. UQ for such estimators appears limited to practically-oriented Bayesian papers, although we suspect that it may be possible to translate some of the theory pertaining to histograms or Bernstein polynomials to this type of basis. Note that it is necessary to normalize each B-spline so it integrates to 1, thereby preserving the "mixture-of-basis-densities" view of (7). As another technical note, here attention is restricted to compactly-supported densities and estimators.

Shen and Ghosal [190] considered the hierarchical Bayesian setup (as defined at the beginning of Section 4), with K having a suitable discrete prior and $\boldsymbol{b} \mid K$ having a conditional K-dimensional Dirichlet prior. Like most practically-oriented papers with a hierarchical framework, they noted that the prior on K must be truncated for computation. They gave a closed-form expression for the posterior mean of f and claimed similar expressions existed for higher posterior moments, allowing them to construct approximate credible intervals (presumably by a Gaussian-style "mean $\pm$ 2*standard deviation" approximation). Their expression for the posterior mean is a ratio of sums, each of which has a number of terms increasing exponentially in n for splines of degree ≥ 1. Thus, the authors suggested randomly sampling a reasonable number of summands, say 3000, to approximate it. This is not an issue for splines of degree 0, as many terms cancel out due to the basis functions having non-overlapping supports. In this case, the estimator is simply a histogram and simplicity arises at the expense of smoothness. Shen and Ghosal found in their simulation study that the credible intervals were more appealing with cubic splines than with constant ones, although both had some difficulty capturing some of the true density's shape.

Edwards, Meyer and Christensen [45] compared Petrone-style Bayesian formulations [161, although Edwards et al. modified the MCMC] using both the Bernstein basis [see also 31] and B-splines for estimating the spectral density of a stationary time series. This use case differs from the probability density estimation considered here, but some of their ideas are nevertheless interesting for our purposes. In addition to pointwise credible intervals, they also considered

simultaneous bands generated from median absolute deviations:

$$\hat{f}(x) \pm \xi_\alpha \text{MAD}[f(x)], \tag{13}$$

where $\hat{f}$ is the posterior median, the pointwise MAD's are taken over MCMC draws, and ξ_α is the $1-\alpha$-quantile (obtained via MCMC draws) of $\sup_x \left(\left| f(x) - \hat{f}(x) \right| / \text{MAD}[f(x)] \right)$. In a simulation study, they found that such bands had vastly superior coverage using B-splines instead of the Bernstein basis. Pointwise intervals for B-splines tended to be wider, but both these and simultaneous bands captured intricate shape details more effectively than when the Bernstein basis was used. This is because the compact support of B-splines allows them to more effectively capture sharp peaks. The authors noted, however, that B-splines resulted in longer computation times than the Bernstein basis. Lopes and Dias [129] used a semiparametric Bayesian model for densities, combining a mixture of normalized B-splines with Dirichlet-distributed coefficients with a mixture of parametric densities. As usual, a straightforward Gibbs sampler allowed them to obtain pointwise credible intervals from MCMC output.

4.4. Orthonormal wavelets

Briefly, the idea behind estimation with orthonormal wavelets is to express a square-integrable function f in the form

$$f(x) = \sum_{k \in \mathbb{Z}} \sum_{j \in \mathbb{Z}} c_{kj} \psi_{kj}(x), \tag{14}$$

where $\psi_{kj}(x) = 2^{k/2} \psi\left(2^k x - j\right)$ for some suitable function ψ called the *mother wavelet*. The mother wavelet is such that $\{\psi_{kj}\}$ is an orthonormal basis of $\mathrm{L}^2(\mathbb{R})$, so that $c_{kj} = \int f \psi_{kj}$. For most of the literature discussed in this section, it can be assumed unless otherwise noted that the domain is indeed all of $\mathbb{R}$. However, in some cases it is desirable to modify the wavelets so that they form a basis of, say, $L^2([0,1])$, and there are multiple approaches to this [e.g. 33].

It is often more convenient to express f as

$$f(x) = \sum_{j \in \mathbb{Z}} d_j \phi_{k_0 j}(x) + \sum_{k=k_0}^{\infty} \sum_{j \in \mathbb{Z}} c_{kj} \psi_{kj}(x), \tag{15}$$

where $\phi_{kj}(x) = 2^{k/2} \phi\left(2^k x - j\right)$ for a *scaling function* or *father wavelet* ϕ such that $\{\phi_{k_0 j}, \psi_{kj} : j \in \mathbb{Z}, k \geq k_0\}$ is also an orthonormal basis for $\mathrm{L}^2(\mathbb{R})$. The number k_0 corresponds to the "coarsest" level of detail under consideration. In most literature explored below, it is either left arbitrary or set to 0 when the domain is $\mathbb{R}$. When modifying the wavelets for use on $[0,1]$, the dimensionality of the basis will depend on k_0 and, for some methods, the "regularity" of ψ [33]. In this setting, k_0 may therefore be chosen to provide an appropriate set of basis functions [e.g. 14, 25].

The simplest wavelet example is the *Haar wavelet*, where $\phi = \mathbb{1}_{[0,1)}$ and $\psi = \mathbb{1}_{[0,1/2)} - \mathbb{1}_{[1/2,1)}$. In general, ϕ and ψ must be selected to mutually satisfy certain functional equations. For further detail on wavelet theory and examples, refer to Kaiser's excellent book on the subject [103].

In practice, to estimate a density with wavelets, one must truncate the sum over k in the second term of (15) to some upper limit K. In this respect, K is a bandwidth or "resolution": higher values introduce thinner wavelets into the sum that capture finer details, thereby reducing bias and increasing variance. In this respect, wavelets differ from other basis expansion methods, in which the shapes of the basis functions themselves change with K. As previously mentioned, the coefficients in the wavelet expansion are simply inner products between the density and the basis functions. Thus, to obtain a point estimate $\hat{f}$, the natural choice is to estimate d_j and c_{kj} by their empirical versions: the sample means of $\phi_{k_0 j}(\boldsymbol{X})$ and $\psi_{kj}(\boldsymbol{X})$, respectively. It is now clear that density estimators based on the Haar wavelet are simply histograms with evenly-spaced bins.

4.4.1. Frequentist L^∞ inference

Giné and Nickl [69] derived some theoretical results for confidence bands over a compact subinterval, taken w.l.o.g to be $[0,1]$, by treating certain types of wavelet estimators in a unified framework with KDE's. In their approach, $\boldsymbol{X}$ is split into two subsamples: one of which is used for a data-driven bandwidth selection procedure (as always, their arguments involved undersmoothing to ensure correct coverage), with the other used to obtain the estimate $\hat{f}$ with this bandwidth. Letting K denote the number obtained from the bandwidth selection procedure (the details of which can be read in [69]), their framework encompasses both

1. kernel density estimators with kernel $\kappa(x,y) = \kappa(x-y)$ and bandwidth 2^{-K}; and
2. wavelet estimators in the form of (15) with $k_0 = 0$, and the sum over k in the second term truncated to K terms. To unify these estimators with KDE's, the authors invoked a *projection kernel* defined in terms of the father wavelet: $\kappa(x,y) = \sum_k \phi(x-k)\phi(y-k)$.

For a final piece of notation, let $c = \sup_x \int \kappa^2(x,y)\mathrm{d}y$. Giné and Nickl showed a result somewhat similar to the asymptotic KDE result in (6): for f_0 bounded away from zero on an open interval containing $[0,1]$, under some technical conditions the estimators in their framework satisfy

$$\mathbb{P}\left[A_n\left(\sqrt{\frac{n2^{-K}}{c}}\sup_{x\in[0,1]}\left|\frac{\hat{f}(x)-f_0(x)}{\sqrt{\hat{f}(x)}}\right| - d_n\right) < z\right] \to e^{-e^{-z}} \tag{16}$$

for suitable (known) sequences A_n and d_n. Just as in Section 3.2, it is straightforward to use this limit to get asymptotically-correct confidence bands. The

authors further showed that these bands are honest and nearly[4] adaptive in a range of Hölder balls over all but a nowhere-dense (w.r.t. the Hölder norm) subset of the function space. However, they noted that their work is theoretically oriented and therefore cautioned against using these bands without assessing their finite-sample performance. Furthermore, the only wavelets they showed to fit into their framework were the *Battle-Lemarié* wavelets of order $1, 2, 3$, and 4. The scaling function for the Battle-Lemarié wavelet of order r is a B-spline of order r [36], so for $r = 1$ it reduces to the Haar wavelet.

Because (16) generalizes the histogram results first discussed by Smirnov [194] and the KDE results shown by Bickel and Rosenblatt [9], results of this type are often called *Smirnov-Bickel-Rosenblatt theorems*. Bull [16] showed that a Smirnov-Bickel-Rosenblatt result holds in the white noise model using symlets and Daubechies wavelets. The Daubechies wavelet of order r is a Haar wavelet for $r = 1$ and has increasing regularity for higher orders, but unlike the Battle-Lemarié wavelet it has the advantage of being compactly supported [16, 103]. Bull verified their results for orders $6 \leq r \leq 20$, using bases on both $\mathbb{R}$ and $[0, 1]$. Although the white noise model is not the focus of this review, they noted that these results could translate to the density estimation context via some of the Gaussian process theory in [69]. Indeed, the notion of equivalence between the white noise model and density estimation is established [e.g. 154], but the details are beyond the scope of this paper.

It was noted above that a Smirnov-Bickel-Rosenblatt confidence band could achieve honesty and adaptivity under certain conditions and restrictions on the function space. More broadly, discussion of these concepts often uses wavelet theory as a starting point, due to the nice theoretical properties of an orthonormal basis. Hoffmann and Nickl [95] considered another approach to ensuring the existence of adaptive and honest confidence bands in *finitely many* nested Hölder balls: removing subsets of functions from the lower-regularity ones to ensure "separation" from the smoother classes. By connecting this idea to hypothesis tests for the smoothness of f_0, they showed that, in the case of finitely many smoothness levels, such separation conditions are necessary and sufficient for the existence of honest and adaptive bands, and that these conditions are weaker than those imposed by [69]. The constructive part of their argument involved a uniform band centered at an estimator satisfying certain properties; their paper and the references therein suggested that a wavelet estimator would be a good choice for both $L^2(\mathbb{R})$ and $L^2([0, 1])$. Unfortunately, the radii of these bands depend on properties of the Hölder balls that are unlikely to be known in practice, rendering application implausible. Nevertheless, these results are useful to inform theoretical discussion of the behaviour of confidence sets. Bull [15] considered inference on a union of Hölder balls with diameters and regularities both varying over a continuum. The conditions they imposed on the function sets are similar to those in [69], but somewhat weaker. Specifically, they required the densities under consideration to be *self-similar*. Briefly,

[4] Their diameters shrink at a rate which is nearly optimal, save for the presence of an extra logarithmic factor.

self-similarity is a property of a function's wavelet expansion ensuring that it exhibits similar regularity at both small and large scales. Note that the general polished tail condition of [179] (see Section 4.1.3) is a generalization of this. Bull showed that this restriction excludes only a "negligible" set of functions in both the topological and probabilistic[5] sense, and that it is necessary and sufficient to achieve honest and adaptive confidence bands over a continuous union of Hölder balls. Refer to Sections 8.3.3 – 8.3.4 of [70] for a more in-depth discussion of the role self-similarity plays in nonparametric inference.

Bull described a rather complex procedure to construct such a uniform band centered at a truncated empirical wavelet estimator, using Daubechies wavelets or symlets of order $6 \leq r \leq 20$, modified to form a basis of $L^2([0,1])$. The procedure exploits self-similarity to estimate the true smoothness of f_0. Unlike the construction of Giné and Nickl [69], it does not require sample-splitting.

4.4.2. A practical approach

None of the literature discussed thus far in this section concerns itself with applications to real data. To the extent that there have been constructive results, they have tended in most cases to be rather complicated. For an example of somewhat more practically-oriented material, Chernozhukov, Chetverikov and Kato [30] developed $1-\alpha$ confidence bands of the form

$$\hat{f}_{\hat{l}}(x) \pm \hat{\sigma}_{\hat{l}}(x)\,(\hat{c}_n(\alpha) + c'_n) \tag{17}$$

over a compact subset of $\mathbb{R}^d$. The subscript $\hat{l}$ is a particular value of l, which is used to denote bandwidth (l replaces the usual letter K here for more streamlined notation as in the original paper). Much like Giné and Nickl [69], these authors cast both KDE's and wavelet estimators into the larger framework of estimators $\hat{f}_l$ based on some kernel κ_l (note that, unlike [69], they folded the bandwidth into the definition of the kernel). In fact, their framework also encompasses estimators based on *nonwavelet* projection kernels using other orthonormal bases such as Legendre polynomials. They considered univariate kernels and wavelets, and extended to the multivariate case by using elementwise products. Returning to (17), $\hat{\sigma}_l$ is an estimate of the standard deviation of $\hat{f}_l$, obtained using sample mean analogues of the relevant expectations [e.g. 57]. Letting $\mathcal{L}_n$ denote the space of possible bandwidths, $\hat{c}_n(\alpha)$ is an estimate of the $1-\alpha$ quantile of

$$\sup_{l \in \mathcal{L}_n, x} \left| \frac{\hat{f}_l(x) - \mathbb{E}\left[\hat{f}_l(x)\right]}{\sqrt{\mathrm{Var}\left[\hat{f}_l(x)\right]}} \right|.$$

The authors suggested obtaining $\hat{c}_n(\alpha)$ by using the *Gaussian multiplier bootstrap*: whereas the normal bootstrap takes repeated samples of size n from the

[5] By considering a natural prior distribution on the space of functions.

empirical distribution of $\boldsymbol{X}$, this version repeatedly samples n i.i.d. standard normal variables $\xi_1, \ldots, \xi_n$. Subsequently,

$$\sup_{l \in \mathcal{L}_n, x} \left| \frac{1}{n} \sum_{i=1}^{n} \xi_i \frac{\kappa_l\left(X_i, x\right) - \hat{f}_l(x)}{\hat{\sigma}_l(x)} \right| \tag{18}$$

is calculated, and $\hat{c}_n(\alpha)$ is taken to be the $1-\alpha$ quantile of this quantity over bootstrap replications. The numbers c_n' and $\hat{l}$ are chosen based on a separate application of the Gaussian multiplier bootstrap: the former is a scaled quantile of a different Gaussian multiplier process, and the latter is based on a modified application of the popular *Lepskiĭ's method* [121]. Chernozhukov, Chetverikov and Kato showed that - under some conditions on the necessary intermediate quantities, $\mathcal{L}_n$, and κ_l - the bands (17) constructed in this way are asymptotically honest and adaptive over a range of Hölder balls, subject to global upper and lower bounds on the densities and a modified version of the "self-similarity" notion mentioned previously. Furthermore, they showed that the worst-case coverage probability of their bands converges to the nominal level at a polynomial rate, asymptotically faster than the logarithmic rate associated with Smirnov-Bickel-Rosenblatt results. These theoretical results hold for KDE's with compactly-supported kernels and estimators using either compact or Battle-Lemarié wavelets, with regularity conditions based on the maximal degree of Hölder smoothness to which one wishes to adapt. In their supplementary material, the authors conducted a small simulation study. Although most of the intermediate quantities used to construct (17) must meet certain conditions (primarily in terms of their behaviour with respect to n), they simply experimented with predetermined numerical values for their simulations.

4.4.3. Frequentist L^2 inference

Robins and van der Vaart [172] investigated the construction of L^2 confidence sets for conventional frequentist wavelet estimators. For a given wavelet basis[6], let $\theta(f)$ denote the expansion coefficients for an arbitrary f, and let $\hat{f}$ be the usual empirical wavelet density estimator. Then their confidence sets are of the form

$$\left\{ f \in \mathcal{F} : \left\| \theta(f) - \theta\left(\hat{f}\right) \right\|_2 \leq \sqrt{\frac{\hat{\tau}_{K,n,\theta}}{\alpha} + \hat{R}_{K,n}\left(\theta\left(\hat{f}\right)\right)} + 2\hat{B}_K \right\}, \tag{19}$$

where $K = K(n)$ is a suitably-increasing bandwidth and the terms $\hat{\tau}$, $\hat{B}$, and $\hat{R}$ are estimates for variance, bias, and $\left\| \theta\left(f_0\right) - \theta\left(\hat{f}\right) \right\|_2$, respectively. They used sample splitting, assuming that the data used to calculate $\hat{f}$ is independent from that used for the other terms. These sets were shown to be honest

[6] Actually, Robins and van der Vaart considered general orthonormal bases, not just wavelets. However, it seems appropriate to discuss their paper in this context, and to handwave some of the notation and technicalities for the sake of brevity.

at level $1-\alpha$, and adaptive to the fullest extent allowed by the theory without further restrictions[7]. Robins and van der Vaart mainly concerned themselves with the theoretical properties of such sets in various contexts. In practice, they may be difficult to construct due to the (likely unknown) quantities required for the calculation of the various terms in Equation 19. Bull and Nickl [17] further expanded upon the results of Robins and van der Vaart in the $L^2([0,1])$ case, showing that honest and adaptive L^2 confidence sets over a wider range of regularity classes and Sobolev ball radii are possible by discretizing the smoothness range and using the "separation" approach from [69]. They constructed such a set in their proofs, somewhat similar in form to (19). Although they acknowledged the possibility of replacing some of the unknown terms in their construction by certain data-driven approximations, they did not consider applications to real data. Lerasle [122] provided a different approach to L^2 confidence balls, the full intricacies of which are omitted here. They used a model selection approach to determine the best approximation space (they dealt more generally with projection estimators on linear subspaces of an L^2 space, but for our purposes it suffices to consider the special case of wavelet estimators where the selection is for the truncation level) and a resampling method to estimate an L^2 norm needed in the radius of the set, thereby avoiding the sample splitting needed by some of the other literature discussed here. They showed that their confidence balls have the same adaptation properties as in [172], and that they are additionally *non-asymptotic*: they have correct coverage probability for *any* sample size n, not just in the limit.

4.4.4. Some extensions and Bayesian ideas

Lounici and Nickl [131] defined a wavelet-based deconvolution density estimator analogous to the kernel one described in Section 3, based on a deconvolution kernel using Fourier transforms of the error density and wavelet basis functions. They used concentration inequalities and Rademacher processes to construct a confidence band, the radius of which is a complicated expression depending on the unknown density of the observed noisy data (although they noted that it can be replaced by the deconvolution estimator in practice). Under some conditions on the bandwidth, error density, and smoothness of f, it is possible to control the probability with which the bands cover f_0 over all of $\mathbb{R}$.

For another somewhat unconventional theoretical example, Kerkyacharian, Nickl and Picard [105] developed an estimator for densities on homogenous compact manifolds such as spheres. Their estimator is based on a *needlet* expansion of the density, where needlets form a basis with multiresolution properties similar to wavelets. They proposed a confidence band of random uniform width, discussed its (limited) adaptivity, and showed that its coverage probability can be controlled by undersmoothing.

[7]For instance, if $\mathcal{F}$ is a Sobolev ball of regularity r, the "fullest extent" in this L^2 context means adaptation over any nested Sobolev balls of regularity $s \in [r, 2r]$ [17, 70]. Recall that the L^∞ context is even more restrictive than this [70].

Bayesian UQ literature for density estimators of this type is generally scarce, but some Bayesian results for histograms by Castillo and Nickl [25] can be more easily explained with the machinery of Haar wavelets. They used wavelet expansions of roughly the form (15) with $k_0 = 0$, the sums over j restricted to $j = 0, \ldots, 2^k - 1$ to ensure the Haar system forms a basis of $L^2([0,1])$, and the sum over k truncated to some upper limit K. This is equivalent to the basis function estimator (7) with 2^K piecewise constant basis functions instead of K. In the latter form, Castillo and Nickl placed a 2^K-dimensional Dirichlet prior on the histogram coefficients $\boldsymbol{b}$, with $K = K_n$ chosen as a deterministic function of n and the assumed Hölder regularity of f_0 (for regularities in the range $(1/2, 1]$). They proposed credible sets C based on a "multiscale" approach:

$$C = \left\{ f : \max_{j,k} \frac{\left| \left\langle f - \hat{f}, \psi_{kj} \right\rangle \right|}{w_k} \leq \frac{R_n}{\sqrt{n}} \right\}, \tag{20}$$

where the inner product is the standard one on $L^2([0,1])$, $\hat{f}$ is the usual empirical wavelet estimator of f_0, w_k is a sequence such that $w_k/\sqrt{k} \to \infty$ as $k \to \infty$, and R_n is such that $\Pi(f \in C \mid \boldsymbol{X}) = 1 - \alpha$. Note that R_n can be computed explicitly due to the conjugacy of the Dirichlet prior, since the likelihood depends only on the counts of observations in each "bin". Castillo and Nickl showed that the posterior over densities satisfies a sort of nonparametric Bernstein-von Mises property, and that these sets therefore have asymptotically correct frequentist coverage: $P_0(C \ni f_0) \to 1 - \alpha$ as $n \to \infty$. With a further refinement to their definition, their L^∞-diameters also contract at a nearly-optimal rate in the "big-O in $\mathbb{P}_0$" sense. Unlike many of the other methods in this section, honesty and adaptivity are not implied here as the authors did not show the asymptotics to be uniform over all f_0 in some class. Although the choice of bandwidth K depends on the regularity of the unknown f_0, they suggested that one could estimate a suitable bandwidth under a self-similarity assumption as in [69]. The geometry of these sets does not lend itself to visualizable error bounds. Instead, one can simulate from the posterior with MCMC, discard the 5% of function draws with the highest values for the multiscale quantity on the left-hand side of (20), and plot the remaining 95% to get a visual representation of the sets. This is the approach taken in, for example, the simulation study of [167], who considered similar theoretical ideas for Bayesian UQ in the context of white noise and conjectured that they may be applicable to densities.

5. Adaptive basis expansion methods for log densities

An adaptive basis expansion does not have to be applied to the density itself as in the preceding section. Rather, it can serve as a model of the logarithm of the

density, provided normalizing constant c is incorporated:

$$\log f(x) = \sum_{j=1}^{K} b_j B_{j,K}(x) - c, \tag{21}$$

$$c = \log \int \exp \left[\sum_{j=1}^{K} b_j B_{j,K}(t) \right] \mathrm{d}t.$$

Modelling the logarithm as a sum has a few nice consequences. In particular, it allows f to be viewed as a member of an exponential family with sufficient statistics $\sum_i B_{j,K}(X_i)$, which makes it very easy to obtain an MLE $\hat{f}$ by maximizing $\sum_i \log f(X_i)$ with respect to $\boldsymbol{b}$ using (21) [e.g. 110]. Additionally, it is no longer necessary to constrain the coefficients.

5.1. Logsplines

One of the best-studied methods of this type is *logspline density estimation*. Assuming the density is supported on an interval and letting L and U denote its endpoints, let $\{B_{j,K} : j = 1, \ldots, K\}$ be a B-spline basis with knot sequence $L < t_1 < \ldots < t_m < U$ (recall Section 4.3). Although cubic splines are the most common choice, lower orders are possible; in particular, using splines of "order" 1 (equivalently, degree 0) corresponds to a histogram [197]. It is common to put some constraint on the tail behaviour of the estimate when using cubic splines, especially (but not exclusively) when $(L, U) = \mathbb{R}$, in which case the MLE $\log \hat{f}$ is typically required to be linear on $(L, t_1]$ and $[t_m, U)$ [91, 110]. If the support is a compact interval, another option is to require $\left(\log \hat{f}\right)''$ to be zero at L and U to reduce variance near the endpoints [112].

Stone [197] discussed some asymptotic theory for the maximum likelihood logspline density estimator, assuming the support is a compact interval ($[0, 1]$ w.l.o.g.) and the knots are equally spaced. They showed that, when K increases to ∞ with n,

$$\frac{\hat{f}(x) - \overline{f}(x)}{\widehat{\mathrm{SE}}\left(\hat{f}(x)\right)} \xrightarrow{d} \mathcal{N}(0, 1)$$

for all $x \in [0, 1]$, where $\widehat{\mathrm{SE}}\left(\hat{f}(x)\right)$ is a standard error estimate involving values of the basis functions and derivatives of c with respect to $\boldsymbol{b}$ (actual expression omitted for brevity), and $\overline{f}$ is the deterministic logspline density obtained by maximizing the *expected* (with respect to f_0) log-likelihood. In a related technical report [196], Stone noted that this result can be used to obtain asymptotic confidence intervals for f_0, provided K increases with respect to n at a suitable rate depending on some underlying differentiability assumptions on f_0. As in many other cases, K must increase faster than the error-optimizing rate for this, leading to undersmoothing.

A more comprehensive and practical treatment of pointwise inference for logsplines was given by Kooperberg and Stone [112]. They considered more involved knot placement schemes: one that involves stepwise selection, addition, and deletion, ultimately selecting the number of knots to optimize a generalized AIC [see 112, and references therein]; and a free knot placement scheme where knot locations and coefficients are *jointly* maximized, with the dimensionality again chosen by AIC. In either case, it is possible to estimate a standard error for $\log \hat{f}$ and use it to get approximate Gaussian pointwise intervals for the log density. By exponentiating the endpoints of these, Kooperberg and Stone obtained approximate confidence intervals for f_0. The only difference between the two knot selection procedures in this regard is the dimensionality of the gradients and Hessians required for the standard error estimate: the free knot procedure requires more components, since it is necessary to include derivatives with respect to knot locations. Additionally, for the stepwise procedure (in which the knots are considered fixed), the authors considered confidence intervals obtained via the bootstrap: either using percentile intervals, or plugging a bootstrap estimate of the standard error into the Gaussian interval approximation. The final UQ option they considered was a fully Bayesian approach, in which they put a hierarchical prior on K, knot placement (conditional on K), and coefficients $\boldsymbol{b}$ (conditional on knot placement and K). They simply took simulation quantiles from a reversible-jump MCMC procedure as pointwise credible intervals. In their simulation study, Kooperberg and Stone found that, for non-bootstrap methods, intervals based on the free knot procedure had higher coverage than those based on the stepwise procedure, but all non-bootstrap frequentist approaches consistently undercovered. Bootstrap methods based on the stepwise procedure were much better, although the percentile bootstrap tended to overcover (i.e. the intervals were perhaps too wide). Using a bootstrap standard error estimate with the stepwise procedure therefore appeared to be the best option, especially due to computational savings since fewer resamples were required than for the percentile bootstrap. They reserved analysis of the Bayesian approach for a real dataset, where they found that the credible intervals were much narrower than the "bootstrap standard error" confidence intervals, suggesting that the Bayesian approach may undercover. In a different publication [111], Kooperberg and Stone expanded somewhat on these results. There they found that the non-bootstrap frequentist intervals could achieve appropriate coverage on average when their widths were modified by some uniform scaling factor. Factors between 1.34 and 1.55 sufficed in their simulations depending on the specifics of the standard error calculations, but it was not clear how well these would generalize. They also found once again that the Bayesian intervals appeared too small when applied to practical data, even with a larger prior covariance on the coefficients. Hansen and Kooperberg [91, in rejoinder to discussions] noted their challenges with UQ in Bayesian logspline estimation: they found it difficult to select priors that led to good point estimates *and* sensible credible intervals. More broadly, some authors have expressed skepticism about the usefulness of UQ for logspline density estimation, stating their view that pointwise confidence intervals do not generally provide useful shape information [110, 140].

5.2. General orthonormal bases

A few theoretical Bayesian papers discussed in Section 4 also provided analogous results for log density basis methods. Castillo and Nickl [25] modelled log densities with wavelets modified to form a basis of $L^2([0,1])$ instead of $L^2(\mathbb{R})$. The coefficients were given independent and identical priors - either Gaussian, Laplace, or something heavier-tailed - with a scale parameter depending on the Hölder regularity of $\log f_0$ (assumed to be > 1). Similarly to their histogram approach described in Section 4.4, the authors used a deterministic bandwidth choice and showed that multiscale credible sets of the form (20) have correct asymptotic frequentist coverage, with near-optimal diameter contraction possible with further refinements. The same comments about practicality made in Section 4.4 apply here. In a similar vein, Rousseau and Szabó [179] considered density estimators (supported on $[0,1]$) of the form (21) with an orthonormal basis of $L^2([0,1])$ such that $B_1 \equiv 1$, with the subscript K removed since they did not consider basis functions changing with K. Among other technical conditions omitted here for brevity, they assumed $\log f_0$ has (up to a normalizing factor) an infinite series representation in terms of this basis; equivalently, that the true density is a member of an infinite-dimensional exponential family. With a suitable prior on $\boldsymbol{b} \mid K$ (a normal distribution with independent components is one example satisfying their conditions), the authors showed that (12) holds with Hellinger balls for both empirical and hierarchical approaches to K, just as it does for the histogram model. As in that case, honesty and near-adaptivity (up to a logarithmic factor) results hold over functions in a Sobolev ball of regularity $> 1/2$ satisfying their general polished tail condition. Unfortunately, their results remain difficult to put into practice due to the existence of the "blow-up factor" in the diameter of the sets.

6. Roughness penalty methods

Some of the frequentist estimators considered in Sections 4 – 5 were MLE's. In the i.i.d. case, one chooses $\hat{f}$ to maximize

$$\sum_{i=1}^{n} \log f(X_i)$$

over all f in some predetermined class of possible estimators - generally those that can be expressed in the form of (7) or (21) - so that obtaining the estimate is simply a matter of optimizing the coefficients. In some cases it is advantageous to impose a further restriction on $\hat{f}$ to reduce variance or otherwise impose some desirable "baseline" shape properties. In this case, instead choose $\hat{f}$ to maximize

$$\sum_{i=1}^{n} \log f(X_i) - \lambda J(f) \tag{22}$$

over the estimator class, where the functional J is some *roughness penalty*. This term forces $\hat{f}$ or $\log \hat{f}$ (depending on the context) to more closely resemble

a function in the null space of J to an extent controlled by the *smoothing parameter* λ. A common choice for J is the integrated square of some linear differential operator: for instance, if $J : f \mapsto \int \left(\mathrm{D}^3 \log f\right)^2$, then as $\lambda \to \infty$, $\log \hat{f}$ is forced towards a quadratic shape, and therefore $\hat{f}$ towards a Gaussian [193]. For brevity, this case may be described as "penalizing [the size of] the third derivative" [166] on the log scale.

As indicated above, roughness penalties most commonly appear in the context of basis expansion methods, particularly spline fitting. When using splines with equally-spaced knots that do not repeat at the endpoints [47], an integrated squared k^{th}-order derivative penalty can be approximated by the sum of squared k^{th}-order differences between the coefficients. This simpler penalty gives rise to so-called P-splines, devised by Eilers and Marx [46]. In any case, such penalties are equivalent to quadratic forms in the basis function coefficients - for instance, the associated matrix for the aforementioned third derivative penalty consists of inner products between the third derivatives of the basis functions.

A Bayesian approach to roughness penalties is quite natural: it comes from viewing (22) as a log-posterior, with the first and second terms respectively corresponding to likelihood and prior. In this respect, the Bayesian methods of the previous two sections technically fit into this framework, but the focus in this section is on literature with a stronger emphasis on specific shape and smoothness restrictions imposed by the prior or penalty. The benefits of expressing penalties as quadratic forms as described above is now apparent: such a penalty is equivalent to an improper Gaussian prior on the spline coefficients (e.g. the P-spline penalty corresponds to a random walk), with λ commonly given a Gamma hyperprior [e.g. 118]. Note that this type of prior is only suitable when modelling the log-density with basis functions - when using a basis expansion for the density itself, care must be taken to ensure that it is nonnegative and integrates to one. Some examples of this approach are given in Section 6.2.

6.1. Penalty methods for log-scale basis expansions

Although roughness penalty density estimators had already been developed by Good and Gaskins [72], Silverman [193] appears to have provided some of the earliest results for the approximate distributions of such estimators. Letting $g = \log f$ and taking $J(g)$ (in a slight abuse of notation) to be the integrated square of some m^{th}-order linear differential operator on g, they considered the estimator $\hat{g} := \log \hat{f}$ which minimized (22) over all g such that

1. g has piecewise differentiable $(m-1)^{\text{th}}$ derivatives,
2. $J(g) < \infty$, and
3. $\int e^g < \infty$.

Silverman showed that, for bounded f_0 on a bounded univariate domain, $\hat{g}$ is asymptotically normal under suitable conditions on the higher-order derivatives of $\log f_0$ and the rate at which $\lambda \to 0$ as a function of n and m. In principle

this result could lead to some type of pointwise confidence intervals, but Silverman did not pursue this further. The mean and covariance functions for the limiting Gaussian process depend on eigenvalues of an inner product space of estimators, and it is not clear how to approximate these in practice. O'Sullivan [155] expanded further on Silverman's original ideas for univariate densities on compact intervals, and justified approximating $\hat{g}$ by cubic B-splines with knots at order statistics of $\boldsymbol{X}$. They proposed to calculate λ by approximations to either a cross-validation score or an AIC-type quantity, and penalized the second derivatives of the log-densities. For uncertainty quantification, O'Sullivan adapted an idea from the non-parametric regression setting [210]: treating (22) as a log-posterior for the coefficients $\boldsymbol{b}$ in order to obtain "approximate Bayesian pointwise intervals". In the density case, O'Sullivan took a second-order Taylor series approximation of the unpenalized likelihood component $\sum \log f(X_i)$. This lead to an approximate Gaussian log-posterior, from which they derived pointwise intervals on the log scale of the form

$$\log \hat{f}(x) \pm 2\sqrt{\frac{2}{n}\boldsymbol{B}(t)^{\mathrm{T}}\left[\hat{\boldsymbol{H}} + 2\lambda\boldsymbol{\Omega}\right]^{-1}\boldsymbol{B}(t)}, \tag{23}$$

where $\boldsymbol{B}(t)$ is a vector of basis function evaluations, $\hat{\boldsymbol{H}}$ is the Hessian (with respect to $\boldsymbol{b}$) of the unpenalized likelihood at $\hat{\boldsymbol{b}}$, and Ω is the matrix of inner products associated to the roughness penalty. Presumably, confidence intervals for f_0 could be obtained by exponentiating the above expression. O'Sullivan did not comment on the performance of these intervals in their simulation study, but noted that they were found to have good coverage properties in the non-parametric regression setting by Wahba [210].

There are other formulations besides the Silverman approach for density estimation with roughness penalties. One such Bayesian approach came from Lambert and Eilers [117], who essentially used logistic regression to produce a smoothed estimate of a histogram. Suppose the density is supported on a bounded interval, which is partitioned into J bins. Let u_j and m_j respectively denote the center of, and number of observations in, the j^{th} bin I_j. Then Lambert and Eilers proposed the model

$$(m_1, \ldots, m_J) \sim \text{Multinomial}(n, \boldsymbol{\pi}), \tag{24}$$

$$\pi_j = \frac{\exp\left[\sum_{k=1}^{K} b_k B_k(u_j)\right]}{\sum_{l=1}^{J} \exp\left[\sum_{k=1}^{K} b_k B_k(u_l)\right]}, \tag{25}$$

$$\boldsymbol{b}_{-K} \sim \mathcal{N}\left(0, (\tau\Lambda)^{-1}\right);$$

where the B_k's are B-splines with equally-spaced knots, $b_K = -\sum_{k=1}^{K-1} b_k$ for identifiability, Λ is a matrix of finite difference coefficients encoding a P-spline penalty, and τ is a precision parameter with a gamma hyperprior. For $x \in I_j$, one can take $f(x) = \pi_j / \ell(I_j)$ as a density estimate, where ℓ denotes the length

of the interval. This penalized spline structure, combined with a high number of reasonably narrow bins, ensures the appearance of smooth estimates. Lambert and Eilers proposed this framework as a flexible way to handle grouped data by dividing the support into a smaller number of "wide bins" and replacing (24) with a multinomial model for wide bin counts, the probabilities for which are sums of the corresponding fine-grid π-values. Using a modified Langevin-Hastings algorithm to generate posterior samples, Lambert and Eilers applied this model to simulated and real data, using a moderately-sized cubic spline basis ($K = 20$). Unsurprisingly, their pointwise credible intervals (obtained from MCMC draws) exhibited higher variance when using larger "wide bins". In an earlier technical report, the same authors considered extensions of this model to multivariate densities by simply using products of B-spline bases, possibly allowing different dimensionalities and roughness penalties in each dimension [116].

6.2. Penalty methods for direct basis expansions

Roughness penalties can also be applied when modelling the density itself, rather than the log density, with basis functions. Komárek, Lesaffre and Hilton [108] considered such a formulation to estimate the error density in accelerated failure time models. Rather than splines, they used Gaussian densities at fixed locations, which they noted to be the limiting case for B-splines as their degree tends to infinity. The number of basis functions in their model is determined by the desired distance between their means (which serve the same purpose as equally-spaced knots for splines), as is their standard deviation. To ensure their estimates were valid densities, the authors used a softmax transformation to obtain the coefficients $\boldsymbol{b}$:

$$b_k = \frac{e^{a_k}}{\sum_{l=1}^{K} e^{a_l}}. \tag{26}$$

For identifiability, it is necessary to fix, say, $a_K = 0$; a few other constraints on $\boldsymbol{a}$ are also necessary to ensure identifiability of other parameters in the failure time model. The roughness penalty, based on second- or third-order finite differences, is imposed directly on $\boldsymbol{a}$. Estimation and inference follow from similar ideas as in O'Sullivan [155]: Komárek, Lesaffre and Hilton took a penalized maximum-likelihood estimate choosing the smoothing parameter by an approximate cross-validation score, and used a second-order Taylor expansion to obtain approximate pointwise "posterior" intervals for the density. They noted that in a simulation study (which they did not show), this method of constructing pointwise intervals yielded better coverage results than asymptotic methods. Komárek and Lesaffre [109] used a Bayesian version of this construction to model the errors and random effects in an accelerated failure time model with interval-censored data. As one might expect, the "logistic-scale" coefficients $\boldsymbol{a}$ in (26) were given (aside from a single identifiability constraint) a Gaussian prior

with a (third-order) finite difference covariance structure, the scale of which is controlled by a smoothing parameter with a diffuse Gamma prior. Specifying the model in this way leads to related closed forms for estimated survival functions and densities of onset and event times. These functions can be simulated in an MCMC run, leading to pointwise credible intervals and means corresponding to posterior predictive functions. The simulation studyconducted by Komárek and Lesaffre [109] showed that such credible intervals did a good job of capturing the true densities of event and onset times, although their smoothness varied with different combinations of true random effect and error densities. Sharef et al. [189] provided an even more flexible Bayesian approach of this type to estimate the frailty density in a proportional hazards frailty model. They used a mixture of normalized B-splines and an optional parametric term, constrained to ensure the density has mean one. The authors considered the use of fixed splines, as well as a reversible-jump MCMC procedure allowing the number and location of knots (and therefore, of basis functions) to vary adaptively. For the latter, they put some truncated discrete prior on the number of knots, with their locations given a discrete uniform prior over a larger set of "candidate knots". Conditioned on dimensionality, they expressed the coefficients for the spline part of the model as in (26). They considered multiple choices for a smoothness-imposing prior on $\boldsymbol{a} \mid K$, listed below.

1. Simply taking the components of $\boldsymbol{a}$ to be i.i.d. Gaussians. The authors used this prior with adaptive knot selection, since the latter procedure controls smoothness automatically.
2. Taking $\boldsymbol{a}$ to be Gaussian with a covariance structure corresponding to second-order finite differences. The authors noted that this is only guaranteed to enforce smoothness for equally-spaced (fixed) knots.
3. Directly penalizing the second derivative of the spline mixture. This amounts to using a log-prior that is a quadratic form in $\exp(\boldsymbol{a})$ (with an associated matrix of inner products between B-spline second derivatives), divided by $(\sum_k e^{a_k})^2$.

In all cases, the prior for $\boldsymbol{a}$ has a scale parameter with an inverse-Gamma prior to control smoothing.The authors applied their approach to both simulated and real data, quantifying uncertainty with pointwise credible intervals from MCMC quantiles. Their simulation study showed that the adaptive knot selection approach without parametric component effectively captured the true frailty densities, although it required a sufficient quantity of data to do so (in particular, too few data clusters lead to wide pointwise intervals that did not adequately capture true shape information). On a real dataset with a modest number of clusters, they compared the fixed-knot version of their model (with second derivative penalty) to the adaptive knot procedure with different prior choices for the parametric component weight and number of knots. They found that the adaptive version with parametric components encouraged more smoothness in the posterior mean density and its credible intervals, to an extent determined by the choices of priors. However, the fixed-knot version with second derivative penalty performed best in terms of a modified Deviance Information

Criterion.

This section concludes with a rather novel frequentist approach from Sardy and Tseng [184] which is better-suited to densities that may not be smooth in the sense of piecewise differentiability. They used estimators which are either piecewise linear between the order statistics of $\boldsymbol{X}$, or piecewise constant between their midpoints (equivalently, splines of degree 1 or 0, respectively), and *total variation* as their roughness penalty. The penalty is easily computed since their estimators ensure piecewise monotonicity, so that total variation is simply the sum of absolute differences between function values at consecutive order statistics. The authors devised two approaches for selecting the smoothing parameter: a universal one (depending only on sample size, not sample values) engineered to control the behaviour of $\hat{f}$ when the true density is uniform; and one based on a sparsity ℓ_1 information criterion, in which λ and $\hat{f}$ are jointly estimated. They used the latter approach on real datasets with some tied values due to rounding, and obtained 95% pointwise confidence intervals by bootstrapping. The pointwise intervals had reasonable width and shape, and the authors noted that they may allude to the existence of additional modes not captured in the "point estimates" of the densities.

7. Random measure mixture methods

This section explores uncertainty quantification for the canonical nonparametric Bayesian method of density estimation. In the general case, this method employs (conditional) mixtures of the form

$$f(\cdot \mid G) = \int \kappa(\cdot \mid \theta, \phi) \, \mathrm{d}G(\theta), \tag{27}$$

where κ is some kernel with parameters θ and ϕ, and the integral constitutes a mixture over the domain of θ with respect to a *random* probability distribution G. The bulk of the nonparametric Bayesian literature uses infinite-dimensional discrete mixing distributions:

$$G(\cdot) = \sum_{i=1}^{\infty} w_i \delta_{Z_i}(\cdot), \tag{28}$$

where the weights and locations of the atoms - respectively, w and Z - are random sequences. The centrepiece of this Bayesian mixture model is the infinite-dimensional prior on G: a "distribution on distributions". As it pertains to density inference, the locations and weights are usually independent, with the former distributed according to some continuous "base measure" and the latter having a prior from one of two commonly-used broad classes.

1. *Normalized random measures with independent increments*, or NRMI's [168], in which unnormalized weights are generated from a Poisson point process [100] and subsequently normalized. The measure with unnormalized weights is a *completely random measure* (CRM).

2. *Gibbs-type random measures* of type[8] $\sigma \in (0,1)$, which are equivalent to *σ-stable Poisson-Kingman processes* [71, 124]. Briefly, these arise from NRMI's with intensity measure corresponding to the σ-stable subordinator[65, p. 604] by conditioning the distribution of the weights on their sum T, then mixing over an arbitrary distribution for T [163].

Assuming independence between weights and locations, each approach is a special case of the larger set of *Poisson-Kingman models* [163, 124], which are in turn a type of *species sampling model*. The *normalized generalized gamma* (NGG) processes comprise the intersection of these approaches [124], whereas the *Pitman-Yor* process [164] is an example of the second but not the first, as noted by Favaro and Teh [52]. It is well-known that both the NGG and Pitman-Yor processes admit the *Dirichlet process* as a limiting case when the parameter $\sigma \to 0$ [as mentioned in 128, for instance]; many Bayesian density inference papers are specifically devoted to so-called Dirichlet process mixtures. For the interested reader, Chapter 14 of Ghosal and van der Vaart [65] is an excellent exploration of the relationships between such discrete nonparametric priors.

For any of these priors on G, it is easily seen that its specification in the form (28) leads to another expression equivalent to (27):

$$f(\cdot \mid G) = \sum_{i=1}^{\infty} w_i \kappa(\cdot \mid Z_i, \phi). \tag{29}$$

Discussion of the theoretical aspects of UQ, such as asymptotic coverage probability, appears scarce in the literature for such estimators. Instead, the focus is on practical generation of uncertainty sets (usually pointwise credible intervals) from posterior samples obtained via MCMC. As one might expect, difficulty arises here due to the nonparametric nature of the quantity of interest - in particular, since the posterior distribution of G (this section hereafter adopts the bracket notation of Gelfand and Smith [61], denoting this posterior by $[G \mid \boldsymbol{X}]$) will be infinite-dimensional. The key to most ideas for MCMC sampling of this model is to reformulate it in a hierarchical way:

$$\begin{aligned} X_i &\sim \kappa(\cdot \mid \theta_i, \phi), \\ \theta_i &\sim G, \\ G &\sim P(\cdot \mid \psi). \end{aligned} \tag{30}$$

If there *are* additional hyperparameters ϕ and ψ, they are typically given their own independent priors, but these are not a main focus here. The latter encodes all parameters of the prior for G: for instance, for a Dirichlet Process prior with Gaussian base measure it may include the concentration parameter, as well as the location and scale of said base. Note that by the almost-sure discreteness of

[8] Other Gibbs-type random measures are possible for different values of σ. For $\sigma < 0$, they are mixtures (over the dimensionality) of finite-dimensional symmetric Dirichlet distributions; for $\sigma = 0$, they are mixtures (over the concentration parameter) of Dirichlet processes [71]. However, these are not typically seen in the density inference literature.

G, there is positive probability that $\theta_i = \theta_j$ for some $i \neq j$. In this respect, the model imposes a random partitioning or clustering of the data, where each cluster is comprised of all observations with the same θ value. With this formulation in mind, the known MCMC strategies divide into two main groups - *marginal* and *conditional*, depending on the way in which the infinite-dimensional parameter G is handled. The sections below briefly explain, and discuss the UQ implications for, each of these groups.

7.1. Marginal sampling methods

Marginal methods rely on integrating G out of the model and being able to obtain approximate samples from $[\theta \mid \boldsymbol{X}]$. Algorithms for this purpose are readily available when using the Dirichlet Process prior; see Neal [147] for a seminal review of them. In this case, it is easy to obtain a Monte Carlo estimate of the posterior mean density (denoted here as $f(\cdot \mid \boldsymbol{X})$, in keeping with the rest of the Bayesian notation in this section), as discussed by Escobar and West [48]. Letting $\theta*$ denote the parameter for a hypothetical new observation and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$, note that $f(\cdot \mid \boldsymbol{\theta}) = \int f(\cdot \mid \theta*) \, \mathrm{d}\Pi(\theta* \mid \boldsymbol{\theta})$. The integrand is simply the kernel κ, and the distribution $[\theta* \mid \boldsymbol{\theta}]$ is readily available. Assuming the base measure of the Dirichlet process is conjugate to the kernel (as in the Gaussian case, for instance), this integral has an analytic closed form. From there, the Monte Carlo estimate of $f(\cdot \mid \boldsymbol{X}) = \int f(\cdot \mid \boldsymbol{\theta}) \, \mathrm{d}\Pi(\boldsymbol{\theta} \mid \boldsymbol{X})$ is an average of the above quantity over posterior MCMC draws of $\boldsymbol{\theta}$. By the same token, it is easy in the conjugate case to quantify uncertainty with respect to $[f(\cdot \mid \boldsymbol{\theta}) \mid \boldsymbol{X}]$. This is essentially the approach suggested by Wang and Dunson [213] to find pointwise confidence intervals, although they further simplified inference by using a greedy algorithm to find an optimal partition of the data. They noted that the deterministic nature of their algorithm results in an underestimation of uncertainty.

Inference of this nature ether ignores or marginalizes out uncertainty in the weights of G. For marginal samplers, this seems to be fairly standard practice when obtaining posterior density estimates to construct credible sets. Shi et al. [192] used one of Neal's nonconjugate algorithms [147] and obtained posterior density draws by taking the mean of the $\kappa(\cdot \mid \theta_i)$'s for each MCMC draw of $\boldsymbol{\theta}$ [see also their R package 191, and its source code[9]]. This is equivalent to taking a mixture of cluster-specific kernels, each weighted by the number of observations in its corresponding cluster. Using kernel- and data-specific scaling and a low-information prior, Shi et al. obtained pointwise credible intervals for simulated and real data. Their framework can accommodate censored data, with pointwise uncertainty increasing in the presence of censoring as expected. In their simulation studies, the credible intervals did a good job of capturing the true densities, covering them throughout the domains for all one-dimensional examples and at roughly 98% of domain points for their two-dimensional example. Favaro and Teh [52] and Favaro, Lomeli and Teh [51] devised marginal

[9] Available at https://github.com/cran/DPWeibull.

Gibbs samplers for NRMI's and a specific subclass of σ-stable Poisson-Kingman models, respectively. For both classes, their density draw computations appear[10] to be based on truncation, and marginalization of the distribution of G given $\boldsymbol{\theta}$ and the auxiliary variables of the sampler. To elaborate, the density draws are a sum of cluster-specific kernels, each given weight proportional to its (conditional) expected mass; and ten "new" kernels with parameters taken from the prior base measure, each given weight proportional the expected *total* mass divided by ten. In the latter paper, the authors showed a pointwise credible interval for the density of a dataset of galaxy velocities, noting that the results were satisfactory and consistent with previous work.

It could be argued that the aforementioned approaches to density inference are inherently "incomplete". Indeed, marginalizing or otherwise deterministically approximating the random weights of G fails to account for some of the uncertainty in (29). If the goal is full uncertainty quantification in this regard, the focus must be on $[f(\cdot \mid G) \mid \boldsymbol{X}]$ if possible. As noted by Gelfand and Kottas [60], it holds that

$$[\boldsymbol{\theta}, G \mid \boldsymbol{X}] \propto [\boldsymbol{\theta} \mid \boldsymbol{X}]\,[G \mid \boldsymbol{\theta}]\,. \tag{31}$$

This reveals the key to fully meaningful inference with a marginal sampler: for each MCMC draw $\boldsymbol{\theta}_b \sim [\boldsymbol{\theta} \mid \boldsymbol{X}], b = 1, \ldots, B$, if it is possible to draw $G_b \sim [G \mid \boldsymbol{\theta}_b]$, then the quantities $\{f(\cdot \mid G_b)\}$ constitute a posterior sample from $[f(\cdot \mid G) \mid \boldsymbol{X}]$. Gelfand and Kottas [60] noted that this is easy for the Dirichlet process prior by conjugacy, since $[G \mid \boldsymbol{\theta}]$ is a Dirichlet process with updated parameters. Of course, in practice the infinite sum in (29) must be somehow truncated to obtain actual density draws. Gelfand and Kottas did this by choosing the number of terms to satisfy a predetermined expected error threshold, then replacing the final weight to ensure that the truncated sum integrates to one. Kottas [114] later used this approach in the context of survival analysis, as did Griffin [74] when comparing different approaches to hyperpriors in the Dirichlet process model. Such methodology is not typically used for more general random measure priors, despite relevant distributional results existing in the literature [52, 51]. This is likely a computational matter: to directly sample the weights $\boldsymbol{w}$ of a random measure, it is typically necessary to employ a *stick-breaking process*, in which they are represented as

$$w_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \tag{32}$$

for certain continuous random variables $\{V_i\}$ on $[0, 1]$. It is well-known that the Dirichlet process with concentration parameter M has a stick-breaking representation of the form (32) with $V_i \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, M)$ [188]. However, such representations for the general classes of random measure considered here are more recent developments, and the densities of the V_i's are quite complicated [54, 55].

[10] Based also on their source code at https://github.com/BigBayes/BNPMix.java.

7.2. Conditional samplers

In contrast to the approaches described above, conditional methods *do* produce posterior samples of the weights in (28), allowing for "full" inference on functionals such as (29). There are several ways to avoid the problem of having to sample infinitely many weights. Early conditional samplers simply replaced G by a finite approximation, choosing the deterministic truncation level *a priori*. Discussion of such methods is deferred to Section 7.4; this section focuses on alternatives that better incorporate the infinite-dimensional nature of the model. Perhaps the most common approach to this end is to introduce some auxiliary variables such that the full conditionals of G are finite-dimensional. This ensures that Gibbs samplers target the correct posterior without the need for approximation, aside from the inevitable truncation to calculate the density draws themselves.

The retrospective sampler of Papaspiliopoulos and Roberts [157] was one of the earliest methods of this type for Dirichlet process mixtures. It involves the introduction of *allocation variables* $\boldsymbol{K} = (K_1, \ldots, K_n)$ such that $K_i = j$ iff $\theta_i = Z_j$, with Z_j as in (28). At each step of the chain, first draw only $\max(\boldsymbol{K}) := \max_i \{K_1, \ldots, K_n\}$ of the atoms and weights in G. A certain condition involving auxiliary standard uniform variables and the full conditionals of $\boldsymbol{K}$ is then checked. If the condition is met, perform a Metropolis-Hastings update of $\boldsymbol{K}$ and resume sampling as normal; otherwise, simulate additional components of G one at a time (from their priors, as they represent clusters with no allocated observations) until the condition is met. Note that the number of components is therefore variable across iterations. The authors noted that posterior draws for any linear functional of G are equal in distribution to a deterministic function of prior draws and the first $\max(\boldsymbol{K})$ components from one retrospective sampling iteration. Thus, full posterior inference for $f(\cdot \mid G)$ is quite straightforward.

Another popular approach which avoids some of the computational burden of the retrospective algorithm is *slice sampling*, first used in this context by Walker [211]. Briefly, Walker's original idea involved introducing new latent variables $U_i, i = 1, \ldots, n$ such that, with K_i again denoting the allocation variable as above, the joint likelihood for observation i is

$$f(X_i, U_i, K_i = j \mid G) = \kappa(X_i \mid \theta_j)\, \mathbb{1}(U_i < w_j). \tag{33}$$

Integrating out K_i and U_i reduces this to (29). Furthermore, these variables ensure that all full conditionals in the Gibbs sampler - including those for the necessary components of G - are finite-dimensional. Numerous adaptations of the algorithm exist: for instance, Kalli, Griffin and Walker [104] altered it for greater efficiency. Technical details aside, the main point is to simulate the finitely (but randomly) many components of G needed for the other sampler variables; this can exceed n, in which case some components will correspond to clusters with no data allocated. Favaro and Walker [53] adapted the algorithm of Kalli et al. to the larger class of σ-stable Poisson-Kingman models, using their stick-breaking representation to devise a method for sampling the weights. They applied their

method with mixtures of Gaussians with common variance and means from the random measure. Density draws were calculated by first adding together the components obtained from the sampler, then allocating the remaining mass (which the authors noted was usually quite small) to a Gaussian kernel with the sampled posterior variance centered at the prior mean of the base distribution. One can then extract posterior sets from these draws in the usual way. In the same paper discussed in Section 7.1, Favaro and Teh [52] considered a slice sampler for NRMI mixtures; here they sampled the *unnormalized* masses of the random measure. They showed pointwise credible intervals for the densities of some real datasets that were reasonable in shape and variability. Their source code suggests that they used the same formula for density draws here as they did for their aforementioned marginal samplers, with ten additional "new" kernels as described in the previous section.

Although finite-sum approximations are always necessary for density estimation, the approaches described above are noteworthy because the samplers *themselves* introduce no truncation error; all of their full conditionals are truly finite-dimensional. This is not the case for *all* papers which use conditional samplers for density inference. For instance, Barrios et al. [7] used a conditional algorithm for NRMI's that does not induce a finite-dimensional full conditional for G. Instead, they used a representation which allowed them to sample the masses in decreasing order. This allowed them to select the number of components sampled based on a relative error criterion, and to calculate density draws from only these (normalized by the sum of the sampled masses). They obtained pointwise credible intervals for a real dataset, demonstrating that the choice of both kernel and NRMI prior can moderately affect the smoothness of said intervals. Argiento, Bianchini and Guglielmi [1] folded random truncation into a modification of the NRMI prior itself by discarding all unnormalized weights smaller than some threshold ϵ. The resulting random measures have finite and random dimension, and converge in distribution to the corresponding NRMI's as $\epsilon \to 0$. The authors recommended fixing some small value for ϵ (it is possible to place a prior on it, but they warned that the computational cost may be unreasonable). They derived a conditional sampling algorithm, introduced a new class of NRMI's with a Bessel function in the intensity measure, and applied their method to real and simulated data. Their pointwise credible intervals showed a pleasing degree of smoothness and reasonable faithfulness to the true density of a simulated sample. Griffin [75] proposed an adaptive truncation method based on sequential Monte Carlo. The method involves iteratively resampling and increasing the dimension of the approximate model until a discrepancy measure falls below some threshold. Griffin applied this approach to a variety of nonparametric models, including Dirichlet process mixtures. Although they did not show credible intervals for densities, they did so in the context of time series modelling, indicating that density inference is indeed possible in this framework.

7.3. Extensions

7.3.1. Feller-Dirichlet priors

This Bayesian model from Petrone and Veronese [162] generalizes the Dirichlet process mixture model, although it also serves as an extension of Petrone's ideas [160, 161] from Section 4.2. Recall from that section that Petrone put a prior on K and introduced latent variables $Y_1, \ldots, Y_n$ from a random distribution F with DP prior such that $X_i \mid Y_i, K, F \sim \text{Beta}\left(\lceil KY_i \rceil, K - \lceil KY_i \rceil + 1\right)$. The *Feller-Dirichlet prior* generalizes this by replacing the latter Beta densities by some kernels $g_K\left(\cdot; Y_i\right)$, leading to a density model of the form

$$f\left(\cdot \mid K, F\right) = \int g_K\left(\cdot; \theta\right) \mathrm{d}F\left(\theta\right).$$

Petrone and Veronese provided several examples beyond the original Bernstein model that are suitable for data on $[0, \infty)$ or $\mathbb{R}$. For instance, take $g_K\left(\cdot; \theta\right)$ to be an inverse Gamma density with parameters $(K, K\theta)$ with a Gamma base measure for the prior on F, or use a $\mathcal{N}\left(\theta, \sigma^2/K\right)$ density for the kernel with a Gaussian base measure. These examples illuminate the idea that the Feller-Dirichlet prior - a "mixture of DP mixtures" - bridges the gap between Dirichlet process mixture models and the Bernstein polynomial models explored previously. For inference, Petrone and Veronese truncated the DP to a large finite number of components and used a Gibbs sampler similar to that of Ishwaran and Zarepour [99] to obtain density estimates and pointwise credible intervals.

7.3.2. Extensions for non-i.i.d. data

Several extensions to the random measure density model also exist for data structures besides an i.i.d. sample $\boldsymbol{X}$, most of which are based on the Dirichlet process instead of the more general measures. Müller and Rodriguez [145] and the references therein provide an excellent overview of such extensions; this section details some examples for which uncertainty quantification has been done in literature. In broad terms, these examples all involve inference for a family of densities $\{f\left(\cdot \mid G_t\right) : t \in \mathcal{T}\}$, where the random measures are indexed by some set $\mathcal{T}$ and share some form of dependence. In many cases, this will mean modelling the density for a "response variable" X with associated covariate t, effectively building yet another bridge between density estimation and nonparametric regression.

The dependent Dirichlet process (DDP) first introduced by MacEachern [134] is the basis for many useful models. DDP mixtures are similar in construction to (29), except that the weights $\{w_{tj}\}$ and locations $\{Z_{tj}\}$ may both vary with $t \in \mathcal{T}$. For instance, De Iorio et al. [38] considered a model for survival analysis when there are covariates t_i associated with each observation X_i. The weights do not vary with t, but the Z_{tj}'s correspond to location-scale pairs with the former component equal to a linear model in t: for example, if $t = (u, v)$ for categorical u

and continuous v, then $Z_{tj} = \left(m_j, A_{uj}, B_j v, \sigma^2\right)$ for $j \in \mathbb{N}$. Inference proceeds by reformulating the model into the conventional DP mixture framework, replacing the top line of the hierarchy in (30) by

$$X_i \mid \theta_i, t_i \sim \mathcal{N}\left(\theta_i d_i\right) \tag{34}$$

where d_i is a design vector so that $\theta_i d_i = \left(m_j + A_{uj} + B_j v, \sigma^2\right)$ when $\theta_i = j$ and $t_i = (u, v)$. De Iorio et al. used this so-called *linear DDP* to analyze the densities of log survival times with various combinations of treatments and other factors. They showed some pointwise credible intervals for survivor and hazard functions, and although they did not do so for densities, it should be no more difficult. However, their inferential approach [as described in 37] is the same as that suggested by Escobar and West [48], where the weights are marginalized so that inference is based on $[f(\cdot \mid t, \theta) \mid \boldsymbol{X}]$ as opposed to $[f(\cdot \mid G_t) \mid \boldsymbol{X}]$.

The above formulation is somewhat similar to the *density regression* model considered by Dunson, Pillai and Park [43] for modelling the density of a continuous response variable. The model assumes a set of continuous covariates associated to each observation and a structure similar to (34), except that the random measure governing the θ-value for an observation now depends on the corresponding covariate vector t: it is a finite mixture of n i.i.d. Dirichlet processes, with the i^{th} weight based on the distance between t and t_i. Dunson et al. used a marginal sampler, so that posterior inference for the predictive density of a "new" observation (given some covariate vector) was once again based only on the finite-dimensional parameters. Draws for these densities have closed forms due to conjugacy: they are mixtures of cluster-specific kernels and one using posterior draws of the hyperparameters of the base distribution. For both real and simulated data, the authors showed pointwise credible intervals for such densities conditioned on various values for the covariates. In the latter case, the intervals did a good job of capturing the true densities. Dunson and Park [42] subsequently developed the *kernel stick-breaking process* (KSBP) to model an uncountable collection of probability distributions (with particular focus on the density regression application), generalizing and expanding upon some of the ideas in [43]. In this model, the covariate-dependent distribution for an observation's θ-value is an *infinite* mixture of "basis" random measures (typically either point masses or draws from a Dirichlet process) with stick-breaking mixture weights. To induce dependence on the covariates, the beta random variables defining the stick-breaking process are weighted by kernels evaluated at the covariate value and centered at random locations with some arbitrary prior distribution. Dunson and Park's MCMC algorithm for pointwise UQ was a hybrid between marginal and conditional: like [43], they marginalized over the basis random measures; but at the t^{th} step of the chain they sampled M_t mixture weights, where M_t is the highest index of an occupied cluster across the first t iterations. The authors repeated the same simulation study as in [43], showing that the pointwise credible intervals from the KSBP model enveloped the true densities. Norets and Pelenis [153] explored the same simulated data model, showing how changes in the KSBP hyperparameters affected the quality of inference.

The formulation in the preceding paragraph directly model the conditional density of X given t by specifying a covariate-dependent random measure. Alternatively, it is possible to first model the *joint* distribution of X and t as a mixture of a kernel $\kappa(X, t \mid \theta, \psi) = \kappa(X \mid t, \theta)\,\kappa(t \mid \psi)$ with respect to a random distribution on the product space for (θ, ψ), then obtain the desired (conditional) density estimates by standard calculations. This approach is used by Park and Dunson [158], who put a Dirichlet process prior on the product measure; and Wade et al. [209], who gave separate DP priors to G_θ and $G_{\psi|\theta}$ to allow for greater flexibility. Both used marginal samplers for inference (again, with uncertainty only in terms of the finite-dimensional parameters), with the latter finding pointwise credible intervals to be much narrower and more accurate than those resulting from a DP on the product measure.

Returning to the DDP, note that it is also a suitable starting point when there are multiple sets of observations from different discrete time points, in which case the density is a random process evolving through time. Nieto-Barajas et al. [150] used this approach in such a context, making the atom locations independent of time but introducing dependence into the weights through their stick-breaking construction. They achieved the latter by introducing latent variables Y_{tj} dependent on the stick-breaking proportion V_{tj}, such that $V_{(t+1)j}$ is in turn dependent on Y_{tj} and the usual Dirichlet process is recovered by marginalizing out the Y's. They applied this construction in a mixed-effects model for protein activation over time, using a partially marginalized algorithm which exploited conjugacy to sample only atoms corresponding to clusters containing data. Müller and Rodriguez [145] showed densities with pointwise credible intervals from the same application, presumably using the same algorithm. Gutiérrez, Mena and Ruggiero [82] used a different approach to introduce dependence in the stick-breaking process: with random probability p having a Beta prior, they sampled $V_{(t+1)j}$ from its usual distribution, and set it equal to V_{tj} otherwise. They used slice sampling for inference, but did not specify if their density draws incorporated any components beyond those sampled (recall that this was the case for the Favaro-authored papers in Section 7.2). Their simulation study showed that their method was much more effective than one based on spline regression at capturing the true shape of their density, but their pointwise credible intervals did a much better job at enveloping the true density at later time points than at earlier ones.

Finally, there may be multiple samples $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m$ for which it makes sense "share information", assigning mutually dependent densities to each sample. The hierarchical methods discussed in Müller and Rodriguez [145] and its references are perhaps the most natural ways of doing this, but there does not appear to be existing literature which specifically conducts UQ with these methods. Griffin, Kolossiatis and Steel [76] developed an interesting model: starting with p underlying i.i.d. CRM's, the mixing distribution for each density is the normalized sum of some sample-specific subset of the underlying measures. Griffin et al. called this the *correlated NRMI* model, and implemented it with a combination of slice sampling and a split-merge step (in which clusters are moved between the underlying measures to address posterior multimodality). Although

the main purpose of their model was assessing differences between distributions, they did show pointwise intervals for survival functions fitted from interval-censored data; as always, it seems reasonable to assume that density inference is possible by similar means.

7.4. Finite mixtures

As previously mentioned, one way around the difficulties of infinite-dimensional models is to simply truncate the sum in (29) at some level N. This case leads to a vector of weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ on the $N-1$-dimensional probability simplex. This was the approach taken by the early conditional samplers of Ishwaran and Zarepour [99] and Ishwaran and James [98], who considered generalized Dirichlet priors on $\boldsymbol{w}$ to approximate random measures with stick-breaking representations (namely, those for which the stick-breaking variables V_j in (32) have beta distributions). For instance, to approximate a Dirichlet process mixture with concentration parameter α, they would either give $\boldsymbol{w}$ a symmetric Dirichlet prior with parameters α/N; or truncate its stick-breaking representation, setting $V_N = 1$ to ensure the N weights summed to one. They gave asymptotic justifications (as N grows large) for both options. With the conditional samplers devised in these papers, approximate posterior inference is obviously possible. Of course, extensions to the types of data structures considered in Section 7.3 are possible. For instance, Chung and Dunson [32] modelled covariate-dependent densities using truncated random measures with stick-breaking weights derived from a probit model. Their structure for the weights incorporated a variable selection component, resulting in a rather flexible density regression framework. Finucane et al. [56] conducted a meta-analysis of child nutrition data by modelling the study-specific densities of interest with finite mixtures of normals, using probit model stick-breaking weights which incorporated individual time and location effects. Norets and Pelenis [152] modelled the joint distribution of a response variable and covariates with a finite Gaussian mixture, obtaining the conditional response densities with standard calculations. Their model allows for any number of discrete variables by mapping them to continuous latent variables. The pointwise credible intervals obtained in these papers showed reasonably good uncertainty quantification, although the choice of a fixed finite number of components naturally reduces their flexibility somewhat.

The focus thus far in this section has been overwhelmingly Bayesian. Frequentist approaches to mixture models do exist in the literature, but it is rare to see them consider density UQ as it is defined here. Roeder [176] provided one rather novel exception for mixture-of-Gaussians estimators with finitely supported mixing distributions. Given some bandwidth h for the Gaussian kernel κ, the mixing distribution $\hat{G}_h$ is uniquely chosen to optimize an asymptotically normal statistic based on sample spacings[11]. This statistic is nonincreasing in

[11] Roeder noted the analogy between such estimators and KDE's, the difference being the sample-dependent mixing distribution used. Similar connections and generalizations were briefly explored in Section 3.3.

h, and so it is possible to find a range of h-values such that the statistic falls within the $(\alpha/2)$- and $(1-\alpha/2)$-quantiles of the standard Normal distribution. The confidence set defined by Roeder is then the set of all estimators $f\left(\cdot \mid \hat{G}_h\right)$ as h varies through this range. This set is comprised entirely of finite mixtures (although the number of components for each is random), is easy to visualize, and provides correct coverage if the true density is assumed to be a mixture of Gaussians.

In addition to the KDE connection, it is easy to see parallels between finite mixtures and some of the basis expansion methods discussed earlier. Indeed, even if one were to put a prior on N (e.g. Norets and Pati [151], whose inference involved modelling conditional densities using covariate-dependent multinomial logit mixture weights), the model would be similar in principle to the fully Bayesian approaches in Section 4. Thus, beyond what has already been explored, there is little else to discuss here. The interested reader may refer to Gelman et al. [62] for some more details on working with models of this type.

8. Other methods

This section explores uncertainty quantification for an assortment of density estimation methods for which literature is too scarce to warrant separate sections.

8.1. Nearest neighbour methods

This classical density estimator is closely related to the KDE and is applicable to any density on $\mathbb{R}^d$. Let $k = k(n)$ be an integer increasing with sample size n, let $\|\cdot\|$ be some norm on $\mathbb{R}^d$ (typically Euclidean, but some other norms also satisfy the required conditions for some of the results discussed here), and for an arbitrary point $x \in \mathbb{R}^d$ let $R(k,x)$ be the $\|\cdot\|$-distance between x and the k^{th}-closest value in $\boldsymbol{X}$. Then for a kernel K, the *nearest neighbour density estimator* as defined by Mack [135] is

$$\hat{f}(x) = \frac{1}{R(k,x)^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R(k,x)}\right). \tag{35}$$

Unless otherwise stated, all results in this section require K to equal 0 outside of the unit $\|\cdot\|$-ball. A particularly common case arises from the uniform kernel:

$$\hat{f}(x) = \frac{k}{nV(k,x)}, \tag{36}$$

where $V(k,x)$ is the volume of the $\|\cdot\|$-ball centered at x with radius $R(k,x)$. Nearly all of the literature on NN density inference is theoretical, and closely mirrors the results discussed previously for KDE's[12]. For instance, Theorem

[12] Unfortunately, even the drawbacks are similar: most asymptotic results relevant to inference require a choice of k which is *not* optimal w.r.t. the mean square error [e.g. 34]

9.3.7 in Csörgő [34] is essentially a Smirnov-Bickel-Rosenblatt result for univariate NN estimators. Unlike the KDE and wavelet theorems, their formulation would lead to confidence bands over a certain *random* interval defined by order statistics of the sample, but they noted that this interval converges to the full support as $n \to \infty$.

Moore and Yackel [144] provided what appear to be the first asymptotic normality results for (35), showing that the limiting distribution could be made to center at f_0 under some conditions on its properties and the asymptotic behaviour of k. They also noted that the asymptotic variance of the NN estimator is smaller than that of the KDE at points x where $f_0(x)$ is small, claiming that this makes it more efficient for estimating density tails. Mack and Rosenblatt [136] expanded on this, noting that the NN estimator can be much more biased than the KDE in the tails, with the opposite relations holding for large values of $f_0(x)$. These observations, combined with the non-monotonic dependence of asymptotic bias on k, make error analysis here somewhat less straightforward than it is for the KDE.

Mack [135] derived slightly different asymptotic normality results than Moore and Yackel, centering at $\mathbb{E}\left[\hat{f}\right]$ instead of f_0. This allows for less restrictive conditions: for instance, theirs are the only results here which do not require K to vanish outside of the unit ball. Pointwise Gaussian limits centered on f_0 with some variant of usual conditions (among others, as needed) are also available for univariate NN density estimates with non-i.i.d. data structures, such as randomly right-censored data [142], observations from an α-mixing sequence [125, only for the uniform kernel], or randomly left-truncated samples [216, who actually implemented confidence intervals in practice using a plug-in estimator of the asymptotic variance].

A technical report by Rodríguez [173] [see also 174] made an interesting connection between NN estimators of the form (36) and KDE's: the former allocates the fixed mass k/n to the random volume $V(k,x)$, while the latter can be rewritten to show that it essentially does the opposite, spreading a random mass over a fixed volume. This observation motivated Rodríguez to view the two estimators as endpoints on a "continuum" of estimators of the form

$$\hat{f}(x) = \frac{c \int K\left(\frac{x-t}{\mu}\right) \mathrm{d}F_n(t)}{\int_0^1 h^d(t) \mathrm{d}\omega(t)},$$

where ω is a distribution on $[0,1]$ with mean c, and the (possibly random) number μ and function h meet certain technical conditions. Rodríguez showed how KDE's and uniform NN estimators arise as special cases and described everything in-between as "double smoothing": in the numerator (resp. denominator), the mass (resp. volume) given by F_n (resp. h^d) is smoothed with K (resp. ω). Rodríguez proved asymptotic normality for certain subclasses of these estimators in this report, as did Biau et al. [8] for another variant. Both cases are generalized NN estimators, and the results hold even using the "optimal" $k(n)$ with given asymptotic biases. It is possible to eliminate the bias and center at

f_0 with a suboptimal k_n, although the conditions for this are less restrictive here than in [144] at the expense of stricter smoothness assumptions on f_0.

8.2. Logistic Gaussian process estimators

This approach is usually Bayesian and involves density estimates of the form

$$f(x) = \frac{e^{g(x)}}{\int e^{g(u)} \mathrm{d}u}, \tag{37}$$

where the latent function g is given a zero-mean *Gaussian process (GP)* prior with hyperparameters γ governing the covariance kernel. The "logistic" transformation of g ensures that the estimates are valid densities: nonnegative and integrating to one. Riihimäki and Vehtari [170] explored some approaches for approximate Bayesian inference with this model with 1- or 2-dimensional densities. Technically, they assumed that g would be the sum of a Gaussian process and a parametric polynomial component, but they integrated out the coefficients for the latter so that the basis function values and hyperparameters could simply fold into the mean and variance of the GP. Similarly to Lambert and Eilers [117] (see Section 6.1), Riihimäki and Vehtari discretized the model, replacing the actual data with observation counts in a fine, equally-spaced partition of the domain. Assuming that the partition consists of J subregions and letting $\boldsymbol{m}$ and $\boldsymbol{g}$ respectively denote the vectors of observation counts and latent function values within each subregion, the likelihood $\mathbb{P}(\boldsymbol{m} \mid \boldsymbol{g})$ is essentially the same as (24 – 25) [117], except the B-spline values in (25) are replaced by the latent function values g_j for $j = 1, \ldots, J$. In turn, the prior $\Pi(\boldsymbol{g} \mid \gamma)$ for the latent values is simply the multivariate normal distribution induced by evaluating the GP prior at the center points of the subregions. The main object of inference is then the conditional posterior of $\boldsymbol{g}$ given the observation counts and hyperparameters (and, technically, conditioned on the chosen partition as well),

$$\Pi(\boldsymbol{g} \mid \boldsymbol{m}, \gamma) \propto \mathbb{P}(\boldsymbol{m} \mid \boldsymbol{g}) \Pi(\boldsymbol{g} \mid \gamma). \tag{38}$$

This posterior is not analytically tractable, so approximate methods must be used to employ this model in practice. As an alternative to MCMC, Riihimäki and Vehtari proposed the use of a *Laplace approximation*: a Gaussian distribution for $\boldsymbol{g}$ based on a second-order Taylor approximation to the log of (38). In order to quantify uncertainty in f, samples must be drawn from this approximate Gaussian posterior and transformed via (37). To this end, the authors showed that importance sampling can improve performance, and rejection sampling can also be incorporated to ensure appropriate tail behaviour if necessary. The model is completed by putting a prior on γ, but Riihimäki and Vehtari also considered the possibility of ignoring the uncertainty in these hyperparameters: marginalizing the approximate Laplace posterior over $\boldsymbol{g}$, maximizing it with respect to γ, and simply plugging in the resulting approximate MAP point estimate for γ. They found that their method performed (in terms of mean

log predictive density, evaluated with cross-validation for real data or w.r.t. the true distribution for simulated data) comparably with MCMC targeting the true joint posterior of $(\boldsymbol{g}, \gamma)$, as well as the Dirichlet process mixture models of Griffin [74]. The pointwise credible intervals for real and simulated data provided reasonable practical visualization for uncertainty quantification. However, one of their simulations showed that densities with varying amounts of smoothness throughout the domain can be challenging, as the MAP parameters needed to capture more narrow features can result in excessive roughness elsewhere. The authors also showed how their method can extend to density regression, modelling densities conditional on covariate values.

8.3. Pólya trees

The Pólya tree (PT) prior is a Bayesian nonparametric method for constructing a random probability measure, discussed in [119] and the first few references therein. The construction is based on a recursive partitioning of the domain and is most easily explained when the domain is an interval in $\mathbb{R}$. At the m^{th} level of partitioning, the interval is split into 2^m subintervals. It is common to set the partition boundaries to the dyadic quantiles of some base measure G_0 (i.e. $G_0^{-1}(j/2^m), j = 0, \ldots, 2^m$), thus "centering" the random measures drawn from the PT prior around this base [119, 145]. Associate to each m^{th}-level subinterval a binary number $\epsilon = \epsilon_1 \ldots \epsilon_m \in \{0,1\}^m$, and define a set of Beta random variables $\{Y_\epsilon : \epsilon \in \{0,1\}^m\}$ such that the $Y_{\epsilon_1 \ldots \epsilon_{m-1} 0}$'s are mutually independent and $Y_{\epsilon_1 \ldots \epsilon_{m-1} 1} = 1 - Y_{\epsilon_1 \ldots \epsilon_{m-1} 0}$. Finally, consider a random probability measure that assigns mass

$$\prod_{j=1}^{m} Y_{\epsilon_1 \ldots \epsilon_j}.$$

to B_ϵ, where B_ϵ is the subinterval associated to binary number $\epsilon = \epsilon_1 \ldots \epsilon_m$. Iterating this process over all $m \in \mathbb{N}$ results in a draw from the Pólya tree prior (so named because the recursive partitioning defines a tree with nodes corresponding to subsets), defined by the sequence of partitions and Beta parameters. A special case for the latter gives rise to the Dirichlet process, but they can also be tailored to almost surely produce absolutely continuous distributions [e.g. 119, 146], which is obviously more appealing for density inference.

It is possible to extend this construction to d-dimensional domains, for instance by using the construction of Hanson [92]. At the m^{th} level, the domain is partitioned into 2^{md} subsets, indexed by base-2^d numbers $\epsilon = \epsilon_1 \ldots \epsilon_m \in \{0, \ldots, 2^d - 1\}^m$ [145]. These subsets are formed by taking Cartesian products of the subintervals used in the univariate construction, then applying a suitable affine transformation. Probabilities are assigned to each subset in an analogous way to the univariate case, except that for a fixed $\epsilon_1 \ldots \epsilon_{m-1}$, the variables $\left\{Y_{\epsilon_1 \ldots \epsilon_{m-1} e}, e = 0, \ldots, 2^d - 1\right\}$ have a 2^d-dimensional Dirichlet distribution. Literature on multivariate PT's rarely entails any density UQ, so the remainder of this section focuses primarily on the univariate case.

Castillo [24] provided theoretical results for posterior inference with such priors on the unit interval, with partition boundaries at the dyadic rationals. In particular, they showed that, when f_0 is Hölder with regularity $\beta \in (0, 1]$ and bounded away from zero, a type of *Bernstein-von Mises result* holds (i.e. the posterior weakly [25] converges in $\mathbb{P}_0$-probability to a Gaussian process) when the Beta parameters of the prior grow suitably fast with m. The posterior must be centered at some estimator for f_0 for this to hold: either the posterior mean or, when the Beta parameters grow suitably slowly depending on β (note that this corresponds to "undersmoothing" of the posterior), a "canonical" estimate based on the Haar wavelet expansion of the empirical measure. Castillo noted that this result can lead to similar results to some of those discussed earlier for wavelet estimators [25]: namely, multiscale credible bands similar to (20) with Pólya trees should have correct frequentist coverage.

Practical implementations of Pólya tree models involve truncating the partitioning at some finite "terminal" level, rather than continuing it infinitely. By a well-known conjugacy result [e.g. 146, 92], the posterior for the PT prior is simply an updated PT, with the same partition and updated Beta (or Dirichlet, in the multivariate case) parameters for the Y_ϵ's. With the aforementioned truncation, densities samples from this posterior can be obtained by allocating the mass proportion within each terminal subset either uniformly [65, chapter 3] or according to the density of the base measure [as in 92]. The resulting densities will be discontinuous at the partition boundaries [146, 65] and are therefore perhaps not as "well-behaved" as one may prefer. In a survival model with longitudinal data and a PT prior on event times, Zhang, Müller and Do [218] addressed this issue by applying kernel smoothing to the actual posterior PT draws to obtain event time densities. There are other ways around this which change the structure of the model itself: mixing the prior over the parameters of the base distribution [92], adding random "jitter" to the partition boundaries [156], or mixing a kernel with respect to a PT measure [12]. Surprisingly, literature employing such methods does not tend to address UQ for densities. On the other hand, Nieto-Barajas and Müller [149] did so for their *rubbery Pólya tree* (rPT) prior, introducing dependence amongst the $Y_{\epsilon_1 \ldots \epsilon_{m-1} 0}$'s at level m (i.e. all "left nodes" in the tree at a given depth) through latent variables. The construction resembles that used to introduce dependence for time-series DDP's by Nieto-Barajas et al. [150] as discussed in Section 7.3, and recovers the usual PT prior by marginalizing over the latent variables. Conditional conjugacy allows for an easy Gibbs sampler, which the Nieto-Barajas and Müller implemented by truncating the partitioning process at some depth (using a depth between 5–8 in all experiments) and allocating the mass uniformly within each of the terminal subsets. Pointwise credible intervals in their simulation study fully contained the true densities, but were not smooth. Indeed, the rPT only "smooths" the estimates in the sense of reducing jump sizes between masses in neighbouring partition sets. Its dependence structure addresses variability, not continuity. Nieto-Barajas and Müller suggested mixing (either over a kernel w.r.t. a rPT prior, or over the parameters of the rPT's base distribution) when more smoothness is desired, but did not attempt uncertainty quantification with

such models.

A different extension of the model came from Hanson, Zhou and Inácio De Carvalho [93], when each X_i is observed at a spatial location t_i. Their object of interest was the predictive density (i.e. marginalizing over G) for a new X, and they proposed to modify the usual formula by weighting the contribution of each observation by some distance between their locations and that of the new X. Uncertainty was with respect to the (hyper)parameters of the PT prior and the distance function and was quantified with MCMC output. Their pointwise credible intervals appeared quite smooth; it is unclear whether this is the result of an actual procedure or merely the plotting functions used.

8.4. Multiscale estimators

This rather novel Bayesian approach from Canale and Dunson [23] uses multiscale mixtures of Bernstein polynomials as estimates:

$$f(\cdot) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{sh} \text{Beta}(\cdot; h, 2^s - h + 1). \tag{39}$$

The weights π_{sh} are constructed in terms of a stochastic process defined on an infinite binary tree. For the h^{th} node at tree depth s ($h = 1, \ldots, 2^s$), let S_{sh} be the probability of stopping at that node and R_{sh} be the probability (conditional on *not* stopping) of moving to the right daughter of node (s, h). These probabilities define a sort of "random climb" on the branches of the tree, which at each step either stops with some probability or else moves on to the next depth, randomly choosing either the left or right path. The weight π_{sh} is then the probability of the process taking the path to node (s, h) (starting from the root of the tree) and then stopping there. For instance, $\pi_{12} = (1 - S_{00}) R_{00} S_{12}$, and $\pi_{23} = (1 - S_{00}) R_{00} (1 - S_{12}) (1 - R_{12}) S_{23}$. The specification of the model is completed with priors $S_{sh} \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, a)$ and $T_{sh} \overset{\text{i.i.d.}}{\sim} \text{Beta}(b, b)$, where a and b can be fixed or given their own hyperpriors.

Canale and Dunson noted that this model induces an interesting multiscale clustering on the data: two data points may be assigned to the same tree node at some depth s, meaning they are sufficiently similar to be clustered together at this scale; but may occupy different nodes at a depth $r > s$, so that they are separated at a higher "resolution". For practical inference, they truncated or "pruned" to a maximum tree depth S by simply setting the stopping probabilities $S_{Sh} = 1$ for all h. Using a slice sampling approach, they devised an MCMC algorithm for inference, which alternates between two steps: assigning each observation to a tree node (equivalently, to a "multiscale cluster") given the π_{sh}'s, and updating the S_{sh}'s and R_{sh}'s given these allocations. Posterior density samples can then be obtained by plugging these probabilities into (39), truncated accordingly. Although Canale and Dunson did not show any credible intervals for densities in their paper, the corresponding R package for this type of model implements them readily [22].

8.5. Shape-restricted methods

If an *a priori* assumption can be made about the shape of the true density f_0 (for instance, that it is monotone or unimodal), one may wish to incorporate this into estimation and inference. A solid body of literature exists on the use of such shape constraints in nonparametric estimation, but only a subset of this literature specifically considers UQ for densities.

Perhaps the best-studied shape constraint is monotonicity, in which f_0 is assumed to be non-decreasing. In the frequentist setting, the so-called *Grenander estimator* is the canonical choice for estimation of f_0, and is also the MLE subject to the monotonicity constraint [73]. Letting F_n denote the empirical distribution function of $\boldsymbol{X}$, let $\hat{F}$ be the *least concave majorant* of F_n: the smallest concave c.d.f. such that $\hat{F} \geq F_n$ throughout the entire support (typically assumed w.l.o.g. to be $[0,1]$, or $[0,\infty)$) [77]. The Grenander estimator $\hat{f}$ is then the left derivative of $\hat{F}$, which turns out to be a step function with jumps at sample values and $\hat{f}(x) = 0$ for $x \leq 0$ and $x > X_{(n)}$ [165]. Prekasa Rao [165] derived a pointwise limiting distribution for this estimator, showing that with suitable standardization it is asymptotically equivalent to a particular functional of Brownian motion[13]. Groeneboom and Jongbloed [80] leveraged this fact to derive the asymptotic distribution of a likelihood ratio test statistic for $f_0(x)$ when f_0 has nonzero derivative in a neighbourhood of $x \in (0,\infty)$. The limiting distribution is that of a different functional of Brownian motion derived by Banerjee and Wellner [6]. The authors of that paper did not find an analytic form for this distribution, but provided estimates of its quantiles from simulation-based methods. Groeneboom and Jongbloed used these estimated quantiles to obtain pointwise confidence intervals with asymptotically correct coverage by inverting their likelihood ratio test. They also considered pointwise bootstrap intervals based on a boundary-corrected kernel (under-)smoothing of the Grenander estimator. The use of the bootstrap in this way is at least partially justified by an asymptotic normality result for this smoothed Grenander estimator [79] (indeed, there are a few modifications to this method that result in smooth, asymptotically normal estimators; see also [203]). Unfortunately, the bootstrap is unsuitable for inference with the unaltered estimator, due to the inconsistency results shown by Kosorok [113] and Sen, Banerjee and Woodroofe [187] and demonstrated in practice by the latter. However, both papers showed that consistency can be restored with a smoothed bootstrap (i.e. resampling from a modified kernel estimate of f_0, rather than from the empirical distribution). Kosorok further showed that smoothed bootstrap methods could be used to define an L^1-ball of functions centered at $\hat{f}$ with correct asymptotic coverage, based on the known asymptotic normality of the L^1-error [77]. Recall, however, that such sets are limited in visual interpretability. Uniform confidence bands were briefly considered by Durot, Kulikov and Lopuhaä [44], who derived a Gumbel limiting distribution similar to the Smirnov-Bickel-Rosenblatt results

[13]The full details of this functional are omitted here, but its distribution is commonly known as the *Chernoff distribution*, which commonly arises in shape-constrained nonparametric inference.

of Section 4.4.1. However, they believed that the technicalities required for data-driven construction of such a band were not worth exploring further. A recent preprint by Deng, Han and Zhang [39] proposed another method to construct pointwise intervals, based on the adaptation of an analogous method for inference in isotonic regression (see their manuscript for details). They suggested that their method, which involves suitable estimates of nuisance parameters in the limiting distribution, could be tailored to adapt to the smoothness of f_0 more readily than the method of Groeneboom and Jongbloed [80], but both methods require simulation-based estimates for quantiles of the complicated limiting distribution.

As an alternative to frequentist inference methods based on the Grenander estimate (or some modification thereof), Bayesian methods are also available. For instance, Martin [138] proposed an empirical prior in which the density is modelled as a finite scale mixture of uniform densities. The mixture weights and uniform density scales are respectively given Dirichlet and Pareto priors, both of which are calibrated so the prior over densities is centered at some predetermined mode. This mode (and the dimensionality of the mixture) can either arise from a *sieve* MLE (i.e. the MLE over a space whose size depends on n) or the Grenander estimate. Using simulated data and MCMC, Martin compared the pointwise credible intervals from this model to those obtained from a Dirichlet process mixture, and found that the empirical model resulted in higher coverage probability and shorter intervals on average.

The second most common shape constraint explored in the literature is arguably log-concavity, in which $\log f_0$ is assumed to be concave. As in the monotone case, frequentist UQ for log-concave densities typically centers on the MLE $\hat{f}$. Rufibach [182] showed that the log of $\hat{f}$ is piecewise linear with breaks at sample values, and $\hat{f}$ supported on the range of $\boldsymbol{X}$. Balabdaoui, Rufibach and Wellner [5] obtained a limiting distribution for this estimator under some regularity conditions on f_0. Much like the monotone case, the MLE for log-concave densities converges in distribution to a certain functional of Brownian motion, scaled by nuisance parameters that depend on the value of f_0 and its derivatives. Azadbakhsh, Jankowski and Gao [2] translated these results into practical means of constructing pointwise confidence intervals. They estimated the necessary quantiles of the limiting distribution by simulation, and considered several methods (kernel-based and plug-in) to estimate the f_0-dependent nuisance parameters. The intervals thus obtained performed reasonably well in a simulation study, although the best overall results came from standard bootstrap percentile intervals. Compared to the bootstrap intervals, the pointwise intervals based on asymptotics generally had a somewhat higher propensity for undercoverage in some parts of the domain, and for overcoverage (i.e. coverage probability exceeding the desired nominal level, leading to wider intervals than necessary) in other parts. Despite these promising empirical results, the authors cautioned that there were no theoretical results justifying bootstrap methods for this purpose.

Mariucci, Ray and Szabó [137] developed a Bayesian model for log-concave

densities f:

$$f(x) = \frac{e^{w(x)}\mathbb{1}_{[a,b]}(x)}{\int_a^b e^{w(u)}\mathrm{d}u},$$
$$w(x) = \gamma_1 \sum_{j=1}^{m} p_j \frac{\min\{\theta_j, x-a\}}{\theta_j} - \gamma_2 (x-a). \tag{40}$$

The function w is piecewise linear with m break-points, where m is a predetermined number dependent on sample size. The weights $p_1, \ldots, p_m$ can either be given a Dirichlet prior, or a prior based on truncating the stick-breaking representation of the Dirichlet process (32). The support $[a, b]$ can be deterministic (based on n), empirical ($a = X_{(1)}$, $b = X_{(n)}$), or hierarchical (a and $b - a$ given their own priors). Priors on $\gamma_1 \geq 0$, $\gamma_2 \in \mathbb{R}$, and $\theta_1, \ldots, \theta_m \in [0, b-a]$ complete the model, and posterior density draws can be obtained from MCMC samples of these parameters using (40). See Mariucci, Ray and Szabó for technical details, as well as motivation for (40). Pointwise credible intervals obtained with MCMC did a good job capturing true densities in their simulation studies, although in some cases they underperformed somewhat around boundaries or modes. The authors also evaluated the coverage probability of the intervals in one example, showing a tendency for undercoverage in some parts of the domain but overall reasonable performance with increasing sample sizes.

Similarly to [165] and [5], complicated limiting distributions have been derived for density estimation under different shape constraints. Examples include monotonicity with right-censored data [97], convexity [78], and s-concavity [90]. In principle these limiting distributions could be used to derive practical UQ methods for densities as in the examples described above, but there does not appear to be any literature directly doing so.

8.6. Connections to nonparametric regression

Various parts of this paper have suggested that some uncertainty quantification ideas from other nonparametric models could apply for density estimation. Indeed, there exist a great deal of theoretical results showing that many such models are "equivalent" in a sense involving asymptotic convergence of their risks [e.g. 154, 13, and references therein, especially those by Lucien Le Cam]. Brown et al. [14] offered a practical way of leveraging these ideas. They proposed the *root-unroot algorithm* to estimate a density on, say, $[0, 1]$ via nonparametric regression. The algorithm proceeds as follows.

1. Divide the domain, assumed wlog to be $[0, 1]$, into T equal subintervals.
2. For $j = 1, \ldots, T$, let $Y_j = \sqrt{Q_j + 1/4}$, where Q_j is the count of X_i's in the j^{th} subinterval and the offset of $1/4$ gives optimal bias and variance properties.
3. Treat the Y_j's as response variables and use any suitable method to fit the corresponding smooth regression function $\widehat{m}$.

4. Take $\hat{f}(\cdot) = [\widehat{m}(\cdot)]^2 / \int [\widehat{m}(t)]^2 \, \mathrm{d}t$ as the density estimate.

Wang [212] used the root-unroot algorithm for Bayesian density inference, using *integrated nested Laplace approximations* (INLA) for the posterior of the regression model. The details of INLA - first given by Rue, Martino and Chopin [181] - are omitted here, but it suffices to note that it uses Gaussian approximations and numerical integration to approximate the posterior, allowing for inference without MCMC being necessary. In the above algorithm, Wang took $\hat{m}$ to be the posterior mean from the INLA model. Letting γ denote the normalizing integral in the denominator, they divided the INLA quantiles of m by γ to obtain approximate pointwise credible intervals for f. Such intervals did an excellent job capturing true density shapes in their simulation studies.

More broadly, one may exploit the connections described here to quantify density uncertainty with any number of methods originally devised for nonparametric regression. Examples include confidence bands based on coverage of surrogate functions [63], or on relaxed notions of coverage that still try to minimize the extent to which the band excludes the true function but allow for nice adaptivity properties [20].

9. Simulation study

Recall Figure 1 from Section 2, which shows select combinations of density estimation and UQ methods for a simulated dataset. Having described many such methods in the preceding sections, a more thorough discussion of the figure is presented here.

The dataset $\boldsymbol{X}$ is a sample of size $n = 1000$ from the mixture density $f_0 = 0.5\mathcal{N}\left(\frac{1}{2}, \frac{1}{49}\right) + 0.5\mathcal{N}\left(\frac{5}{7}, \frac{1}{490}\right)$. This is a bimodal, everywhere-positive density with almost all of its mass contained in the interval $[0, 1]$, and its magnitude and curvature are close to zero at the boundaries of this interval. Thus, it "approximately satisfies" the assumptions made by many of the UQ methods discussed here, while having a fairly interesting shape which provides a good test for UQ methods.

The methods applied to $\boldsymbol{X}$ and shown in Figure 1 are as follows.

1. KDE with pointwise bias-corrected confidence intervals as in Calonico, Cattaneo and Farrell [21], and fixed-width bootstrap confidence bands based on the same bias correction [29]. The bandwidth was selected to minimize estimated integrated MSE (instead of pointwise MSE as in the former reference) in order to ensure a smooth estimator.
2. Adaptive basis expansion with Bernstein polynomials as in Petrone [160], with pointwise credible intervals and credible bands based on median absolute deviations [45].
3. Logspline estimation with stepwise knot selection [110] and exponentiated pointwise Gaussian confidence intervals using bootstrap standard error estimates [111].
4. A Dirichlet process mixture of Gaussians with a Normal-Inverse Gamma base measure. A marginal MCMC sampler was used (see Section 7.1)

but pointwise credible intervals incorporated "full uncertainty" by using posterior draws of the Dirichlet process obtained by conditional conjugacy [60].

All Bayesian methods were based on output from an appropriate MCMC sampler, and the level for all UQ methods was taken to be $1 - \alpha = 0.95$. The simulation study was conducted with the R programming language [199], and further details can be found in the supplementary material for this manuscript [141].

As noted in Section 2, the bands are expectedly wider than the pointwise intervals for both estimation methods shown on the top row of Figure 1. Note that the confidence sets for the KDE are not centered at the estimator due to the bias correction, and are in fact closer to the true density. However, they still fail to fully reach the height of the main mode. Certainly no conclusions can be made about the coverage probability of any UQ method when it is applied to only a single dataset, but further simulations (not shown; see [141]) suggested that this deficiency is typical for samples taken from the true density f_0, even when using pointwise instead of integrated MSE to select bandwidths. In fact, sample sizes in the millions were necessary to attain good coverage probability at the main mode, although the performance was much better at the smaller mode even for $n = 1000$. To some degree this is to be expected as the coverage error depends on higher-order derivatives of f_0 [21], but it is infeasible to fully predict this error in practice. This leads to an important point to be made about the difference between asymptotic and finite-sample behaviour: although Calonico, Cattaneo and Farrell [21] showed that these confidence intervals have coverage error ultimately decaying at the optimal rate with respect to n, there are no concrete guarantees for any finite sample size when using data-driven methods.

Recall from Section 3.2 that Cheng and Chen [29] provided bootstrap methods for both fixed- and variable-width bias-corrected confidence bands for KDE's. Here the former was used, as the latter involves bootstrapping a quantity which can have a zero denominator when using a compact kernel, as was the case here [141]. In contrast, the credible band used for the Bernstein polynomial estimator has variable width (see (13)). However, the band shown in the top-right plot of Figure 1 extends over the subinterval $[0.01, 0.99]$, as the band taken over all of $[0, 1]$ was far too wide to be graphically meaningful. This is because a sizeable proportion of MCMC draws had absolute deviations near the boundaries that were much larger than the MAD there, so that the quantile ξ_α in (13) was very large. In turn, the MAD was comparatively small at the boundaries because, like f_0 itself, most MCMC draws had tail values near zero. These examples demonstrate that variable-width bands may not be an ideal choice unless f_0 is bounded suitably far away from zero.

Interpretation of the bottom row of Figure 1 is straightforward. The pointwise intervals for the logspline estimator are noticeably less smooth than those for the other estimation methods. Recall that the width of the interval (on the log scale) is determined by the pointwise sample variance of bootstrap density estimates

[111]; evidently this induces some roughness. The pointwise credible intervals for the DP mixture are quite similar to those for the Bernstein polynomial estimator: both are quite narrow and smooth and encompass f_0 throughout nearly the entire domain.

10. Conclusion

There is a vast, sprawling body of work on density uncertainty quantification, dating back over half a century and spanning across many different methods for both estimation and inference. Reviewing the literature - from classical approaches like KDE's and histograms, to the spline methods of the late twentieth century, to modern nonparametric methods - one notices that the gap between theoretical and practical ideas seems to have widened over time. KDE's and related methods are extremely well-studied, with a litany of theoretical and practical results for all relevant types of UQ. Turning the focus to the past two decades of developments, one sees that UQ in the literature for random mixtures is entirely practical, with almost no regard for asymptotic properties; conversely, the advanced wavelet-based papers comprising the core of new theoretical developments often include no data studies whatsoever. It seems natural to wonder whether it is possible to "bridge the gap": perhaps introducing greater theoretical justification for some of the most commonly-used practical methods, or facilitating applications of some of the more obscure asymptotic arguments. However, such developments may be hampered by issues intrinsic to the problems at hand, such as the known complexities of asymptotics in nonparametric Bayesian inference Rousseau [e.g. the review of 178]. The importance of these considerations is certainly a subjective matter, and as modern practitioners turn their focus to larger datasets and more overt "data science" approaches, there is perhaps a case to be made that applications could provide "all the proof we need".

Based on the simulation study described in Section 9, Figure 1 shows finite-sample results for a few of the methods discussed throughout this paper. The code for these experiments is available in the supplementary material [141], and there is certainly merit to further comparative analysis beyond that considered here.

Of interest for future work are extensions to frameworks beyond a single i.i.d. sample, particularly hierarchical modelling of multiple related densities. Bayesian nonparametric methods are emerging as a promising approach to such frameworks, and we are eager to explore the improvements which further developments can provide.

Acknowledgements

The authors wish to thank the following people for clarifying some ideas in personal correspondence: Eric Cator, Zhong Guan, Maria Lomeli, Omiros Papaspiliopoulos, Yushi Shi, Richard Nickl, and Bin Wang. They also wish to thank the anonymous referees for their insights and improvements to this paper.

Supplementary Material

Supplement to "A Review of Uncertainty Quantification for Density Estimation": further details on the methods used in the simulation study, as well as R code for these and further studies.

References

[1] Argiento, R., Bianchini, I. and Guglielmi, A. (2016). Posterior sampling from ϵ-approximation of normalized completely random measure mixtures. *Electronic Journal of Statistics* **10** 3516–3547.

[2] Azadbakhsh, M., Jankowski, H. and Gao, X. (2014). Computing confidence intervals for log-concave densities. *Computational Statistics and Data Analysis* **75** 248–264.

[3] Babu, G. J., Canty, A. J. and Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference* **105** 377–392.

[4] Babu, G. J. and Chaubey, Y. P. (2006). Smooth estimation of a distribution and density function on a hypercube using Bernstein polynomials for dependent random vectors. *Statistics & Probability Letters* **76** 959–969.

[5] Balabdaoui, F., Rufibach, K. and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *The Annals of Statistics* **37** 1299–1331.

[6] Banerjee, M. and Wellner, J. A. (2001). Likelihood ratio tests for monotone functions. *The Annals of Statistics* **29** 1699–1731.

[7] Barrios, E., Lijoi, A., Nieto-Barajas, L. E. and Prünster, I. (2013). Modeling with Normalized Random Measure Mixture Models. *Statistical Science* **28** 313–334.

[8] Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L. and Rodríguez, C. (2011). A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics* **5** 204–237.

[9] Bickel, P. J. and Rosenblatt, M. (1973). On Some Global Measures of the Deviations of Density Function Estimates. *The Annals of Statistics* **1** 1071–1095.

[10] Bissantz, N., Dümbgen, L., Holzmann, H. and Munk, A. (2007). Non-Parametric Confidence Bands in Deconvolution Density Estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69** 483–506.

[11] Bouezmarni, T. and Rombouts, J. V. K. (2010). Nonparametric density estimation for positive time series. *Computational Statistics & Data Analysis* **54** 245–261.

[12] Branscum, A. J. and Hanson, T. E. (2008). Bayesian Nonparametric Meta-Analysis Using Polya Tree Mixture Models. *Biometrics* **64** 825–833.

[13] Brown, L. D., Carter, A. V., Low, M. G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *The Annals of Statistics* **32** 2074–2097.
[14] Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probability Theory and Related Fields* **146** 401–433.
[15] Bull, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics* **6** 1490–1516.
[16] Bull, A. D. (2013). A Smirnov-Bickel-Rosenblatt Theorem for Compactly-Supported Wavelets. *Constructive Approximation* **37** 295–309.
[17] Bull, A. D. and Nickl, R. (2013). Adaptive confidence sets in L^2. *Probability Theory and Related Fields* **156** 889–919.
[18] Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics* **18** 179–189.
[19] Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics* **32** 1805–1840.
[20] Cai, T. T., Low, M. and Ma, Z. (2014). Adaptive Confidence Bands for Nonparametric Regression Functions. *Journal of the American Statistical Association* **109** 1054–1070.
[21] Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018). On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association* **113** 767–779.
[22] Canale, A. (2017). msBP: An R Package to Perform Bayesian Nonparametric Inference Using Multiscale Bernstein Polynomials Mixtures. *Journal of Statistical Software* **78** 1–19.
[23] Canale, A. and Dunson, D. B. (2016). Multiscale Bernstein polynomials for densities. *Statistica Sinica* **26** 1175–1195.
[24] Castillo, I. (2017). Pólya tree posterior distributions on densities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **53** 2074–2102.
[25] Castillo, I. and Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics* **42** 1941–1969.
[26] Chacón, J. E. and Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. CRC Press.
[27] Chen, S. X. (1996). Empirical Likelihood Confidence Intervals for Nonparametric Density Estimation. *Biometrika Trust* **83** 329–341.
[28] Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology* **1** 161–187.
[29] Cheng, G. and Chen, Y.-C. (2019). Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics* **13** 2194–2256.
[30] Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics* **42** 1787–1818.

[31] Choudhuri, N., Ghosal, S. and Roy, A. (2004). Bayesian Estimation of the Spectral Density of a Time Series. *Journal of the American Statistical Association* **99** 1050–1059.
[32] Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes Conditional Distribution Modeling With Variable Selection. *Journal of the American Statistical Association* **104** 1646–1660.
[33] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis* **1** 54–81.
[34] Csörgő, M. (1983). An Invariance Principle for Nearest-Neighbor Empirical Density Functions. In *Quantile Processes with Statistical Applications* 9, 137–143. Society for Industrial and Applied Mathematics.
[35] Csörgo, M. and Révész, P. (1981). *Strong approximations in probability and statistics*. Academic Press.
[36] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.
[37] De Iorio, M., Müller, P., Rosner, G. L. and Maceachern, S. N. (2004). An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association* **99** 205–215.
[38] De Iorio, M., Johnson, W. O., Müller, P. and Rosner, G. L. (2009). Bayesian Nonparametric Nonproportional Hazards Survival Modeling. *Biometrics* **65** 762–771.
[39] Deng, H., Han, Q. and Zhang, C.-H. (2020). Confidence intervals for multiple isotonic regression and other monotone models Technical Report.
[40] Dias, R. (2011). Nonparametric Estimation: Smoothing and Visualization Technical Report.
[41] Dümbgen, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *The Annals of Statistics* **26** 288–314.
[42] Dunson, D. B. and Park, J.-H. (2008). Kernel Stick-Breaking Processes. *Biometrika* **95** 307–323.
[43] Dunson, D. B., Pillai, N. and Park, J.-H. (2007). Bayesian Density Regression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69** 163–183.
[44] Durot, C., Kulikov, V. N. and Lopuhaä, H. P. (2012). The limit distribution of the L_∞-error of Grenander-type estimators. *The Annals of Statistics* **40** 1578–1608.
[45] Edwards, M. C., Meyer, R. and Christensen, N. (2019). Bayesian nonparametric spectral density estimation using B-spline priors. *Statistics and Computing* **29** 67–78.
[46] Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science* **11** 89–121.
[47] Eilers, P. H. C., Marx, B. D. and Durbán, M. (2015). Twenty years of P-splines. *Statistics & Operations Research Transactions SORT* **39** 149.
[48] Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90** 577–588.

[49] Fan, J. (1991). Asymptotic normality for deconvolution kernel density estimators. *Sankhya : The Indian Journal of Statistics* **53** 97–110.

[50] Fan, Y. and Liu, Y. (1997). A Note on Asymptotic Normality for Deconvolution Kernel Density Estimators. *Sankhyā: The Indian Journal of Statistics, Series A* **59** 138–141.

[51] Favaro, S., Lomeli, M. and Teh, Y. W. (2015). On a class of σ-stable Poisson-Kingman models and an effective marginalized sampler. *Statistics and Computing* **25** 67–78.

[52] Favaro, S. and Teh, Y. W. (2013). MCMC for Normalized Random Measure Mixture Models. *Statistical Science* **28** 335–359.

[53] Favaro, S. and Walker, S. G. (2013). Slice Sampling σ-Stable Poisson-Kingman Mixture Models. *Journal of Computational and Graphical Statistics* **22** 830–847.

[54] Favaro, S., Lomeli, M., Nipoti, B. and Teh, Y. W. (2014). On the stick-breaking representation of σ-stable Poisson-Kingman models. *Electronic Journal of Statistics* **8** 1063–1085.

[55] Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Prünster, I. and Teh, Y. W. (2016). On the Stick-Breaking Representation for Homogeneous NRMIs. *Bayesian Analysis* **11** 697–724.

[56] Finucane, M. M., Paciorek, C. J., Stevens, G. A. and Ezzati, M. (2015). Semiparametric Bayesian Density Estimation With Disparate Data Sources: A Meta-Analysis of Global Childhood Undernutrition. *Journal of the American Statistical Association* **110** 889–901.

[57] Fiorio, C. V. (2004). Confidence intervals for kernel density estimation. *The Stata Journal* **4** 168–179.

[58] Freedman, D. and Diaconis, P. (1981). On the Maximum Deviation Between the Histogram and the Underlying Density. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **58** 139–167.

[59] Gawronski, W. and Stadtmüller, U. (1981). Smoothing Histograms by Means of Lattice-and Continuous Distributions. *Metrika* **28** 155–164.

[60] Gelfand, A. E. and Kottas, A. (2002). A Computational Approach for Full Nonparametric Bayesian Inference Under Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **11** 289–305.

[61] Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Source: Journal of the American Statistical Association* **85** 398–409.

[62] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). Finite mixture models. In *Bayesian Data Analysis* 3 ed. 22, 519–544. Chapman and Hall/CRC.

[63] Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *The Annals of Statistics* **36** 875–905.

[64] Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics* **29** 1264–1280.

[65] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

[66] Giné, E., Koltchinskii, V. and Sakhanenko, L. (2003). Conver-

gence in distribution of Self-Normalized Sup-Norms of Kernel Density Estimators. In *High-Dimensional Probability III, Progress in Probability* (J. Hoffmann-Jørgensen, J. A. Wellner and M. B. Marcus, eds.) **55** 241–253. Birkhäuser, Basel.

[67] Giné, E., Koltchinskii, V. and Sakhanenko, L. (2004). Kernel density estimators: convergence in distribution for weighted sup-norms. *Probability Theory and Related Fields* **130** 167–198.

[68] Giné, E. and Mason, D. M. (2007). On local U-statistic processes and the estimation of densities of functions of several sample variables. *The Annals of Statistics* **35** 1105–1145.

[69] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics* **38** 1122–1170.

[70] Giné, E. and Nickl, R. (2015). Adaptive Inference. In *Mathematical Foundations of Infinite-Dimensional Statistical Models* 607–666. Cambridge University Press, Cambridge.

[71] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* **138** 5674–5685.

[72] Good, I. J. and Gaskins, R. A. (1971). Nonparametric Roughness Penalties for Probability Densities. *Biometrika* **58** 255–277.

[73] Grenander, U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal* **1956** 125–153.

[74] Griffin, J. E. (2010). Default priors for density estimation with mixture models. *Bayesian Analysis* **5** 45–64.

[75] Griffin, J. E. (2016). An adaptive truncation method for inference in Bayesian nonparametric models. *Statistics and Computing* **26** 423–441.

[76] Griffin, J. E., Kolossiatis, M. and Steel, M. F. J. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75** 499–529.

[77] Groeneboom, P., Hooghiemstra, G. and Lopuhaa, H. P. (1999). Asymptotic Normality of the L_1 Error of the Grenander Estimator. *The Annals of Statistics* **27** 1316–1347.

[78] Groeneboom, P., Jongbloed, G. and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *The Annals of Statistics* **29** 1653–1698.

[79] Groeneboom, P., Jongbloed, G. and Witte, B. I. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *The Annals of Statistics* **38** 352–387.

[80] Groeneboom, P. and Jongbloed, G. (2015). Nonparametric confidence intervals for monotone functions. *The Annals of Statistics* **43** 2019–2054.

[81] Guan, Z. (2016). Efficient and robust density estimation using Bernstein type polynomials. *Journal of Nonparametric Statistics* **28** 250–271.

[82] Gutiérrez, L., Mena, R. H. and Ruggiero, M. (2016). A time dependent Bayesian nonparametric model for air quality analysis. *Computational Statistics and Data Analysis* **95** 161–175.

[83] Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields* **89** 447–455.
[84] Hall, P. (1992). Effect of Bias Estimation on Coverage Accuracy of Bootstrap Confidence Intervals for a Probability Density. *The Annals of Statistics* **20** 675–694.
[85] Hall, P. (1993). On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* **55** 291–304.
[86] Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics* **41** 1892–1921.
[87] Hall, P. and Kang, K.-H. (2001). Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *The Annals of Statistics* **29** 1443–1468.
[88] Hall, P. and Owen, A. B. (1993). Empirical Likelihood Confidence Bands in Density Estimation. *Journal of Computational and Graphical Statistics* **2** 273–289.
[89] Hall, P. and Titterington, D. M. (1988). On Confidence Bands in Nonparametric Density Estimation and Regression. *Journal of Multivariate Analysis* **27** 228–254.
[90] Han, Q. and Wellner, J. A. (2016). Approximation and estimation of s-concave densities via Rényi divergences. *The Annals of Statistics* **44** 1332–1359.
[91] Hansen, M. H. and Kooperberg, C. (2002). Spline Adaptation in Extended Linear Models. *Statistical Science* **17** 2–51.
[92] Hanson, T. E. (2006). Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association* **101** 1548–1565.
[93] Hanson, T., Zhou, H. and Inácio De Carvalho, V. (2018). Bayesian Nonparametric Spatially Smoothed Density Estimation. In *New Frontiers of Biostatistics and Bioinformatics* (Y. Zhao and D.-G. Chen, eds.) 4, 87–105. Springer International Publishing.
[94] Hengartner, N. W. and Stark, P. B. (1995). Finite-Sample Confidence Envelopes for Shape-Restricted Densities. *The Annals of Statistics* **23** 525–550.
[95] Hoffmann, M. and Nickl, R. (2011). On adaptive inerence and confidence bands. *The Annals of Statistics* **39** 2383–2409.
[96] Horowitz, J. L. (2001). The Bootstrap. In *Handbook of Econometrics*.
[97] Huang, Y. and Zhang, C.-H. (1994). Estimating a Monotone Density from Censored Observations. *The Annals of Statistics* **22** 1256–1274.
[98] Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* **96** 161–173.
[99] Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87** 371–390.
[100] James, L. F., Lijoi, A. and Prünster, I. (2009). Posterior Analysis

for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics* **36** 76–97.

[101] Jones, M. C. (1993). Simple boundary correction for density estimation kernel. *Statistics and Computing* **3** 135–146.

[102] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association* **91** 401–407.

[103] Kaiser, G. (2011). *A Friendly Guide to Wavelets*. Birkhäuser Boston, Boston.

[104] Kalli, M., Griffin, J. E. and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing* **21** 93–105.

[105] Kerkyacharian, G., Nickl, R. and Picard, D. (2012). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probability Theory and Related Fields* **153** 363–404.

[106] Kim, B. K. and Van Ryzin, J. (1980). On the asymptotic distribution of a histogram density estimator. In *Colloquia Mathematica Societatis Janos Bolyai, 32. Nonparametric Stat. Inference* 483–499.

[107] Kim, B. K. and Van Ryzin, J. (1985). A bivariate histogram density estimator: Consistency and asymptotic normality. *Statistics & Probability Letters* **3** 167–173.

[108] Komárek, A., Lesaffre, E. and Hilton, J. F. (2005). Accelerated Failure Time Model for Arbitrarily Censored Data With Smoothed Error Distribution. *Journal of Computational and Graphical Statistics* **14** 726–745.

[109] Komárek, A. and Lesaffre, E. (2008). Bayesian Accelerated Failure Time Model With Multivariate Doubly Interval-Censored Data and Flexible Distributional Assumptions. *Journal of the American Statistical Association* **103** 523–533.

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

[111] Kooperberg, C. and Stone, C. J. (2003). Confidence Intervals for Logspline Density Estimation. 285–295. Springer, New York, NY.

[112] Kooperberg, C. and Stone, C. J. (2004). Comparison of Parametric and Bootstrap Approaches to Obtaining Confidence Intervals for Logspline Density Estimation. *Journal of Computational and Graphical Statistics* **13** 106–122.

[113] Kosorok, M. R. (2008). Bootstrapping the Grenander estimator. **1** 282–292.

[114] Kottas, A. (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference* **136** 578–596.

[115] Laloë, T. and Servien, R. (2016). A note on the asymptotic law of the histogram without continuity assumptions. *Brazilian Journal of Probability and Statistics* **30** 562–569.

[116] Lambert, P. and Eilers, P. H. C. (2006). Bayesian multi-dimensional

density estimation with P-splines Technical Report.
[117] Lambert, P. and Eilers, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis* **53** 1388–1399.
[118] Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics* **13** 183–212.
[119] Lavine, M. (1992). Some Aspects of Polya Tree Distributions for Statistical Modelling. *The Annals of Statistics* **20** 1222–1235.
[120] Leblanc, A. (2010). A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics* **22** 459–475.
[121] Lepskiĭ, O. V. (1992). Asymptotically Minimax Adaptive Estimation. I: Upper Bounds. Optimally Adaptive Estimates. *Theory of Probability & Its Applications* **36** 682–697.
[122] Lerasle, M. (2012). Adaptive non-asymptotic confidence balls in density estimation. *ESAIM - Probability and Statistics* **16** 61–85.
[123] Li, K.-C. (1989). Honest Confidence Regions for Nonparametric Regression. *The Annals of Statistics* **17** 1001–1008.
[124] Lijoi, A., Prünster, I. and Walker, S. G. (2008). Investigating nonparametric priors with Gibbs structure. *Statistica Sinica* **18** 1653–1668.
[125] Liu, Y. and Zhang, Y. (2010). The consistency and asymptotic normality of nearest neighbor density estimator under α-mixing condition. *Acta Mathematica Scientia* **30** 733–738.
[126] Lo, A. Y. (1987). A Large Sample Study of the Bayesian Bootstrap. *The Annals of Statistics* **15** 360–375.
[127] Lo, S. H., Mack, Y. P. and Wang, J. L. (1989). Density and Hazard Rate Estimation for Censored Data via Strong Representation of the Kaplan-Meier Estimator. *Probability Theory and Related Fields* **80** 461–473.
[128] Lomelí, M., Favaro, S. and Teh, Y. W. (2017). A Marginal Sampler for σ-Stable Poisson-Kingman Mixture Models. *Journal of Computational and Graphical Statistics* **26** 44–53.
[129] Lopes, H. F. and Dias, R. (2011). Bayesian Mixture of Parametric and Nonparametric Density Estimation: A Misspecification Problem. *Brazillian Review of Econometrics* **31** 19–44.
[130] Louani, D. (1998). On the asymptotic normality of the kernel estimators of the density function and its derivatives under censoring. *Communications in Statistics - Theory and Methods* **27** 2909–2924.
[131] Lounici, K. and Nickl, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *The Annals of Statistics* **39** 201–231.
[132] Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics* **25** 2547–2554.
[133] Low, M. G. and Ma, Z. (2015). Discussion of "Frequentist coverage of adaptive nonparametric Bayesian credible sets". *Ann. Statist.* **43** 1448–1454.
[134] MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science* **1** 50–55.

[135] Mack, Y. P. (1980). Asymptotic Normality of Multivariate k-NN Density Estimates. *Sankhyā: The Indian Journal of Statistics, Series A* **42** 53–63.
[136] Mack, Y. P. and Rosenblatt, M. (1979). Multivariate k-Nearest Neighbor Density Estimates. *Journal of Multivariate Analysis* **9** 1–15.
[137] Mariucci, E., Ray, K. and Szabó, B. (2020). A Bayesian nonparametric approach to log-concave density estimation. *Bernoulli* **26** 1070–1097.
[138] Martin, R. (2019). Empirical Priors and Posterior Concentration Rates for a Monotone Density. *Sankhyā : The Indian Journal of Statistics* **81** 493–509.
[139] Masry, E. (1993). Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes. *Journal of Multivariate Analysis* **44** 47–68.
[140] Mâsse, B. R. and Truong, Y. K. (1999). Conditional Logspline Density Estimation. *The Canadian Journal of Statistics* **27** 819–832.
[141] McDonald, S. and Campbell, D. Supplement to "A Review of Uncertainty Quantification for Density Estimation" Technical Report.
[142] Mielniczuk, J. (1986). Some Asymptotic Properties of Kernel Estimators of a Density Function in Case of Censored Data. *The Annals of Statistics* **14** 766–773.
[143] Mokkadem, A. and Pelletier, M. (2006). Confidence bands for densities, logarithmic point of view. *Alea* **2** 231–266.
[144] Moore, D. S. and Yackel, J. W. (1977). Large sample properties of nearest neighbor density function estimators. In *Statistical Decision Theory and Related Topics* 269–279. Elsevier.
[145] Müller, P. and Rodriguez, A. (2013a). Dependent Dirichlet Processes and Other Extensions. In *Nonparametric Bayesian Inference* 53–75.
[146] Müller, P. and Rodriguez, A. (2013b). Pólya Trees. In *Nonparametric Bayesian Inference* 43–51. Institute of Mathematical Statistics; American Statistical Association.
[147] Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **9** 249–265.
[148] Neumann, M. H. (1998). Strong Approximation of Density Estimators from Weakly Dependent Observations by Density Estimators from Independent Observations. *The Annals of Statistics* **26** 2014–2048.
[149] Nieto-Barajas, L. E. and Müller, P. (2012). Rubbery Polya Tree. *Scandinavian Journal of Statistics* **39** 166–184.
[150] Nieto-Barajas, L. E., Müller, P., Ji, Y., Lu, Y. and Mills, G. B. (2012). A Time-Series DDP for Functional Proteomics Profiles. *Biometrics* **68** 859–868.
[151] Norets, A. and Pati, D. (2017). Adaptive Bayesian estimation of conditional densities. *Econometric Theory* **33** 980–1012.
[152] Norets, A. and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics* **168** 332–346.
[153] Norets, A. and Pelenis, J. (2017). Posterior consistency in conditional

density estimation by covariate dependent mixtures. *Econometric Theory* **30** 606–646.
[154] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *The Annals of Statistics* **24** 2399–2430.
[155] O'Sullivan, F. (1988). Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators. *SIAM Journal on Scientific and Statistical Computing* **9** 363–379.
[156] Paddock, S. M., Ruggeri, F., Lavine, M. and West, M. (2003). Randomized Polya tree models for nonparametric Bayesian inference. *Statistica Sinica* **13** 443–460.
[157] Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95** 169–186.
[158] Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica* **20** 1203–1226.
[159] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33** 1065–1076.
[160] Petrone, S. (1999a). Random Bernstein Polynomials. *Scandinavian Journal of Statistics* **26** 373–393.
[161] Petrone, S. (1999b). Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics* **27** 105–126.
[162] Petrone, S. and Veronese, P. (2002). Non parametric mixture priors based on an exponential random scheme. *Statistical Methods & Applications* **11** 1–20.
[163] Pitman, J. (2003). Poisson-Kingman Partitions. *Lecture Notes-Monograph Series* **40** 1–34.
[164] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25** 855–900.
[165] Prekasa Rao, B. L. S. (1969). Estimation of a Unimodal Density. *Sankhyā: The Indian Journal of Statistics, Series A* **31** 23–36.
[166] Ramsay, J. O. and Siverman, B. W. (2005). *Functional Data Analysis. Springer Series in Statistics*. Springer-Verlag, New York.
[167] Ray, K. (2017). Adaptive Bernstein-von Mises theorems in Gaussian white nosie. *The Annals of Statistics* **45** 2511–2536.
[168] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics* **31** 560–585.
[169] Révész, P. (1972). On empirical density function. *Periodica Mathematica Hungarica* **2** 85–110.
[170] Riihimäki, J. and Vehtari, A. (2014). Laplace Approximation for Logistic Gaussian Process Density Estimation and Regression. *Bayesian Analysis* **9** 425–448.
[171] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probability Theory and Related Fields* **98** 21–45.
[172] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric

confidence sets. *The Annals of Statistics* **34** 229–253.
[173] Rodríguez, C. C. (1986). On a New Class of Density Estimators Technical Report, State University of New York at Albany.
[174] Rodríguez, C. C. (2003). Optimal recovery of local truth. In *AIP Conference Proceedings* **567** 89–115. AIP Publishing.
[175] Rodriguez, C. and Van Ryzin, J. (1992). Large sample properties of maximum entropy histograms. *IEEE Transactions on Information Theory* **32** 751–759.
[176] Roeder, K. (1990). Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. *Journal of the American Statistical Association* **85** 617–624.
[177] Rosenblatt, M. (1976). On the Maximal Deviation of k-Dimensional Density Estimates. *The Annals of Probability* **4** 1009–1015.
[178] Rousseau, J. (2016). On the Frequentist Properties of Bayesian Nonparametric Methods. *The Annual Review of Statistics and Its Applications* **3** 211–231.
[179] Rousseau, J. and Szabó, B. (2019). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *Annals of Statistics (to appear)*.
[180] Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics* **9** 130–134.
[181] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 319–392.
[182] Rufibach, K. (2006). Log-concave density estimation and bump hunting for IID observations, PhD thesis, Verlag nicht ermittelbar.
[183] Sakhanenko, L. (2015). Asymptotics of suprema of weighted Gaussian fields with applications to kernel density estimators. *Theory of Probability & its Applications* **59** 415–451.
[184] Sardy, S. and Tseng, P. (2010). Density Estimation by Total Variation Penalized Likelihood Driven by the Sparsity ℓ_1 Information Criterion. *Scandinavian Journal of Statistics* **37** 321–337.
[185] Schick, A. and Wefelmeyer, W. (2004). Root n consistent density estimators for sums of independent random variables. *Journal of Nonparametric Statistics* **16** 925–935.
[186] Schick, A. and Wefelmeyer, W. (2006). Pointwise convergence rates and central limit theorems for kernel density estimators in linear processes. *Statistics & Probability Letters* **76** 1756–1760.
[187] Sen, B., Banerjee, M. and Woodroofe, M. (2010). Inconsistency of bootstrap: the Grenander estimator. *The Annals of Statistics* **38** 1953–1977.
[188] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–650.
[189] Sharef, E., Strawderman, R. L., Ruppert, D., Cowen, M. and Halasyamani, L. (2010). Bayesian adaptive B-spline estimation in propor-

tional hazards frailty models. *Electronic Journal of Statistics* **4** 606–642.
[190] Shen, W. and Ghosal, S. (2015). Adaptive Bayesian Procedures Using Random Series Priors. *Scandinavian Journal of Statistics* **42** 1194–1213.
[191] Shi, Y. (2020). DPWeibull: Dirichlet Process Weibull Mixture Model for Survival Data. R package version 1.5.
[192] Shi, Y., Martens, M., Banerjee, A. and Laud, P. (2019). Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models. *Bayesian Analysis* **14** 677–702.
[193] Silverman, B. W. (1982). On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method. *The Annals of Statistics* **10** 795–810.
[194] Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR* **74** 189–191.
[195] Stadtmüller, U. (1983). Asymptotic Distributions of Smoothed Histograms. *Metrika* **30** 145–158.
[196] Stone, C. J. (1986). Asymptotic properties of logspline density estimation Technical Report, Department of Statistics, University of California, Berkeley, California.
[197] Stone, C. J. (1990). Large-Sample Inference for Log-Spline Models. *The Annals of Statistics* **18** 717–741.
[198] Szabó, B., van der Vaart, A. W. and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics* **43** 1391–1428.
[199] R Core Team (2020). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
[200] Tenbusch, A. (1994). Two-Dimensional Bernstein Polynomial Density Estimators. *Metrika* **41** 233–253.
[201] Uh, H.-w. (2003). Kernel Deconvolution, PhD thesis, University of Amsterdam.
[202] van de Wiel, M. A., Te Beest, D. E. and Münch, M. M. (2019). Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics* **46** 2–25.
[203] Van Der Vaart, A. W. and Van Der Laan, M. J. (2003). Smooth estimation of a monotone density. *Statistics: A Journal of Theoretical and Applied Statistics* **37** 189–203.
[204] van Es, A. J. and Uh, H. W. (2004). Asymptotic normality of nonparametric kernel type deconvolution density estimators: crossing the Cauchy boundary. *Nonparametric Statistics* **16** 261–277.
[205] van Es, B. and Uh, H.-W. (2005). Asymptotic Normality of Kernel-Type Deconvolution Estimators. *Scandinavian Journal of Statistics* **32** 467–483.
[206] Van Kerm, P. (2003). Adaptive kernel density estimation. *The Stata Journal* **3** 148–156.
[207] Vermeesch, P. (2005). Statistical uncertainty associated with histograms in the Earth sciences. *Journal of Geophysical Research* **110**.

[208] Vitale, R. A. (1975). A Bernstein Polynomial Approach to Density Function Estimation. In *Statistical Inference and Related Topics* 87–99. Academic Press.

[209] Wade, S., Dunson, D. B., Petrone, S. and Trippa, L. (2014). Improving Prediction from Dirichlet Process Mixtures via Enrichment. *Journal of Machine Learning Research* **15** 1041–1071.

[210] Wahba, G. (1983). Bayesian "Confidence Intervals" for the Cross-Validated Smoothing. *Journal of the Royal Statistical Society. Series B (Methodological)* **45** 133–150.

[211] Walker, S. G. (2007). Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics-Simulation and Computation* **36** 45–54.

[212] Wang, X.-F. (2013). Bayesian nonparametric regression and density estimation using integrated nested Laplace approximations. *Journal of Biometrics and Biostatistics* **25**.

[213] Wang, L. and Dunson, D. B. (2011). Fast Bayesian Inference in Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* **20**.

[214] Wang, B. and Wertelecki, W. (2013). Density estimation for data with rounding errors. *Computational Statistics and Data Analysis* **65** 4–12.

[215] Yeh, A. B. (1996). Bootstrap percentile confidence bands based on the concept of curve depth. *Communications in Statistics Part B: Simulation and Computation* **25** 905–922.

[216] Zamini, R., Fakoor, V. and Sarmad, M. (2014). Asymptotic Behaviors of Nearest Neighbor Kernel Density Estimator in Left-truncated Data. *Journal of Sciences, Islamic Republic of Iran* **25** 57–67.

[217] Zhang, C.-H. (1990). Fourier Methods for Estimating Mixing Densities and Distributions. *The Annals of Statistics* **18** 806–831.

[218] Zhang, S., Müller, P. and Do, K.-A. (2010). A Bayesian Semiparametric Survival Model with Longitudinal Markers. *Biometrics* **66** 435–443.

[219] Zu, Y. (2015). A Note on the Asymptotic Normality of the Kernel Deconvolution Density Estimator with Logarithmic Chi-Square Noise. *Econometrics* **3** 561–576.